PRECONDITIONED ITERATIVE METHODS AND PERTURBATION ANALYSIS FOR A CLASS OF SADDLE POINT PROBLEMS

Ph.D. THESIS

By

PINKI KHATUN

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

Preconditioned Iterative Methods and Perturbation Analysis for a Class of Saddle Point Problems

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree

of

DOCTOR OF PHILOSOPHY

by

Pinki Khatun

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

I hereby certify that the work which is being presented in the thesis entitled **Preconditioned Iterative Methods and Perturbation Analysis for a Class of Saddle Point Problems** in the partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** and submitted in the **Department of Mathematics, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from **August 2020** to **March 2025** under the supervision of **Prof. Sk. Safique Ahmad, Professor, Indian Institute of Technology Indore**.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Pinki Khatun 29/05/2025

Signature of the student with date

(Pinki Khatun)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

29/05/2025

Signature of thesis supervisor with date

(Prof. Sk. Safique Ahmad)

Pinki Khatun has successfully given her Ph.D. Oral Examination held on 29 May, 2025.

29/05/2025

Signature of thesis supervisor with date

(Prof. Sk. Safique Ahmad)

ACKNOWLEDGEMENTS

My Ph.D. journey at IIT Indore has been an enriching and transformative experience, marked by profound learning, introspection, and growth. While deeply personal, this journey would not have been possible without the support, encouragement, and contributions of many individuals, to whom I extend my deepest gratitude.

First and foremost, I would like to begin by expressing my heartfelt gratitude to my supervisor, **Prof. Sk. Safique Ahmad**, for his invaluable guidance, unwavering support, and constant encouragement throughout my doctoral journey. His profound knowledge and insightful suggestions have played a pivotal role in shaping the course of my research. I am deeply thankful for the numerous enlightening discussions on research topics and for his mentorship, which has significantly enhanced my reading and writing skills.

I am deeply grateful to the **Director of IIT Indore** and the **Head of the Mathematics Department** for fostering an environment conducive to research. I want to extend my thanks to the **Council of Scientific and Industrial Research (CSIR)**, Govt. of India (grant no. 09/1022(0098)/EMR-I) for providing me with financial support. I would also like to extend my heartfelt appreciation to my PSPC members, **Prof. Niraj Kumar Shukla** and **Dr. Vijay Kumar Sohani**, for their constructive feedback and invaluable suggestions, which have greatly contributed to the development of my work.

My time at IIT Indore has been profoundly enriched by the incredible people I have had the privilege to meet. I am especially grateful to **Md Sajid** for his unwavering support and encouragement. His constant motivation and generous willingness to review my manuscript, offering invaluable feedback and insightful suggestions, have been truly instrumental in this journey. Additionally, I would also like to extend my heartfelt gratitude to my research group members, **Gyan Sir** and **Neha**, for their steadfast support throughout this endeavor. I am also sincerely thankful to **Shreyas**, **Anuradha**, **Abdul**, **Abdur Rahaman**, **Mushir**, **Dipendu**, **Himanshi**, **Navneet**, **Sahil** and **Abhishek** for their companionship and making this journey at IIT Indore truly memorable.

I come from a small village where access to higher education is rare, especially for girls. I am deeply grateful to my father, **Ali Hossen Miah**, and my mother, **Sahana Bibi**, whose unwavering faith in me and extraordinary resilience empowered me to pursue my dreams. Despite countless challenges, they provided me with every opportunity to focus on my education, and for that, I will always be indebted to them. I would like to express my gratitude to my amazing nephews, **Nabab** and **Ehsaan**, whose joy and

laughter provided me with much-needed moments of relaxation and happiness throughout this journey.

Finally, I am deeply indebted and grateful to the Most Merciful and the Most Beneficent, "**Almighty Allah**", without whose help I would not have been able to accomplish this journey.

(Pinki Khatun)

DEDICATION

To my parents

ABSTRACT

Saddle point problems (SPPs) have gained significant attention due to their diverse applications in computational science and engineering domains. This underscores the need for their efficient and robust solution methods. However, round-off and truncation errors in existing numerical approaches restrict solutions to approximations, raising critical concerns about their accuracy, sensitivity and reliability. To overcome these challenges, this thesis introduces preconditioned iterative methods for solving SPPs efficiently and employs perturbation analysis to assess the sensitivity and stability of the computed solutions. We specifically focus on two types of SPPs: the generalized saddle point problem (GSPP) and the double saddle point problem (DSPP).

Firstly, this thesis focuses on the development of novel iterative methods and preconditioners for DSPPs characterized by three-by-three block structures. Specifically, we propose two classes of shift-splitting iterative methods along with corresponding preconditioners tailored for DSPPs. A comprehensive convergence analysis is provided to establish the theoretical foundations of these methods. Additionally, we conduct a spectral analysis of the preconditioned matrices to better understand their efficiency and effectiveness. To evaluate the efficiency of the proposed preconditioners, we apply them to DSPPs arising from PDE-constrained optimization problems and Stokes equations.

Next, in this thesis, we address several fundamental questions: How sensitive is the solution when structure-preserving perturbations are applied to the coefficient matrix of GSPP or DSPP? Does a backward stable algorithm for solving the GSPP or DSPP also exhibit strong backward stability? What is the nearest GSPP or DSPP for which the approximate solution becomes the exact one?

To address these questions, in this thesis, we study structured backward errors (BEs) and structured condition numbers (CNs) for the GSPP and DSPP. We derive structured BEs for the GSPP and DSPP by preserving key properties of the block matrices of coefficient matrices, such as sparsity and linear structures (e.g., symmetric, Hermitian, Toeplitz, and circulant) within the corresponding perturbation matrices. Through this analysis, we demonstrate that a backward stable algorithm for solving an SPP may not always exhibit strong backward stability. Since the sensitivity of individual solution components in SPPs can vary significantly, we investigate partial normwise condition number (NCN), mixed condition number (MCN), and componentwise condition number (CCN) for both the GSPP and DSPP to evaluate the conditioning of each component independently. Furthermore, we examine structured CNs for both GSPPs and DSPPs by applying

structure-preserving perturbations to the block matrices, thereby capturing the impact of inherent structural properties on the stability of the solutions. We also introduce partial unified CNs for DSPPs, which encompass traditional NCN, MCN, and CCN, and reveal the sensitivity of individual components.

By leveraging the connection between SPPs and various least squares (LS) problems—such as weighted regularized least squares (WRLS) and equality-constrained indefinite least squares (EILS) problems—we apply our developed frameworks for CNs and BEs to derive explicit expressions for CNs and BEs in these contexts.

Finally, we extend our investigation to structured CNs for LS problems and the Moore-Penrose inverse, particularly when the associated matrices are rank-deficient with specific rank structures. To address this, we develop a general framework for computing the upper bounds of the MCN and CCN for rank-deficient structured matrices. This framework significantly improves the efficiency of calculating upper bounds for structured CNs, enabling faster and more accurate computations.

LIST OF PUBLICATIONS

List of Published/Communicated Research Papers from the Thesis

1. S. S. Ahmad and **P. Khatun**, "Structured backward errors for special classes of saddle point problems with applications." *Linear Algebra and its Applications*, 13:90-112, 2025 (SCI, Q1, MCQ: 0.78).
   DOI: https://doi.org/10.1016/j.laa.2025.03.003
2. S. S. Ahmad and **P. Khatun**, "A robust parameterized enhanced shift-splitting preconditioner for three-by-three block saddle point problems." *Journal of Computational and Applied Mathematics,* 459:116358, 2025 (SCI, Q2, MCQ: 0.83).
   DOI: https://doi.org/10.1016/j.cam.2024.116358
3. S. S. Ahmad and **P. Khatun**, "Structured condition numbers for a linear function of the solution of the generalized saddle point problems." *Electronic Transactions on Numerical Analysis,* 60:471-500, 2024 (SCI, Q2, MCQ: 0.87).
   DOI: https://doi.org/10.1553/etna_vol60s471
4. S. S. Ahmad and **P. Khatun**, "Condition numbers for the Moore-Penrose inverse and the least squares problem involving rank-structured matrices." *Linear and Multilinear Algebra,* 4:1-37, 2024 (SCI, Q1, MCQ: 0.68).
   DOI: https://doi.org/10.1080/03081087.2024.2410962
5. S. S. Ahmad and **P. Khatun**, "A class of generalized shift-splitting preconditioners for double saddle point problems." *arXiv preprint arXiv:2408.11750*, ***Revision Submitted*** *in Applied Mathematics and Computation* (SCI, Q1, MCQ: 0.61).
6. S. S. Ahmad and **P. Khatun**, "Partial condition numbers for double saddle point problems." *arXiv preprint arXiv:2502.19792*, ***Under Revision*** *in Numerical Algorithms,* (SCI, Q1, MCQ: 1.21).
7. S. S. Ahmad and **P. Khatun**, "Structured backward errors of sparse generalized saddle point problems with Hermitian block matrices." ***Revision Submitted*** *in Electronic Transactions on Numerical Analysis,* (SCI, Q2, MCQ: 0.87).
8. S. S. Ahmad and **P. Khatun**, "Structured backward error analysis for double saddle point problems." *arXiv preprint arXiv:2408.14019*, ***Under Review***.

TABLE OF CONTENTS

LIST OF FIGURES xix

LIST OF TABLES xxi

NOTATION xxv

ACRONYMS xxvii

Chapter 1 **Introduction** 1

1.1 Applications Leading to the Saddle Point Problems 9

1.1.1 Equality-Constrained Quadratic Programming Problems 9

1.1.2 PDE-Constrained Optimization Problems 10

1.1.3 Augmented Systems in Least Squares Problems 10

1.1.4 Discretization of Equations from Physics 11

1.2 Iterative Methods and Preconditioning for Linear Systems 12

1.2.1 Stationary Iterative Methods 13

1.2.2 Krylov Subspace Methods 13

1.2.3 Generalized Minimal Residual (GMRES) Method 14

1.2.4 Preconditioning 15

1.3 Preliminaries 16

Chapter 2 **A Robust Parameterized Enhanced Shift-Splitting Preconditioner for Double Saddle Point Problems** 19

2.1 Background 19

2.2 The Proposed Parameterized Enhanced Shift-Splitting (PESS) Iterative Method and Preconditioner 22

2.3 Convergence Analysis of the PESS Iterative Method 24

2.4 Spectral Distribution of the PESS Preconditioned Matrix 27

2.5 Local PESS (LPESS) Preconditioner 36

2.6 The Strategy of Parameter Selection 38

2.7 Numerical Experiments 39

2.8 Summary 55

Chapter 3 **A Class of Generalized Shift-Splitting Preconditioners for Double Saddle Point Problems** 57

3.1 Background 57

3.2 Solvability Conditions and Properties of the DSPP 59

3.3 Proposed Generalized Shift-Splitting (GSS) Iterative Method and Preconditioner 62

3.4 Convergence Analysis of the GSS Iterative Method 64

3.5 Two Relaxed Variants of GSS Preconditioner 66

3.6 Discussion on the Selection of Parameters 73

3.7 Numerical Experiments 74

3.8 Summary 79

Chapter 4 **Sparsity Preserving Structured Backward Errors for Saddle Point Problems** 81

4.1 Structured Backward Errors for Generalized Saddle Point Problems 81

4.1.1 Preliminaries 83

4.1.2 Structured BEs for Circulant Structured GSPPs 85

4.1.3 Structured BEs for Toeplitz Structured GSPPs 91

4.1.4 Structured BEs Symmetric-Toeplitz Structured GSPPs 94

4.1.5 Unstructured BEs with Preserving Sparsity Pattern 96

4.1.6 Application to Derive the Structured BEs for the WRLS Problems 98

4.2 Structured Backward Errors of Generalized Saddle Point Problems with Hermitian Block Matrices 99

4.2.1 Basic Definitions and Lemmas 99

4.2.2 Computation of Structured BEs 102

4.3 Structured Backward Errors for Double Saddle Point Problems 109

4.3.1 Preliminaries 110

4.3.2 Derivation of Structured BEs for Case (i) 113

4.3.3 Derivation of Structured BEs for Case (ii) 117

4.3.4 Derivation of Structured BEs for Case (iii) 120

4.4 Numerical Experiments 121

4.5 Summary 127

Chapter 5 **Partial Condition Numbers for Saddle Point Problems** 129

5.1 Partial Condition Numbers for the Generalized Saddle Point Problem 129
5.1.1 Preliminaries 131
5.1.2 Partial CNs for the GSPP when $B = C$ 133
5.1.3 Structured Partial CNs when A is Symmetric and $B = C$ is Toeplitz 136
5.1.4 Structured Partial CNs when A and D have Linear Structures 143
5.1.5 Application to WRLS Problems 150
5.1.6 Numerical Experiments 153
5.1.7 Summary 157
5.2 Partial Condition Numbers for Double Saddle Point Problems 157
5.2.1 Background 158
5.2.2 Preliminaries 159
5.2.3 Partial Unified CNs for the DSPP 161
5.2.4 Structured Partial CNs 166
5.2.5 Deduction of Partial CNs for the EILS Problem 170
5.2.6 Numerical Experiments 172
5.2.7 Summary 176

Chapter 6 **Condition Numbers for Moore-Penrose Inverse and Least Square Problem** 177
6.1 Background 177
6.2 Preliminaries 180
6.3 MCN and CCN for General Parameterized Matrices 181
6.3.1 M-P Inverse of General Parameterized Matrices 181
6.3.2 MNLS Solution for General Parameterized Coefficient Matrices 186
6.4 CNs for Cauchy-Vandermonde (CV) Matrices 189
6.5 Quasiseparable (QS) Matrices 194
6.5.1 CNs Corresponding to QS Representation 195
6.5.2 CNs Corresponding to GV Representation 199
6.5.3 Comparisons Between Different CNs for $\{1, 1\}$-QS Matrices 202
6.5.4 The Structured Effective CNs 204
6.6 Numerical Experiments 206
6.7 Summary 209

Chapter 7 **Conclusions and Scope for Future Work** 211

REFERENCES 213

LIST OF FIGURES

2.7.1 Convergence curves for IT versus RES of PGMRES methods employing BD, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS and LPESS ($s = 12$) preconditioners in Case II for Example 2.7.1. 44

2.7.2 Spectral distributions of $\mathcal{A}$, $\mathscr{P}_{BD}^{-1}\mathcal{A}$, $\mathscr{P}_{IBD}^{-1}\mathcal{A}$, $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$, $\mathscr{P}_{SL}^{-1}\mathcal{A}$, $\mathscr{P}_{SS}^{-1}\mathcal{A}$, $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{RPGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $l = 16$ for Example 2.7.1. 45

2.7.3 Spectral bounds for $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $l = 16$ for Example 2.7.1. 46

2.7.4 Parameter s vs CN of the preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ for Case II with $l = 32$ for Example 2.7.1. 47

2.7.5 Relationship of norm error of solution with increasing noise percentage, employing proposed $\mathscr{P}_{PESS}$ for Case II with $s = 12$ for $l = 16, 32, 48, 64$ and 80 for Example 2.7.1. 48

2.7.6 Convergence curves for IT versus RES of the PGMRES methods by employing IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS and LPESS preconditioners in Case II for Example 2.7.2. 51

2.7.7 Spectral distributions of $\mathcal{A}$, $\mathscr{P}_{IBD}^{-1}\mathcal{A}$, $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$, $\mathscr{P}_{SL}^{-1}\mathcal{A}$, $\mathscr{P}_{SS}^{-1}\mathcal{A}$, $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{RPGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $\mathbf{h} = 1/8$. 52

2.7.8 Spectral bounds for $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $\mathbf{h} = 1/8$ for Example 2.7.2. 53

2.7.9 Characteristic curves for IT of the proposed PESS (left) and LPESS (right) PGMRES methods by varying s in the interval $[1, 100]$ with step size 1 with $\mathbf{h} = 1/16$ in Case I for Example 2.7.2. 54

3.7.1 Eigenvalue distributions of $\mathfrak{B}$, $\mathscr{P}_{\text{BD}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{DS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{SS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ for def_set.pow $= 5$ with $\boldsymbol{\beta} = 0.1$ for Example 3.7.1. 77

3.7.2 Convergence curves of the GSS and RGSS-II RGMRES methods varying the parameters α, β, τ, $\boldsymbol{\omega}$ for Example 3.7.1 with $\boldsymbol{\beta} = 0.1$. 78

3.7.3 Relationship between CNs of the preconditioned matrices $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ varying the parameter $\boldsymbol{\omega}$ in $[1, 100]$ with $\boldsymbol{\beta} = 0.1$ for Example 3.7.1. 78

4.4.1 Different structured and unstructured BEs for $n = 8 : 4 : 100$. 123

LIST OF TABLES

2.2.1 $\mathscr{P}_{PESS}$ as a generalization of the above SS preconditioners for different choices of $\Lambda_1, \Lambda_2, \Lambda_3$ and s. 23

2.7.1 Numerical results of GMRES, BD, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS and LPESS PGMRES methods for Example 2.7.1. 41

2.7.2 Numerical results of PESS-I, LPESS-I, PESS-II and LPESS-II PGMRES methods for Example 2.7.1. 43

2.7.3 Spectral bounds for $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for case II with $l = 16$ for Example 2.7.1. 46

2.7.4 Numerical results of GMRES, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS, and LPESS PGMRES methods for Example 2.7.2. 49

2.7.5 Numerical results of PESS-I, LPESS-I, PESS-II and LPESS-II PGMRES methods for Example 2.7.2. 51

3.7.1 Experimental results of GMRES, BD, BD-MINRES, BP, DS, RDF, SS, GSS, RGSS-I and RGSS-II PGMRES methods for Example 3.7.1 when $\boldsymbol{\nu} = 0.1$ 75

3.7.2 Experimental results of GMRES, BD, BD-MINRES, BP, DS, RDF, SS, GSS, RGSS-I and RGSS-II PGMRES methods for Example 3.7.1 when $\boldsymbol{\nu} = 0.001$ 76

4.4.1 Unstructured and structured BEs for different values of n for Example 4.4.4. 124

4.4.2 Values of structured and unstructured BEs of the approximate solution obtained using GMRES for Example 4.4.7. 127

5.1.1 Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = I_{2n}$ for Example 5.1.1. 155

5.1.2 Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}$ for Example 5.1.1. 155

5.1.3 Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = \begin{bmatrix} \mathbf{0} & I_n \end{bmatrix}$ for Example 5.1.1. 155

5.1.4 Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = I_{m+n}$ for Example 5.1.2. 156

5.2.1 Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_0$ for Example 5.2.1. 174

5.2.2 Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_n$ for Example 5.2.1. 174

5.2.3 Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_m$ for Example 5.2.1. 174

5.2.4 Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_p$ for Example 5.2.1. 175

5.2.5 Comparison of the partial NCN, MCN, and CCN with their structured counterparts for Example 5.2.2. 176

6.6.1 Comparison between upper bounds of structured and unstructured CNs for $M^{\dagger}(\Psi_{\mathbb{CV}})$ and $M^{\dagger}(\Psi_{\mathbb{CV}}, b)$ for Example 6.6.1. 206

6.6.2 Comparison between the upper bounds of unstructured, structured CNs and structured effective CNs for the M-P inverse and the MNLS solution of $\{1,1\}$-QS matrices for Example 6.6.2. 207

6.6.3 Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the M-P inverse of $\{1,1\}$-QS matrices for Example 6.6.3. 207

6.6.4 Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the MNLS solution for $\{1,1\}$-QS matrices for Example 6.6.3. 208

6.6.5 Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the M-P inverse of $\{1,1\}$-QS matrices for Example 6.6.4. 208

6.6.6 Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the MNLS solution for $\{1, 1\}$-QS matrices for Example 6.6.4. 208

NOTATION

Symbols

$\mathbb{R}$	The set of real numbers
$\mathbb{C}$	The set of complex numbers
$\mathbb{R}^{m\times n}$	The set of real matrices of size $m \times n$
$\mathbb{C}^{m\times n}$	The set of complex matrices of size $m \times n$
$\mathbb{HC}^{n\times n}$	The set of all $n \times n$ Hermitian matrices
$\mathcal{S}_n$	The set of all real $n \times n$ symmetric matrices
$\mathbb{SKR}^{n\times n}$	The set of all real $n \times n$ skew-symmetric matrices
$\mathcal{C}_n$	The set of all $n \times n$ circulant matrices
$\mathcal{T}_{m\times n}$	The set of all $m \times n$ Toeplitz matrices
$\mathcal{ST}_n$	The set of all $n \times n$ symmetric-Toeplitz matrices
A^T	The transpose of $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
A^H	The conjugate transpose of $A \in \mathbb{C}^{m\times n}$
A^{-1}	The inverse of $A \in \mathbb{F}^{n\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$A^\dagger$	The Moore-Penrose inverse of $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
I_n	The identity matrix of order n
I	The identity matrix of appropriate dimension
$\mathbf{0}_{m\times n}$	The zeros matrix of size $m \times n$
$\mathbf{0}$	The zero matrix of appropriate dimension
e_i^n	The i^{th} column of I_n
$\boldsymbol{i}$	The imaginary unit
$\mathbf{1}_{m\times n}$	The matrix of size $m \times n$ with all entries are equal to 1
$\mathbf{1}_m$	The vector of size m with all entries are equal to 1
$A \otimes B$	Kronecker product of matrices A and B
$A \odot B$	Hadamard product of matrices A and B
$\lvert A\rvert := [\lvert a_{ij}\rvert]$	The absolute matrix of $A \in \mathbb{R}^{m\times n}$
$\lvert X\rvert \ \leq \lvert Y\rvert$	$\lvert x_{ij}\rvert \ \leq \lvert y_{ij}\rvert$ for $X = [x_{ij}]$, $Y = [y_{ij}] \in \mathbb{R}^{m\times n}$ and $1 \leq i \leq m$ and $1 \leq j \leq n$

$\text{rank}(A)$	Rank of matrix $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\texttt{tr}(A)$	Trace of the matrix $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\sigma(A)$	Spectrum of matrix $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\lambda_{\max}$	Maximum eigenvalue of any matrix $A \in \mathbb{R}^{m\times n}$ with $\sigma(A)$ real
$\lambda_{\min}$	Minimum eigenvalue of any matrix $A \in \mathbb{R}^{m\times n}$ with $\sigma(A)$ real
$\vartheta(A)$	The spectral radius of $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\|x\|_2 := \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$	The 2-norm of $x \in \mathbb{F}^n$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\|x\|_\infty := \max_i \|x_i\|$	The infinity norm of $x \in \mathbb{F}^n$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\|A\|_{\max} := \max_{i,j} \|a_{ij}\|$	The max norm of $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\|A\|_F := \sqrt{\text{tr}(A^H A)}$	The Frobenius norm of $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\|A\|_2 := \max_{\|x\|_2=1} \|Ax\|_2$	The spectral norm of $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} \|a_{ij}\|$	The infinity norm of $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$
$\mathfrak{R}(A)$	The real part of $A \in \mathbb{C}^{m\times n}$
$\mathfrak{I}(A)$	The imaginary part of $A \in \mathbb{C}^{m\times n}$
$A \succ (\succeq)\, 0$	A is symmetric positive (semi)definite matrix
$A \succ B$ $(A \succeq B)$	$A - B \succ \mathbf{0}$ $(A - B \succeq \mathbf{0})$
$\text{diag}(A_1, \ldots, A_p)$	The block diagonal matrix with diagonal blocks $A_1, \ldots, A_p$
$A(i,:)$	Returns i^{th} row of the matrix A
$A(:,j)$	Returns j^{th} column of the matrix A
$\texttt{randn}(m,n)$	Returns an m-by-n normally distributed random matrix
$\texttt{sprandn}(m,n,\mu)$	Returns a normally distributed $m \times n$ sparse random matrix with μmn nonzero entries
$\texttt{sprand}(m,n,\mu)$	Creates a uniformly distributed $m \times n$ sparse random matrix with μmn nonzero entries
$\texttt{toeplitz}(x,y)$	Returns a Toeplitz matrix with x as its first column and y as its first row, where $x \in \mathbb{C}^m$ and $y \in \mathbb{C}^n$
$\texttt{tridiag}(a,b,c)$	The tridiagonal matrix with diagonal entries b, subdiagonal entries a, and superdiagonal entries c

ACRONYMS

SPP	Saddle point problem
GSPP	Generalized saddle point problem
DSPP	Double saddle point problem
SPD	Symmetric positive definite
BD	Block diagonal
IBD	Inexact block diagonal
SS	Shift-splitting
RSS	Relaxed shift-splitting
EGSS	Extensive generalized shift-splitting
PESS	Parameterized enhance shift-splitting
LPESS	Local parameterized enhance shift-splitting
GSS	Generalized shift-splitting
RGSS	Relaxed generalized shift-splitting
APSS	Alternating positive semidefinite splitting
WRLS	Weighted regularized least squares
EILS	Equality-constrained indefinite least squares
BE	Backward errror
M-P	Moore-Penrose
LS	Least squares
CN	Condition number
NCN	Normwise condition number
MCN	Mixed condition number
CCN	Componentwise condition number
(P) GMRES	(Preconditioned) Generalized minimum residual method
PDE	Partial differential equation
CV	Cauchy-Vandermonde
QS	Quasiseparable

CHAPTER 1

Introduction

Numerical approximation methods serve as a cornerstone in science and engineering, translating complex real-world problems into systems of linear equations. Consequently, the ability to efficiently solve these systems serves as a linchpin of computational science, driving advancements in fields ranging from physics and engineering to machine learning and finance. A system of linear equations is generally defined as finding a solution x for a given matrix $A \in \mathbb{C}^{m\times n}$ and a vector $b \in \mathbb{C}^m$, such that

$$Ax = b.$$

Of particular significance are cases where m and n are large, and A exhibits sparsity—a structure that arises naturally in many real-world problems.

Linear systems frequently arise from differential equations and optimization problems involving infinite degrees of freedom. Through appropriate discretization methods, these continuous problems are transformed into algebraic systems with a finite number of degrees of freedom. To construct realistic numerical models, these discrete systems often involve millions of variables, making their solution computationally demanding. A substantial portion of simulation time is dedicated to solving such large-scale linear systems. The associated coefficient matrices are typically sparse, containing a small fraction of nonzero elements, often exhibiting structured patterns. Efficient numerical linear algebra algorithms are crucial for solving these systems with minimal computational cost and memory usage. As scientific computing advances, the demand for robust solvers grows, driving progress in numerical analysis, large-scale simulations, and data-driven applications.

Linear systems in saddle point form have received significant attention owing to their extensive applications in partial differential equation (PDE)-constrained optimization problems [115], computational fluid dynamics [34, 59], least squares estimation problems [29, 30], liquid crystal director models [111], optimal control [110], Maxwell's equations [45], and so on. These systems often manifest in diverse forms, with their coefficient matrices typically exhibiting two-by-two or three-by-three block structures and referred to as saddle point problems (SPPs).

In the following, we explore two important classes of SPPs that have garnered considerable attention in the literature.

Generalized saddle point problem (GSPP): The GSPP is represented by a two-by-two block linear system of the form:

$$\mathcal{M}\boldsymbol{v} := \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B}^T \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{g} \end{bmatrix} =: \mathbf{b}, \tag{1.0.1}$$

$\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{C}^{m\times n}$, $\boldsymbol{D} \in \mathbb{C}^{m\times m}$; $\boldsymbol{f} \in \mathbb{C}^n$ and $\boldsymbol{g} \in \mathbb{C}^m$ are the known vectors; $\boldsymbol{u} \in \mathbb{C}^n$ and $\boldsymbol{p} \in \mathbb{C}^m$ are the solution vectors. The block matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$ satisfy some special structures, such as $\boldsymbol{B} = \boldsymbol{C}$, symmetric, Toeplitz, or have some other linear structures [11, 26]. The GSPP in (1.0.1) encompasses several important cases: the standard SPP ($\boldsymbol{A} = \boldsymbol{A}^T$, $\boldsymbol{B} = \boldsymbol{C}$, $\boldsymbol{D} = \boldsymbol{D}^T$) and the real GSPP ($\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B}, \boldsymbol{C} \in \mathbb{R}^{m\times n}$, $\boldsymbol{D} \in \mathbb{R}^{m\times m}$, $\boldsymbol{f} \in \mathbb{R}^n$, and $\boldsymbol{g} \in \mathbb{R}^m$) [26].

The GSPP or its special cases originate from a wide range of applications. For example: (i) The Karush-Kuhn-Tucker (KKT) system ($\boldsymbol{A} = \boldsymbol{A}^T$, $\boldsymbol{B} = \boldsymbol{C}$, and $\boldsymbol{D} = \mathbf{0}$) is one of the simplest versions of (1.0.1) and arises from the KKT first-order optimality condition [28] in constrained optimization problems [26, 129] given by

$$\min_{\boldsymbol{u}} \frac{1}{2}\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} - \boldsymbol{f}^T\boldsymbol{u}$$
$$\text{subject to} \quad \boldsymbol{B}\boldsymbol{u} = \boldsymbol{g}.$$

(ii) The system (1.0.1) also comes from the finite element discretization of time-harmonic eddy current models [9]. (iii) GSPPs emerge in the weighted regularized least squares (WRLS) problem [24] arising from image restoration and reconstruction problems [106].

Double saddle point problem (DSPP): A general form of the DSPP, also known as three-by-three block saddle point problem, can be written as:

$$\mathfrak{B}\boldsymbol{w} := \begin{bmatrix} A & B^T & \mathbf{0} \\ F & -D & C^T \\ \mathbf{0} & G & E \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} = \begin{bmatrix} f \\ g \\ h \end{bmatrix} =: \mathbf{d}, \tag{1.0.2}$$

where $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$, $E \in \mathbb{R}^{p\times p}$, B, $F \in \mathbb{R}^{m\times n}$, and C, $G \in \mathbb{R}^{p\times m}$. The vectors $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$, and $\boldsymbol{z} \in \mathbb{R}^p$ are unknown, while $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$, and $h \in \mathbb{R}^p$ are known vectors. Let $l = n + m + p$ and we refer $\mathfrak{B}$ as the double saddle point matrix.

In many applications, DSPPs commonly arise with $B = F$ or $C = G$. Additionally, a special case of the DSPP frequently appears with $D = \mathbf{0}$ or $E = \mathbf{0}$, referred to as the unregularized form, in various contexts.

The structure and properties of SPPs in (1.0.1) and (1.0.2) make them a critical focus in numerical linear algebra, particularly for developing efficient, robust, and scalable solution methods tailored to their unique characteristics. However, a key challenge arises from the indefinite nature or poor spectral properties of the coefficient matrices $\mathcal{M}$ and $\mathfrak{B}$, which significantly complicates the task of numerically solving the systems (1.0.1) and (1.0.2). Furthermore, the coefficient matrices are generally sparse and large; therefore, iterative methods become superior to direct methods. The Krylov subspace method is preferable among other iterative methods investigated widely in the literature due to their minimal storage requirement and feasible implementation [139]. Slow convergence of the Krylov subspace methods is a major drawback for a system of linear equations of large dimensions. Moreover, the saddle point matrices $\mathcal{M}$ and $\mathfrak{B}$ can be very sensitive, which leads to a significant slowdown in the solution algorithm. Therefore, it is important to develop novel, efficient, and robust preconditioners for the fast convergence of the Krylov subspace method, which can handle the sensitivity of the problem (1.0.1).

Iterative methods for the classical GSPP are well-established, and over the years, considerable attention has been devoted to developing its numerical solution techniques. These include Uzawa methods [12, 158], Hermitian and skew-Hermitian splitting (HSS)-type methods [10, 13], successive overrelaxation (SOR)-type methods [15, 64], null space methods [65, 126], and shift-splitting (SS)-type strategies [37, 40, 124], etc. For a comprehensive survey of applications, algebraic properties, and iteration methods for GSPPs, we refer to [11] and reference therein.

The DSPP (1.0.2) can be converted into the GSPP (1.0.1) by considering the following partitions:

$$\boldsymbol{A} = \begin{bmatrix} A & B^T \\ F & -D \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} \mathbf{0} & C \end{bmatrix}, \ \boldsymbol{C} = \begin{bmatrix} \mathbf{0} & G \end{bmatrix} \text{ and } \boldsymbol{D} = E, \tag{1.0.3}$$

or

$$\boldsymbol{A} = A, \ \boldsymbol{B} = \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix}, \ \boldsymbol{C} = \begin{bmatrix} F \\ \mathbf{0} \end{bmatrix} \text{ and } \boldsymbol{D} = \begin{bmatrix} -D & C^T \\ G & E \end{bmatrix}. \tag{1.0.4}$$

The first partitioning reveals that (1.0.2) possesses a double saddle point structure, as the (1,1) block itself represents the coefficient matrix of an SPP. Consequently, the system in (1.0.2) is also referred to as a DSPP. On the other hand, the second partitioning demonstrates the same with (2,2) block having saddle point structure. However, the properties of the submatrices in (1.0.3) and (1.0.4) are different from the standard two-by-two block SPP. Notice that the leading block is not symmetric positive definite (SPD)

in the first partitioning. The second partitioning highlights that the $(1,2)$ block is rank deficient and the $(2,2)$ block is indefinite. In contrast, in the standard SPP, $(1,2)$ block is full row rank, and the $(2,2)$ block is generally symmetric positive semidefinite. Therefore, the existing literature for solving the DSPPs (see [26]) may not be applied to solve (1.0.2) directly and it is essential to develop new preconditioners for DSPPs that exploit the specific structure of the double saddle point matrix $\mathfrak{B}$, which is sensitive in nature.

The invertibility conditions for the coefficient matrix in (1.0.2) with $F = B$ and $G = C$ are analyzed in [20]. Furthermore, bounds on the eigenvalues of the double saddle point matrix $\mathfrak{B}$ are discussed in [33]. Recently, several iterative methods and preconditioning techniques have been developed in recent times for solving DSPPs. Block diagonal (BD) and inexact BD (IBD) preconditioners have been explored for various forms of the DSPP (1.0.2) in [21, 33, 75]. Additionally, iterative methods and preconditioners based on alternating positive semidefinite splitting (APSS) have been studied in [43, 125]. Moreover, SOR-type and Uzawa-type methods have been investigated in [74, 77, 76].

Recently, various studies have been done for SS-type preconditioner for the DSPP (1.0.2) with $F = B, G = C$, $D = \mathbf{0}$, and $E = \mathbf{0}$; see [37, 92, 134, 156]. However, these preconditioners have notable drawbacks, including inefficiency in performance. They lack the ability to outperform state-of-the-art preconditioners, such as the widely used BD-type preconditioners, and need further improvements. More importantly, SS-type preconditioners remain largely unexplored for cases where D and E are nonzero or when the diagonal block matrices are non-symmetric. Furthermore, the spectral distribution of SS-type preconditioners has not been thoroughly studied. This thesis aims to bridge these gaps, enhancing the effectiveness of SS-type preconditioners and unlocking their full potential in solving DSPP efficiently.

When applying numerical or iterative methods to solve a problem, machine round-off errors and truncation errors inevitably result in approximate solutions rather than exact ones. This approximate nature raises several critical questions: Are these computed solutions reliable? Are the numerical algorithms stable which are used to compute the solution? For which problem the obtained approximate solution is exact? How does a small change in the input data affect the output of the problem? Addressing these questions is crucial, as neglecting them could lead to results that are meaningless or irrelevant to the original problem.

In the realm of numerical analysis, perturbation theory is extensively used to examine the quality of the computed solution using some numerical methods [73]. The concepts of

condition number (CN) and backward error (BE) are the two most important tools since they are widely employed in assessing the sensitivity and stability of an approximate solution; see [73]. The CN measures how sensitive, in the worst-case scenario, a problem is to a slight change in input data. On the other hand, the BE is used to find a nearly perturbed problem with minimal magnitude perturbations so that the approximate solution becomes an actual solution of the perturbed problem. The minimal distance between the original and perturbed problem is referred to as the BE. We can estimate the forward error (difference between a computed solution and the exact solution) of an approximate solution by combining the BE with the CN.

Most likely, for the first time, Rice [116] introduced the classical theory of CNs. In accordance with [116], the normwise condition number (NCN), which has been extensively considered in the literature, quantifies the input and output data errors using the norms. A notable drawback associated with NCN lies in its inability to capture the inherent structure of badly scaled or sparse input data. Consequently, even minor relative normwise perturbations can have a disproportionate impact on entries that are small or zero, thereby potentially compromising data sparsity. Consequently, the NCN occasionally overestimates the true conditioning of the numerical solution. To address this challenge, mixed condition number (MCN) and componentwise condition number (CCN) have seen a growing interest in the literature [63, 118, 127]. The MCN measures the input perturbations componentwise and the output error using norms, whereas the CCN measures both the error in output data and the input perturbations componentwisely. The BE analysis proposed by Wilkinson [143], has several key applications: for example, BEs are often employed as a stopping criterion for iterative algorithms when solving a problem. For a given problem, if the computed BE of an approximate solution is within the unit round-off error, then the corresponding numerical algorithm is considered backward stable [73].

In many applications, the coefficient matrix blocks of systems (1.0.1) and (1.0.2) exhibit linear structures, such as symmetric, Toeplitz, or symmetric-Toeplitz patterns [32, 60, 124, 163]. This naturally raises an important question: How sensitive is the solution when structure-preserving perturbations are applied to the coefficient matrix of GSPP or DSPP? Moreover, a numerical algorithm is considered strongly backward stable if the computed solution corresponds exactly to the solution of a nearby structure-preserving problem [35, 36]. This leads to a fundamental inquiry: Does a backward stable algorithm for solving (4.3.1) also exhibit strong backward stability? To address this question, the notions of structured CN and BE are introduced specifically for problems

with special structural properties, where both CN and BE are analyzed under structure-preserving constraints imposed on the perturbation matrices.

Higham and Higham [72] explored the CN and BE analysis for approximate solutions of linear systems involving both structured and unstructured matrices. Recently, the investigation of structured BEs and CNs in SPPs has shown significant development. For GSPPs, CNs and perturbation bounds have been examined in [100, 138, 147, 151, 153], focusing on the solution $\boldsymbol{v} = [\boldsymbol{u}^T, \boldsymbol{p}^T]^T$ and its individual components $\boldsymbol{u}$ and $\boldsymbol{p}$. However, the sensitivity of each solution component can vary; hence, analyzing the sensitivity of the individual solution components of $\boldsymbol{v}$ in (1.0.1) or $\boldsymbol{w}$ in (1.0.2) becomes crucial [38]. The traditional CNs lack the ability to reveal the conditioning of a specific part of the solution. Thus, it is crucial to evaluate their individual conditioning properties. Moreover, both structured and unstructured CNs for DSPPs remain largely unexplored in the literature, presenting an avenue for further research. To tackle this situation, in this thesis, we investigate the structured NCN, MCN, and CCN of a linear function of the solution by introducing a matrix L of the GSPP and DSPP. This kind of CN is referred to as partial CN. Different choices of L provide the flexibility to determine the CNs of various solution components of $\boldsymbol{w}$. For example, by selecting $\mathrm{L} = I_l$, $[I_n \;\; \mathbf{0}_{n\times(l-n)}]$, or $[\mathbf{0} \;\; I_m \;\; \mathbf{0}_{m\times p}]$, we can determine the CN of $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$, $\boldsymbol{x}$, or $\boldsymbol{y}$, respectively.

On the other hand, many research has been conducted on both unstructured and structured BE analysis for GSPP and DSPP [44, 129, 146, 162]. However, these studies often fail to preserve for the inherent sparsity patterns of the coefficient matrix in SPPs. Moreover, the existing techniques are not applicable when the block matrices have circulant, Toeplitz, or symmetric-Toeplitz structures and do not even provide structure-preserving minimal perturbation matrices for which the BE is attained. Furthermore, due to the special block structure of the double saddle point matrix, these investigations do not provide exact structured BEs for the DSPP (1.0.2). Recently, structured BEs for DSPP have been investigated in [95, 96, 98]. However, these studies do not preserve the symmetric structures and sparsity pattern of the block matrices. This thesis explores the structured BEs for GSPPs and DSPPs, focusing on preserving the sparsity pattern and the inherent block structure of the coefficient matrices in the perturbation matrices.

The above-discussed linear systems consider the coefficient matrix as nonsingular. Next, consider the problem of finding a solution x for the system of linear equations: $Ax = b$, $A \in \mathbb{C}^{m\times n}$, $b \notin \mathcal{R}(A)$; i.e., the system is inconsistent. In such cases, the

objective is to determine $\mathfrak{X}$ that minimizes the residual in the least squares (LS) sense:

$$\min_{\mathfrak{X}\in\mathbb{C}^n} \|A\mathfrak{X} - b\|_2. \tag{1}$$

This formulation seeks the best approximate solution and is known as the LS problem. The unique minimum norm least square (MNLS) solution to the LS problem is given by $A^\dagger b$, where $A^\dagger$ denotes the Moore-Penrose (M-P) inverse of A.

Over the years, several generalizations of LS problems have been studied to address challenges such as ill-conditioning or rank deficiency of the matrix A. Techniques like Tikhonov regularization, WRLS, and indefinite least squares (ILS) have been developed to stabilize solutions. These generalizations achieve stabilization by introducing regularization terms or weighting matrices, effectively mitigating issues related to ill-posedness and numerical instability. Recently, structured CNs of these problems, including the M-P inverse, LS problem, WRLS problem, and Tikhonov regularization, have garnered significant attention [50, 51, 152]. However, structured CNs for LS problems, where A is rank-deficient and possesses special structures, such as low-rank patterns, remain largely unexplored. Thus, the development of structured BEs for these problems requires further refinement. In this thesis, we address the problem of structured CNs for the M-P inverse and LS problems associated with rank-deficient matrices. Furthermore, leveraging the intrinsic connection between LS problems and SPPs, we extend our developed framework to analyze structured BEs and CNs for these classes of LS problems.

The outline of the of the thesis is as follows. In the remaining part of this chapter, we discuss some applications that lead to GSPP and DSPP, key conceptual ideas that are used throughout the thesis, and essential preliminaries. Additionally, an overview of stationary iterative methods, Krylov subspace methods, and preconditioners are provided.

In **Chapter 2**, we propose the parameterized enhanced shift-splitting (PESS) iterative method and preconditioner for solving the equivalent nonsymmetric DSPP (1.0.2) with $F = B$, $G = C$, $D = \mathbf{0}$, and $E = \mathbf{0}$, formulated as:

$$\mathcal{A}\mathbf{u} := \begin{bmatrix} A & B^T & \mathbf{0} \\ -B & \mathbf{0} & -C^T \\ \mathbf{0} & C & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} = \begin{bmatrix} f \\ -g \\ h \end{bmatrix} =: \widehat{\mathbf{d}}. \tag{1.0.5}$$

The coefficient matrix $\mathcal{A}$ exhibits key properties essential for the convergence of the SS-type iterative method. Specifically, $\mathcal{A}$ is semipositive real, satisfying $\mathbf{u}^T\mathcal{A}\mathbf{u} \geq 0$ for all $\mathbf{u} \in \mathbb{R}^l$, and positive semistable, with all eigenvalues having nonnegative real parts. We also propose the local PESS (LPESS) preconditioner, a localized variant of the PESS

preconditioner, enhanced with a relaxation mechanism. This chapter provides a comprehensive convergence analysis of the PESS iterative method and a spectral study of the PESS and LPESS preconditioned matrices, including detailed spectral bounds to evaluate their effectiveness.

In **Chapter 3**, building on the framework from Chapter 2, we proposed generalized shift-splitting (GSS) preconditioners for the DSPP when $F = B$, $G = C$, and $D = \mathbf{0}$, in the following form obtained by reordering (1.0.2):

$$\mathfrak{B}\widehat{\boldsymbol{w}} := \begin{bmatrix} A & \mathbf{0} & B^T \\ \mathbf{0} & E & C \\ -B & -C^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{z} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} f \\ h \\ g \end{bmatrix}. \tag{1.0.6}$$

This formulation arises in applications such as PDE-constrained optimization problems [115]. The GSS preconditioner accommodates both symmetric and nonsymmetric A and E. To improve its efficiency, we introduced two relaxed versions, along with detailed spectral analyses of the preconditioned matrices.

In **Chapter 4**, we analyze structured BEs for the GSPP (1.0.1) and DSPP (1.0.2) under perturbations that respect sparsity and specific matrix structures, including symmetric, Hermitian, circulant, Toeplitz, and symmetric-Toeplitz. The minimal perturbation matrices are constructed to preserve these structural properties. The developed framework is applied to compute BEs for WRLS problems, supported by numerical experiments that validate the findings and highlight their effectiveness in evaluating the strong backward stability of numerical algorithms.

In **Chapter 5**, we explore unstructured and structured partial NCN, MCN, and CCN for the GSPP (1.0.1) and the DSPP (1.0.2). A general framework is developed to measure the structured CNs of individual components of the solution of the GSPP by preserving the symmetric, Toeplitz, and more general linear structures of block matrices. Furthermore, we introduce the concept of unified partial CNs for the DSPP, which incorporates the traditional NCN, MCN, and CCN into a single framework. Sharp upper bounds for partial CNs are provided, which are free from expensive Kronecker products. Applications of the derived frameworks include:

- To derive structured CNs for the WRLS and Tikhonov regularization. These studies also retrieve some previous studies in the literature [101].
- By leveraging the relationship between DSPP and equality-constrained indefinite least squares (EILS) problems, we derive partial CNs for the EILS problem.

In **Chapter 6**, we analyze the structured MCN and CCN for the M-P inverse and MNLS solutions of rank-structured matrices, including Cauchy-Vandermonde (CV) and $\{1,1\}$-quasiseparable (QS) matrices. A general framework is developed to efficiently compute upper bounds for MCN and CCN of rank-deficient parameterized matrices, enabling faster computation to estimate the structured CNs for CV and $\{1,1\}$-QS matrices.

In **Chapter 7**, we provide a summary of the thesis and a few potential directions for future research.

This thesis aims to address existing research gaps by achieving these objectives. By doing so, it contributes to advancing knowledge in the field.

1.1. Applications Leading to the Saddle Point Problems

This section explores various applications that give rise to the GSPP and DSPP. These problems commonly emerge in areas such as fluid dynamics, PDE-constrained optimization, and finite element discretizations of PDEs.

1.1.1. Equality-Constrained Quadratic Programming Problems

One of the main application areas leading to SPPs is the equality-constrained convex quadratic programming problems (EQPPs). Consider the following EQPP is defined by:

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^n,\,\boldsymbol{z}\in\mathbb{R}^p} \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + r^T\boldsymbol{x} + q^T\boldsymbol{z} \\ &\text{subject to } B\boldsymbol{x} + C^T\boldsymbol{z} = b, \end{aligned}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}, C \in \mathbb{R}^{p\times m}$, $b \in \mathbb{R}^m, r \in \mathbb{R}^n$, and $q \in \mathbb{R}^p$. Let $\boldsymbol{\lambda} \in \mathbb{R}^m$ be the Lagrange multipliers, then the KKT [28] conditions applied to the following Lagrangian:

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + r^T\boldsymbol{x} + q^T\boldsymbol{z} + \boldsymbol{\lambda}^T(B\boldsymbol{x} + C^T\boldsymbol{z} - b),$$

yield:

$$\begin{aligned} \nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda}) &= A\boldsymbol{x} + B^T\boldsymbol{\lambda} + r = \mathbf{0}, \\ \nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda}) &= q + C\boldsymbol{\lambda} = \mathbf{0}, \\ \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda}) &= B\boldsymbol{x} + C^T\boldsymbol{z} - b = \mathbf{0}. \end{aligned}$$

The symbol $\nabla_{\boldsymbol{x}}$ represents the gradient operator with respect to the variable $\boldsymbol{x}$. The above equation leads to the following DSPP:

$$\begin{bmatrix} A & B^T & 0 \\ B & 0 & C^T \\ 0 & C & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \\ \boldsymbol{z} \end{bmatrix} = \begin{bmatrix} -r \\ b \\ -q \end{bmatrix}.$$

1.1.2. PDE-Constrained Optimization Problems

The Poisson control problem: Consider the distributed control problem defined by:

$$\begin{aligned} &\min_{u,f} \ \tfrac{1}{2}\|u-\hat{u}\|^2_{L_2(\Omega)} + \tfrac{\boldsymbol{\nu}}{2}\|f\|^2_{L_2(\Omega)} \\ \text{subject to} \quad &-\Delta u = f \text{ in } \Omega, \\ &u = g \text{ on } \partial\Omega, \end{aligned} \tag{1.1.1}$$

where u is the state, and $\hat{u}$ is the desired state, f is the control, $\Omega = [0,1]\times[0,1]$ is the domain with the boundary $\partial\Omega$, Δ denotes the Laplacian operator in $\mathbb{R}^d$, and $0 < \boldsymbol{\nu} \ll 1$ is the regularization parameter. By discretizing (1.1.1) using the Galerkin finite element method and then applying Lagrange multiplier techniques, we obtain the following system:

$$\begin{bmatrix} \boldsymbol{\beta} M & \mathbf{0} & K^T \\ \mathbf{0} & M & -M \\ -K & M & \mathbf{0} \end{bmatrix} \begin{bmatrix} u \\ f \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ b \\ -d \end{bmatrix}, \tag{1.1.2}$$

where $M \in \mathbb{R}^{n\times n}$ and $K \in \mathbb{R}^{n\times n}$ are SPD mass matrix and discrete Laplacian, respectively. Note that by setting $A = \boldsymbol{\nu} M$, $B = K$, $C = -M$, and $\mathbf{d} = [\mathbf{0},\, b^T,\, -d^T]^T$, (1.1.2) can be expressed in the form of the DSPP (1.0.2), where $n = m = p$.

1.1.3. Augmented Systems in Least Squares Problems

Weighted and regularized least squares (WRLS) problems: Given $K \in \mathbb{R}^{m\times n}$ $(m \geq n)$, a SPD weight matrix $W \in \mathbb{R}^{n\times n}$ and a vector $f \in \mathbb{R}^n$, we consider the WRLS problem [24] arising from image restoration and reconstruction problems [62, 106] of the form:

$$\min_{y\in\mathbb{R}^n} \|\mathbf{M}y - \tilde{\mathbf{d}}\|_2^2, \tag{1.1.3}$$

where $\mathbf{M} = \begin{bmatrix} W^{1/2}K^T \\ \sqrt{\lambda} I_m \end{bmatrix} \in \mathbb{R}^{(m+n)\times m}$, $\widetilde{\mathbf{d}} = \begin{bmatrix} W^{1/2} f \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m+n}$ and $\lambda > 0$ is the regularization parameter. Then, the minimization problem (1.1.3) can be expressed as the following

augmented linear system:

$$\widehat{\mathcal{M}}\begin{bmatrix}\mathbf{r}\\ y\end{bmatrix} := \begin{bmatrix}W^{-1} & K^T\\ K & -\lambda I_m\end{bmatrix}\begin{bmatrix}\mathbf{r}\\ \boldsymbol{y}\end{bmatrix} = \begin{bmatrix}f\\ \mathbf{0}\end{bmatrix}, \tag{1.1.4}$$

where $\mathbf{r} = W(f - K^T y)$. The equivalent augmented system (1.1.4) possesses the GSPP of the form (1.0.1) where A is symmetric and $B = C$ as a Toeplitz (or symmetric-Toeplitz) matrix.

Equality-constrained indefinite least squares (EILS) problems: The EILS problem is an extension of the famous linear least squares problem, having linear constraints on unknown parameters. It can be expressed as follows:

$$\min_{\boldsymbol{y}\in\mathbb{R}^m} (b - M\boldsymbol{y})^T \mathbb{J}\ (b - M\boldsymbol{y}) \quad \text{subject to} \quad C\boldsymbol{y} = d, \tag{1.1.5}$$

where $M \in \mathbb{R}^{n\times m}(n \geq m)$, $C \in \mathbb{R}^{p\times m}$, $b \in \mathbb{R}^n$, $d \in \mathbb{R}^p$ and the signature matrix $\mathbb{J}$ given by

$$\mathbb{J} = \begin{bmatrix}I_{n_1} & \mathbf{0}\\ \mathbf{0} & -I_{n_2}\end{bmatrix},\ n_1 + n_2 = n. \tag{1.1.6}$$

When $\text{rank}(C) = p$ and $\boldsymbol{y}^T(M^T\mathbb{J}M)\boldsymbol{y} > 0$ for all nonzero $\boldsymbol{y} \in \text{null}(C)$, the EILS problem (1.1.5) has a unique solution. The solution of the EILS (1.1.5) problem also satisfies the following the augmented system [137]:

$$\widehat{\mathfrak{B}}\begin{bmatrix}\lambda\\ \boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} := \begin{bmatrix}\mathbf{0} & \mathbf{0} & C\\ \mathbf{0} & \mathbb{J} & M\\ C^T & M^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\lambda\\ \boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} = \begin{bmatrix}d\\ b\\ \mathbf{0}\end{bmatrix}, \tag{1.1.7}$$

where $\boldsymbol{x} = \mathbb{J}\boldsymbol{r}$, $\boldsymbol{r} = b - M\boldsymbol{y}$ and $\lambda = (CC^T)^{-1}CM^T\mathbb{J}\boldsymbol{r}$ is the vector of Lagrange multipliers [31]. Note that the system in (1.1.7) can be equivalently transformed into

$$\begin{bmatrix}\mathbb{J} & M & \mathbf{0}\\ M^T & \mathbf{0} & C^T\\ \mathbf{0} & C & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \lambda\end{bmatrix} = \begin{bmatrix}b\\ \mathbf{0}\\ d\end{bmatrix} =: \widehat{\mathbf{d}}. \tag{1.1.8}$$

1.1.4. Discretization of Equations from Physics

The Stokes equations: The (steady-state) Stokes problem, which models the flow of a viscous fluid, is represented by the following system of PDEs:

$$\begin{cases} -\nu\Delta\boldsymbol{u} + \nabla\boldsymbol{p} &= F \ \text{ in } \ \Omega,\\ \nabla\cdot\boldsymbol{u} &= 0 \ \text{ in } \Omega,\\ \boldsymbol{u} &= \mathbf{0} \ \text{ on } \partial\Omega, \end{cases} \tag{1.1.9}$$

where $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is the domain, μ is the viscosity, $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ the velocity field and $\boldsymbol{p} : \Omega \to \mathbb{R}$ is the pressure field, and $F : \Omega \to \mathbb{R}^d$ is a given body force. Here, ∇ denotes the gradient and $\nabla\cdot$ is the divergence.

Discretization of the Stokes equation (1.1.9) using finite differences or finite elements results in GSPP of the form (1.0.1), where $\boldsymbol{u}$ represents the discrete velocities and $\boldsymbol{p}$ the discrete pressure.

Maxwell's equations: Discretization of Maxwell's equation in electromagnetics also leads to SPPs. Consider the following time-harmonic Maxwell's equations:

$$\begin{cases} \Delta \times \Delta \times \boldsymbol{u} - k^2 \boldsymbol{u} + \nabla \boldsymbol{p} &= F \ \text{ in } \ \Omega, \\ \nabla \cdot \boldsymbol{u} &= \boldsymbol{0} \ \text{ in } \Omega, \\ \boldsymbol{u} \times \boldsymbol{n} &= \boldsymbol{0} \ \text{ on } \partial\Omega, \\ \boldsymbol{p} &= 0 \ \text{ on } \partial\Omega, \end{cases} \tag{1.1.10}$$

where Ω is a subset of $\mathbb{R}^2$ or $\mathbb{R}^3$ and $\Delta\times$ denotes the curl Discretization using N*édélec* finite elements for $\boldsymbol{u}$ and nodal elements for $\boldsymbol{p}$ results in a linear system of the form:

$$\begin{bmatrix} A - kM & B^T \\ B & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} g \\ \boldsymbol{0} \end{bmatrix},$$

where A is the discrete curl-curl operator, B is the negative discrete divergence operator, M is the finite element mass matrix, and $g \in \mathbb{R}^n$ represents the load vector associated with F.

1.2. Iterative Methods and Preconditioning for Linear Systems

SPPs that arise from real-world applications are typically sparse (containing many zero entries) and are of very large dimensions. Direct methods based on matrix factorizations, such as Gaussian elimination and QR decomposition, perform well for small and medium-sized problems; however, for large problems, they start to struggle. Moreover, these decompositions may introduce a large amount of nonzero entries, which are problematic because of high computational costs.

Iterative methods generate successive approximations that converge to the exact solution. These methods are particularly useful for large systems where direct methods are computationally expensive. Next, we give some overview of commonly used iterative methods.

1.2.1. Stationary Iterative Methods

Consider the linear system $\mathcal{A}\boldsymbol{w} = \mathbf{d}$, where we aim to compute a solution using iterative methods. Stationary iterative methods, among the simplest, are based on splitting the coefficient matrix $\mathcal{A}$ into two matrices such that $\mathcal{A} = M - N$, where M is nonsingular. Then, the iterative method is defined as:

$$M\boldsymbol{w}_{k+1} = N\boldsymbol{w}_k + \mathbf{d}. \tag{1.2.1}$$

The matrix M is chosen based on the specific splitting method employed. For instance, in the Jacobi method, M corresponds to the diagonal part of $\mathcal{A}$, while in the Gauss-Seidel method, M is the lower triangular part of $\mathcal{A}$. Other prominent stationary iterative methods include the SOR and Richardson methods (refer to [122]). The convergence of these types of methods depends upon the spectral radius of the iteration matrix $\mathcal{T} = M^{-1}N$.

Lemma 1.2.1. *[122] Any stationary iterative methods of the form (1.2.1) converges to the unique solution of the linear system $\mathcal{A}\boldsymbol{w} = \mathbf{d}$ for any initial guess $\boldsymbol{w}_0$ if and only if the spectral radius of the iteration matrix $\mathcal{T} = M^{-1}N$ is strictly less than one, i.e., $\vartheta(\mathcal{T}) < 1$.*

The main disadvantages of these types of iterative methods are slow convergence and fixed storage cost at each iteration.

1.2.2. Krylov Subspace Methods

In this subsection, we discuss Krylov subspace methods for solving sparse linear systems of the form $\mathcal{A}\boldsymbol{w} = \mathbf{d}$, where $\mathcal{A}$ is nonsingular matrix and $\mathbf{d}$ is the right hand side vector. The Krylov subspace iterative method is preferable among other iterative methods investigated widely in the literature due to their minimal storage requirement and feasible implementation [139].

Let $\boldsymbol{w}_0$ be an initial guess vector and the initial residual vector defined as $\mathbf{r}_0 = \mathbf{d} - \mathcal{A}\boldsymbol{w}_0$. Then, Krylov subspace methods are the iterative solvers where kth approximate solution satisfies the following:

$$\boldsymbol{w}_k \in \boldsymbol{w}_0 + \mathcal{K}_k(\mathcal{A}, \mathbf{r}_0), \quad k = 1, 2, \ldots, \tag{1.2.2}$$

where $\mathcal{K}_k(\mathcal{A}, \mathbf{r}_0)$ is the k^{th} Krylov subspace generated by $\mathcal{A}$ and $\mathbf{r}_0$ and defined as

$$\mathcal{K}_k(\mathcal{A}, \mathbf{r}_0) := \text{span}\left\{\mathbf{r}_0, \mathcal{A}\mathbf{r}_0, \ldots, \mathcal{A}^{k-1}\mathbf{r}_0\right\}.$$

To make the iterate $\boldsymbol{w}_k$ unique, the residual vector satisfies the condition $\mathbf{r}_k = \mathbf{d} - \mathcal{A}\boldsymbol{w}_k \perp \mathcal{L}_k$, where $\mathcal{L}_k$ is called the constraint space. Based on the choices for the $\mathcal{L}_k$, various classes of Krylov subspace methods have been developed. For example:

- $\mathcal{L}_k = \mathcal{K}_k(\mathcal{A}, \mathbf{r}_0)$ gives orthogonal residual methods such as conjugate gradient (CG) methods for SPD linear systems.
- $\mathcal{L}_k = \mathcal{A}\mathcal{K}_k(\mathcal{A}, \mathbf{r}_0)$ gives minimal residual (MINRES) method and generalized minimal residual (GMRES) method to solve symmetric and nonsymmetric linear systems, respectively.

1.2.3. Generalized Minimal Residual (GMRES) Method

The GMRES method is prominently one of the most useful Krylov subspace methods, which was first discussed by Saad [122] for solving a linear system $\mathcal{A}\boldsymbol{w} = \mathbf{d}$, where $\mathfrak{B}$ is nonsingular. By selecting the constraint space $\mathcal{L}_k = \mathcal{A}\mathcal{K}_k(\mathcal{A}, \mathbf{r}_0)$, which is equivalent to minimizing the norm of residual at each iteration, i.e., the approximate solution at kth iteration satisfies

$$\min_{\boldsymbol{w} \in \boldsymbol{w}_0 + \mathcal{K}_k(\mathcal{A}, \mathbf{r}_0)} \|\mathbf{d} - \mathcal{A}\boldsymbol{w}_k\|_2.$$

The GMRES uses the Arnoldi method to obtain an orthonormal basis for the Krylov subspace $\mathcal{K}_k(\mathcal{A}, \mathbf{r}_0)$.

The first vector in the Krylov subspace is $\mathbf{r}_0$, and the first orthonormal vector $v_1 = \mathbf{r}_0/\|\mathbf{r}_0\|_2$. Next, we take $\mathcal{A}v_1$ and orthonormalize it against v_1. After subsequent iterations, we obtain v_{m+1} by othonormalize $\mathcal{A}v_m$ against previous vectors. This gives the following matrix relation:

$$\mathcal{A}V_m = V_{m+1}\bar{H}_m, \tag{1.2.3}$$

where $V_m \in \mathbb{R}^{n \times m}$ consist of v_j as columns and $\bar{H}_m \in \mathbb{R}^{(m+1) \times m}$ is an upper Hessenberg matrix. Since V_m is orthogonal, we have $V_m^T A V_m = H_m$, where H_m consist of first m rows of $\bar{H}_m$.

In the mth step of the Krylov subspace method, the approximate solution is of the form $\boldsymbol{w}_m = \boldsymbol{w}_0 + V_m y$. Then, the residual $\mathbf{r}_m := \mathbf{d} - \mathcal{A}\boldsymbol{w}_m$ satisfies

$$\|\mathbf{r}_m\|_2 = \|\mathbf{d} - \mathcal{A}\boldsymbol{w}_m\|_2 = \|\mathbf{d} - \mathcal{A}\boldsymbol{w}_0 - \mathcal{A}V_m y\|_2 = \|V_{m+1}^T \mathbf{r}_0 - \bar{H}_m y\|_2 = \|\beta_0 e_1 - \bar{H}_m y\|_2.$$

Thus, y is chosen to minimizes $= \|\beta_0 e_1 - \bar{H}_m y\|_2$.

Algorithm 1.2.1 GMRES

Choose an initial guess vector $\boldsymbol{w}_0$, compute $\mathbf{r}_0 = \mathbf{d} - \mathcal{A}\boldsymbol{w}_0$;

$1 : \beta_0 = \|\mathbf{r}_0\|_2$; $v_1 := \mathbf{r}_0/\beta_0$;

$2 :$ for $j = 1, 2, \ldots,$ do

$3 :$ compute $u_j := \mathcal{A}v_j$;

$4 :$ for $i = 1, 2, \ldots, j,$ do

$5 : h_{ij} := (u_j, v_j)$;

$6 : u_j := u_j - h_{ij}v_j$;

$7 :$ end for

$8 : h_{j+1,j} = \|u_j\|_2$;

$9 : v_{j+1} = u_j/h_{j+1,j}$;

$10 :$ end for

$11 :$ Define the $(m+1) \times m$ Hessenberg matrix $\bar{H}_m = [h_{ij}]$;

$12 :$ Compute the minimizer y_m of $\|\beta_0 e_1 - \bar{H}_m y\|_2$ and form the solution

$$\boldsymbol{w}_m = \boldsymbol{w}_0 + V_m y_m.$$

1.2.4. Preconditioning

A major drawback of iterative solvers is their slow convergence and lack of robustness, particularly when the coefficient matrix of the system is ill-conditioned. To accelerate the convergence speed and increase the robustness of the Krylov subspace methods by applying suitable preconditioners. The term *preconditioning* refers to the method of transforming the linear system $\mathcal{A}\boldsymbol{w} = \mathbf{d}$ into an equivalent system that possesses more favorable properties for iterative solution methods. A *preconditioner* is a nonsingular matrix $\mathscr{P}$ that serves as a suitable approximation of $\mathcal{A}$, designed to improve the convergence properties of the chosen Krylov subspace method. When the preconditioner is multiplied from the left side leads to the following left preconditioned system:

$$\mathscr{P}^{-1}\mathcal{A}\boldsymbol{w} = \mathscr{P}^{-1}\mathbf{d}.$$

Alternatively, a preconditioner can be applied from the right side as well, which leads:

$$\mathcal{A}\mathscr{P}^{-1}\boldsymbol{u} = \mathbf{d}, \quad \boldsymbol{u} := \mathscr{P}\boldsymbol{w}.$$

If a preconditioner is available as a factored form as $\mathscr{P} = \mathscr{P}_1\mathscr{P}_2$, two-sided preconditioning leads to the following preconditioned system:

$$\mathscr{P}_1^{-1}\mathcal{A}\mathscr{P}_2^{-1}\boldsymbol{u} = \mathscr{P}_1^{-1}\mathbf{d}, \quad \boldsymbol{u} := \mathscr{P}_2\boldsymbol{w}.$$

When a Krylov subspace method, such as GMRES, is applied to the preconditioned system, it is referred to as the preconditioned GMRES (PGMRES) method.

In general, a preconditioner aims to enhance the spectral properties of the preconditioned matrix $\mathscr{P}^{-1}\mathcal{A}$ (or $\mathcal{A}\mathscr{P}^{-1}$). For symmetric problems, the convergence rate of Krylov subspace methods, such as CG and MINRES, is governed by the distribution of eigenvalues or spectral condition numbers. In the case of nonsymmetric problems, the situation becomes more complex, especially for methods like GMRES. However, when the eigenvalues of the preconditioned matrix are clustered, the convergence rate of the method improves significantly.

1.3. Preliminaries

This section provides fundamental definitions and key results that will be applied throughout this thesis. For the matrix $A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n] \in \mathbb{C}^{m\times n}$, where $\boldsymbol{a}_i \in \mathbb{C}^m$, $i = 1, 2, \ldots, n$, the linear operator vec : $\mathbb{R}^{m\times n} \mapsto \mathbb{R}^{mn}$ is defined by $\text{vec}(A) := [\boldsymbol{a}_1^T, \boldsymbol{a}_2^T, \ldots, \boldsymbol{a}_n^T]^T \in \mathbb{C}^{mn}$. The vec operator satisfies $\|\text{vec}(A)\|_\infty = \|A\|_F$. The sparsity pattern of a matrix $A = [a_{ij}] \in \mathbb{C}^{m\times n}$ is defined as $\Theta_A := \texttt{sgn}(A) = [\texttt{sgn}(a_{ij})]$. where

$$\texttt{sgn}(a_{ij}) = \begin{cases} 1, & a_{ij} \neq 0, \\ 0, & a_{ij} = 0. \end{cases}$$

The Hadamard product of $A, B \in \mathbb{C}^{m\times n}$ is defined as $A \odot B = [a_{ij}b_{ij}] \in \mathbb{C}^{m\times n}$. For any vector $x \in \mathbb{C}^m$, $\mathfrak{D}_x$ is the diagonal matrix defined as $\mathfrak{D}_x := \text{diag}(x) \in \mathbb{C}^{m\times m}$.

Definition 1.3.1. *Let $A = [a_{ij}] \in \mathbb{C}^{m\times n}$ and $B = [b_{ij}] \in \mathbb{C}^{p\times q}$. Then, the Kronecker product of A and B is denoted by $A \otimes B$ and defined as follows:*

$$A \otimes B := \begin{bmatrix} a_{11}B & \ldots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \ldots & a_{mn}B \end{bmatrix} \in \mathbb{C}^{mp\times nq}. \tag{1.3.1}$$

The Kronecker product of matrices satisfies the following properties [66, 84]: For $A \in \mathbb{C}^{r\times m}$, $X \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{s\times p}$, $Y \in \mathbb{C}^{p\times q}$, and $Z \in \mathbb{C}^{n\times p}$, we have

$$\begin{cases} \text{vec}(XZY) = (Y^T \otimes X)\text{vec}(Z), \\ (X \otimes Y)^T = X^T \otimes Y^T, \\ |X \otimes Y| = |X| \otimes |Y|, \\ (A \otimes B)(X \otimes Y) = AX \otimes BY. \end{cases} \tag{1.3.2}$$

Definition 1.3.2. *Let $A \in \mathbb{C}^{m\times n}$. The M-P inverse is the unique matrix $Y \in \mathbb{C}^{n\times m}$ that satisfies the following conditions:*

$$(1)\ AYA = A,\ (2)\ YAY = Y,\ (3)\ (AY)^H = AY\ (4)\ (YA)^H = YA. \tag{1.3.3}$$

The M-P inverse of a matrix $A \in C^{m\times n}$ is always uniquely exist and denoted by $A^\dagger$.

Lemma 1.3.1. *[133] Consider the system of linear equations $Ax = b$, where $A \in \mathbb{C}^{m\times n}, b \in \mathbb{C}^m$. The following results hold:*

(1) The linear system is consistent if and only if $AA^\dagger b = b$. Furthermore, when the system is consistent, the solution with the minimum norm is given by $A^\dagger b$.

(2) The MNLS solutions of the LS problem $\min\limits_x \|Ax - b\|_2$ is given by $A^\dagger b$.

Definition 1.3.3. *[117] Let $\widetilde{x}$ be an approximate solution of the linear system $Ax = b$. Then, the normwise unstructured BE, denoted by $\boldsymbol{\eta}(\widetilde{x})$, is defined as:*

$$\boldsymbol{\eta}(\widetilde{x}) := \min_{(\Delta A,\ \Delta b)\in\mathcal{F}} \left\| \begin{bmatrix} \frac{\|\Delta A\|_F}{\|A\|_F} & \frac{\|\Delta b\|_F}{\|b\|_F} \end{bmatrix} \right\|_2,$$

where $\mathcal{F} = \left\{(\Delta A,\ \Delta b) \middle| (A + \Delta A)\widetilde{x} = b + \Delta b\right\}$.

Rigal and Gaches [117] provided explicit expression for the BE defined above, which is given by

$$\boldsymbol{\eta}(\widetilde{x}) = \frac{\|b - A\widetilde{x}\|_2}{\sqrt{\|A\|_F^2\|\widetilde{x}\|_2^2 + \|b\|_2^2}}. \tag{1.3.4}$$

When $\boldsymbol{\eta}(\widetilde{x})$ is sufficiently small, the approximate solution $\widetilde{x}$ becomes the exact solution to a slightly perturbed system, $(A + \Delta A)\widetilde{x} = b + \Delta b$, where both $\|\Delta A\|_F$ and $\|\Delta b\|_2$ are relatively small. A numerical algorithm to solve a problem is backward stable if the approximate solution of the problem is always the exact solution of a nearby problem [73].

For any matrix $A = [a_{ij}] \in \mathbb{R}^{m\times n}$, we set $|A| := [|a_{ij}|]$, where $|a_{ij}|$ denotes the absolute value of a_{ij}. For two matrices $A, B \in \mathbb{R}^{m\times n}$, the notation $|A| \le |B|$ represents $|a_{ij}| \le |b_{ij}|$ for all $1 \le i \le m$ and $1 \le j \le n$. According to [51, 88], we define the following notations. The componentwise distance between two vectors a and b in $\mathbb{R}^p$ is defined as:

$$d(a, b) = \left\| \frac{a - b}{b} \right\|_\infty = \max_{i=1,2,\ldots,p} \left\{ \frac{|a_i - b_i|}{|b_i|} \right\}. \tag{1.3.5}$$

Let $u \in \mathbb{R}^p$ and $\eta > 0$, consider the sets: $B_1(u, \eta) = \{x \in \mathbb{R}^p : \|x - u\|_2 \le \eta\|u\|_2\}$ and $B_2(u, \eta) = \{x \in \mathbb{R}^p : |x_i - u_i| \le \eta|u_i|,\ i = 1, \ldots, p\}$.

With the above conventions, next, we present the definitions of NCN, MCN, and CCN for a mapping $\boldsymbol{\varphi} : \mathbb{R}^p \mapsto \mathbb{R}^q$ as follows.

Definition 1.3.4. *[51, 63] Let $\boldsymbol{\varphi} : \mathbb{R}^p \mapsto \mathbb{R}^q$ be a continuous mapping defined on an open set $\Omega_{\boldsymbol{\varphi}} \subseteq \mathbb{R}^p$, and $\mathbf{0} \neq u \in \Omega_{\boldsymbol{\varphi}}$ such that $\boldsymbol{\varphi}(u) \neq \mathbf{0}$.*

(i) The NCN of $\boldsymbol{\varphi}$ at u is defined by

$$\mathscr{K}(\boldsymbol{\varphi}, u) = \lim_{\eta \to 0} \sup_{\substack{x \neq u \\ x \in B_1(u,\eta)}} \frac{\|\boldsymbol{\varphi}(x) - \boldsymbol{\varphi}(u)\|_2 / \|\boldsymbol{\varphi}(u)\|_2}{\|x - u\|_2 / \|u\|_2}.$$

(ii) The MCN of $\boldsymbol{\varphi}$ at u is defined by

$$\mathscr{M}(\boldsymbol{\varphi}, u) = \lim_{\eta \to 0} \sup_{\substack{x \neq u \\ x \in B_2(u,\eta)}} \frac{\|\boldsymbol{\varphi}(x) - \boldsymbol{\varphi}(u)\|_\infty}{\|\boldsymbol{\varphi}(u)\|_\infty} \frac{1}{d(x, u)}.$$

(iii) Let $\boldsymbol{\varphi}(u) = [\boldsymbol{\varphi}(u)_1, \ldots, \boldsymbol{\varphi}(u)_q]^T$ be such that $\boldsymbol{\varphi}(u)_i \neq 0$ for $i = 1, 2, \ldots, q$. Then, the CCN of $\boldsymbol{\varphi}$ at u is defined by

$$\mathscr{C}(\boldsymbol{\varphi}, u) = \lim_{\eta \to 0} \sup_{\substack{x \neq u \\ x \in B_2(u,\eta)}} \frac{d(\boldsymbol{\varphi}(x), \boldsymbol{\varphi}(u))}{d(x, u)}.$$

Next, we present the definition of the *Fréchet* derivative, which serves as the foundation for deriving expressions of the CNs.

Definition 1.3.5. *[46] Let $\boldsymbol{\varphi} : \mathbb{R}^p \mapsto \mathbb{R}^q$ be a mapping defined on an open set $\Omega_{\boldsymbol{\varphi}} \subseteq \mathbb{R}^p$. Then $\boldsymbol{\varphi}$ is said to be Fréchet differentiable at $u \in \Omega_{\boldsymbol{\varphi}}$ if there exists a bounded linear operator $\mathbf{d}\boldsymbol{\varphi} : \mathbb{R}^p \mapsto \mathbb{R}^q$ such that*

$$\lim_{h \to \mathbf{0}} \frac{\|\boldsymbol{\varphi}(u + h) - \boldsymbol{\varphi}(u) - \mathbf{d}\boldsymbol{\varphi} h\|}{\|h\|} = 0,$$

where $\|\cdot\|$ denotes any norm on $\mathbb{R}^p$ and $\mathbb{R}^q$.

When $\boldsymbol{\varphi}$ is *Fréchet* differentiable at u, we denote the *Fréchet* derivative at u as $\mathbf{d}\boldsymbol{\varphi}(u)$. The next lemma gives closed-form expressions for the above three CNs when the continuous mapping $\boldsymbol{\varphi}$ is *Fréchet* differentiable.

Lemma 1.3.2. *[51, 63] Under the same hypothesis as in Definition 1.3.4, when $\boldsymbol{\varphi}$ is Fréchet differentiable at u, we have*

$$\mathscr{K}(\boldsymbol{\varphi}; u) = \frac{\|\mathbf{d}\boldsymbol{\varphi}(u)\|_2 \|u\|_2}{\|\boldsymbol{\varphi}(u)\|_2}, \quad \mathscr{M}(\boldsymbol{\varphi}; u) = \frac{\||\mathbf{d}\boldsymbol{\varphi}(u)||u|\|_\infty}{\|\boldsymbol{\varphi}(u)\|_\infty}, \text{ and } \quad \mathscr{C}(\boldsymbol{\varphi}; u) = \left\| \frac{|\mathbf{d}\boldsymbol{\varphi}(u)||u|}{|\boldsymbol{\varphi}(u)|} \right\|_\infty.$$

CHAPTER 2

A Robust Parameterized Enhanced Shift-Splitting Preconditioner for Double Saddle Point Problems*

This chapter proposes a novel parameterized enhanced shift-splitting (PESS) preconditioner to solve the DSPP by considering $F = B$, $G = C$, $D = \mathbf{0}$, and $E = \mathbf{0}$. Additionally, we introduce a local PESS (LPESS) preconditioner by relaxing the PESS preconditioner. Necessary and sufficient criteria are established for the convergence of the proposed PESS iterative method for any initial guess. Furthermore, we meticulously investigate the spectral bounds of the PESS and LPESS preconditioned matrices. Moreover, empirical investigations have been performed for the sensitivity analysis of the proposed PESS preconditioner, which unveils its robustness. Numerical experiments are carried out to demonstrate the enhanced efficiency and robustness of the proposed PESS and LPESS preconditioners compared to the existing state-of-the-art preconditioners.

2.1. Background

We consider the DSPP of the following form [75]:

$$\mathcal{A}\mathbf{u} := \begin{bmatrix} A & B^T & \mathbf{0} \\ -B & \mathbf{0} & -C^T \\ \mathbf{0} & C & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} = \begin{bmatrix} f \\ -g \\ h \end{bmatrix} =: \widehat{\mathbf{d}}, \tag{2.1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is a SPD matrix, $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ are the full row rank matrices. Here, $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$ and $h \in \mathbb{R}^p$ are known vectors. Under the stated assumptions on block matrices A, B and C, $\mathcal{A}$ is nonsingular, which ensures the existence of a unique solution for the system (2.1.1).

After an appropriate partitioning of the coefficient matrix $\mathcal{A}$, the linear system (2.1.1) can be recognized as a standard two-by-two SPP. Over the past few decades, significant efforts have been devoted to developing numerical solution methods for standard SPPs

* S. S. Ahmad and **P. Khatun**, "A robust parameterized enhanced shift-splitting preconditioner for three-by-three block saddle point problems." *Journal of Computational and Applied Mathematics,* 459:116358, 2025.

[26]. However, due to the distinct properties of the submatrices, these methods cannot be directly applied to solve the DSPP (2.1.1).

Recently, various effective preconditioners have been developed in the literature for solving the DSPP (2.1.1). For instance, Huang and Ma [75] have studied the exact BD and IBD preconditioners $\mathscr{P}_{BD}$ and $\mathscr{P}_{IBD}$ in the following forms:

$$\mathscr{P}_{BD} = \begin{bmatrix} A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & S & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & CS^{-1}C^T \end{bmatrix} \quad \text{and} \quad \mathscr{P}_{IBD} = \begin{bmatrix} \widehat{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widehat{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & C\widehat{S}^{-1}C^T \end{bmatrix}, \tag{2.1.2}$$

where $S = BA^{-1}B^T$, $\widehat{A}$ and $\widehat{S}$ are SPD approximations of A and S, respectively. Although the spectrum of the preconditioned matrices $\mathscr{P}_{BD}^{-1}\mathcal{A}$ and $\mathscr{P}_{IBD}^{-1}\mathcal{A}$ have good clustering properties, these preconditioners have certain shortcomings, such as they are time-consuming, expensive, require an excessive number of iterations and CNs are very large. For more details, see the paper [75]. Inspired by the HSS iteration method [13], Salkuyeh et al. [125] split the coefficient matrix $\mathcal{A}$ and presented the APSS iteration method. They proved the unconditional convergence of the iteration method and proposed the corresponding APSS preconditioner to solve the system (2.1.1). Moreover, to improve the effectiveness of the APSS preconditioner, many relaxed and modified versions of APSS have been designed; see [43, 150, 89]. For instance, by relaxing the $(1,1)$ block and introducing a new parameter $\beta > 0$ in the APSS preconditioner, Chen and Ren [43] proposed the following modified APSS (MAPSS) preconditioner:

$$\mathscr{P}_{MAPSS} = \begin{bmatrix} A & B^T & -\frac{1}{\alpha}B^TC^T \\ -B & \alpha I & -C^T \\ \mathbf{0} & C & \beta I \end{bmatrix}, \tag{2.1.3}$$

where $\alpha > 0$ and I stands for the identity matrix of the appropriate dimension. Motivated by the work in [39] and by incorporating the ideas of shifting and the BD preconditioner, the authors in [18] proposed two preconditioners along with their inexact variants. One of these preconditioners is given by:

$$\mathscr{P}_{SL} = \begin{bmatrix} A & B^T & \mathbf{0} \\ -B & C^TC & \mathbf{0} \\ \mathbf{0} & C & I \end{bmatrix}. \tag{2.1.4}$$

Three block preconditioners are developed by Xie and Li [148] for the system (2.1.1) with the equivalent symmetric coefficient matrix. It is also demonstrated in [148] that the

preconditioned matrices possess no more than three distinct eigenvalues. Huang [74] proposed a variant of the Uzawa iterative method for the DSPP by introducing two variable parameters. Additionally, Huang et al. [76] generalized the well-known Uzawa method to solve the linear system (2.1.1) and propose the inexact Uzawa method. In addition to the preconditioners discussed above, recent literature [1, 136, 19] have introduced several other preconditioners to solve the DSPP (2.1.1).

The SS preconditioners were initially developed for a non-Hermitian system of linear equations by Bai et al. [16] and later for two-by-two block SPPs [40, 124, 41]. Recently, Cao [37] enhanced this idea and introduced the following SS preconditioner $\mathscr{P}_{SS}$ and relaxed SS (RSS) preconditioner $\mathscr{P}_{RSS}$ for the DSPP (2.1.1):

$$\mathscr{P}_{SS} = \frac{1}{2}\begin{bmatrix} \alpha I + A & B^T & \mathbf{0} \\ -B & \alpha I & -C^T \\ \mathbf{0} & C & \alpha I \end{bmatrix} \text{ and } \mathscr{P}_{RSS} = \frac{1}{2}\begin{bmatrix} A & B^T & \mathbf{0} \\ -B & \alpha I & -C^T \\ \mathbf{0} & C & \alpha I \end{bmatrix}, \tag{2.1.5}$$

where $\alpha > 0$ and verified unconditionally convergence of the associated SS iterative method. Wang and Zhang [134] generalized the SS preconditioner by introducing a new parameter $\beta > 0$ in the $(3,3)$ block. By merging the lopsided and shift technique, a lopsided SS preconditioner is presented by Zhang et al. [160]. Further, Yin et al. [156] developed the following extensive generalized SS (EGSS) preconditioner:

$$\mathscr{P}_{EGSS} = \frac{1}{2}\begin{bmatrix} \alpha P + A & B^T & \mathbf{0} \\ -B & \beta Q & -C^T \\ \mathbf{0} & C & \gamma W \end{bmatrix} \tag{2.1.6}$$

for the DSPP (2.1.1), where $\alpha, \beta, \gamma > 0$ and P, Q, W are SPD matrices and investigated its convergent properties. By relaxing the $(1,1)$ block and eliminating the prefactor $1/2$, Liang and Zhu [92] have proposed relaxed and preconditioned generalized SS (RPGSS) preconditioner

$$\mathscr{P}_{RPGSS} = \begin{bmatrix} A & B^T & \mathbf{0} \\ -B & \beta Q & -C^T \\ \mathbf{0} & C & \gamma W \end{bmatrix}, \tag{2.1.7}$$

where $Q \in \mathbb{R}^{m\times m}$ and $W \in \mathbb{R}^{p\times p}$ are SPD and $\beta, \gamma > 0$.

Despite exhibiting favorable performance of SS, RSS, and EGSS preconditioners, they do not outperform BD and IBD preconditioners. Therefore, to improve the convergence speed and efficiency of the preconditioners $\mathscr{P}_{SS}$, $\mathscr{P}_{RSS}$ and $\mathscr{P}_{EGSS}$, this chapter presents a PESS preconditioner by introducing a parameter $s > 0$ for DSPPs. As per the direct

correlation between the rate of convergence in Krylov subspace iterative methods and the spectral properties of the preconditioned matrix, we perform in-depth spectral distribution of the PESS and LPESS preconditioned matrices. We summarize the main contributions of this chapter as follows:

- We propose the PESS iterative method along with its associated PESS preconditioner and its relaxed version known as the LPESS preconditioner to solve the DSPP (2.1.1).
- General framework is given on necessary and sufficient criteria for the convergence of the PESS iterative method. These investigations also encompass the unconditional convergence of other exiting SS preconditioners in the literature [37, 156].
- We have conducted a comprehensive analysis of the spectral distribution for the PESS and LPESS preconditioned matrices. Our framework allows us to derive the spectral distribution for the existing SS and EGSS preconditioned matrices.
- We empirically show that our proposed preconditioner significantly reduces the CN of the ill-conditioned saddle point matrix $\mathcal{A}$. Thereby establishing an efficiently solvable, well-conditioned system.
- Numerical experiments show that the proposed PESS and LPESS preconditioners outperform all the compared baseline preconditioners.

The structure of this chapter is as follows. In Section 2.2, we propose the PESS iterative method and the associated PESS preconditioner. Section 2.3 is devoted to investigating the convergence of the PESS iterative method. In Section 2.4, we investigate the spectral distribution of the PESS preconditioned matrix. In Section 2.5, we present the LPESS preconditioner. In Section 2.6, we discuss strategies for selecting the appropriate parameters for the proposed preconditioners. In Section 2.7, we conduct numerical experiments to illustrate the computational efficiency and robustness of the proposed preconditioner. In the end, a brief summary remark is provided in Section 2.8.

2.2. The Proposed Parameterized Enhanced Shift-Splitting (PESS) Iterative Method and Preconditioner

In this section, we proposed a PESS iterative method and the corresponding PESS preconditioner. Let $s > 0$ be a real number. Then, we split the matrix $\mathcal{A}$ in the form

$$\mathcal{A} = (\Sigma + s\mathcal{A}) - (\Sigma - (1-s)\mathcal{A}) =: \mathscr{P}_{PESS} - \mathcal{Q}_{PESS}, \quad \text{where} \tag{2.2.1}$$

$$\mathscr{P}_{PESS} = \begin{bmatrix} \Lambda_1 + sA & sB^T & \mathbf{0} \\ -sB & \Lambda_2 & -sC^T \\ \mathbf{0} & sC & \Lambda_3 \end{bmatrix},$$

$$\mathcal{Q}_{PESS} = \begin{bmatrix} \Lambda_1 - (1-s)A & -(1-s)B^T & \mathbf{0} \\ (1-s)B & \Lambda_2 & (1-s)C^T \\ \mathbf{0} & -(1-s)C & \Lambda_3 \end{bmatrix},$$

$\Sigma = \text{diag}(\Lambda_1, \Lambda_2, \Lambda_3)$ and $\Lambda_1, \Lambda_2, \Lambda_3$ are SPD matrices. The matrix $\mathscr{P}_{PESS}$ is nonsingular for $s > 0$. The special matrix splitting (2.2.1) leads us to the subsequent iterative method for solving the DSPP given in (2.1.1).

The PESS iterative method. Let s be a positive constant and let $\Lambda_1 \in \mathbb{R}^{n\times n}, \Lambda_2 \in \mathbb{R}^{m\times m}$, and $\Lambda_3 \in \mathbb{R}^{p\times p}$ be SPD matrices. For any initial guess $\mathbf{u}_0 \in \mathbb{R}^{n+m+p}$ and $k = 0, 1, 2\ldots$, until a specified termination criterion is fulfilled, we compute

$$\mathbf{u}_{k+1} = \mathcal{T}\mathbf{u}_k + \mathbf{c}, \tag{2.2.2}$$

where $\mathbf{u}_k = [x_k^T, y_k^T, z_k^T]^T, \mathbf{c} = \mathscr{P}_{PESS}^{-1}\mathbf{d}$ and $\mathcal{T} = \mathscr{P}_{PESS}^{-1}\mathcal{Q}_{PESS}$ is called the iteration matrix for the PESS iterative method.

Moreover, the matrix splitting (2.2.2) introduces a new preconditioner $\mathscr{P}_{PESS}$, identified as the PESS preconditioner. This preconditioner generalizes the previous work in the literature; for example, refer to [37, 134, 156] for specific choices of $\Lambda_1, \Lambda_2, \Lambda_3$ and s, which are listed in Table 2.2.1.

Table 2.2.1: $\mathscr{P}_{PESS}$ as a generalization of the above SS preconditioners for different choices of $\Lambda_1, \Lambda_2, \Lambda_3$ and s.

$\mathscr{P}_{PESS}$	Λ_1	Λ_2	Λ_3	s	memo
$\mathscr{P}_{SS}$ [37]	$\frac{1}{2}\alpha I$	$\frac{1}{2}\alpha I$	$\frac{1}{2}\alpha I$	$s = \frac{1}{2}$	$\alpha > 0$
$\mathscr{P}_{GSS}$ [134]	$\frac{1}{2}\alpha I$	$\frac{1}{2}\alpha I$	$\frac{1}{2}\beta I$	$s = \frac{1}{2}$	$\alpha, \beta > 0$
$\mathscr{P}_{EGSS}$ [156]	$\frac{1}{2}\alpha P$	$\frac{1}{2}\beta Q$	$\frac{1}{2}\gamma W$	$s = \frac{1}{2}$	P, Q, W are SPD matrices and $\alpha, \beta, \gamma > 0$

In the implementation of the PESS iterative method or the PESS preconditioner to enhance the rate of convergence of the Krylov subspace iterative method like GMRES, at each iterative step, we solve the following system of linear equations:

$$\mathscr{P}_{PESS}w = r, \tag{2.2.3}$$

where $r = [r_1^T, r_2^T, r_3^T]^T \in \mathbb{R}^{n+m+p}$ and $w = [w_1^T, w_2^T, w_3^T]^T \in \mathbb{R}^{n+m+p}$. Specifying $\widehat{X} = \Lambda_2 + s^2 C^T \Lambda_3^{-1} C$ and $\widetilde{A} = \Lambda_1 + sA + s^2 B^T \widehat{X}^{-1} B$, then $\mathscr{P}_{PESS}$ can be written in the following way:

$$\mathscr{P}_{PESS} = \begin{bmatrix} I & sB^T\widehat{X}^{-1} & \mathbf{0} \\ \mathbf{0} & I & -sC^T\Lambda_3^{-1} \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \widetilde{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widehat{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Lambda_3 \end{bmatrix} \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ -s\widehat{X}^{-1}B & I & \mathbf{0} \\ \mathbf{0} & s\Lambda_3^{-1}C & I \end{bmatrix}. \quad (2.2.4)$$

The decomposition (2.2.4) leads us to the following Algorithm 2.2.1 to determine the solution of the system (2.2.3). The implementation of Algorithm 2.2.1 requires to solve

Algorithm 2.2.1 Computation of w from $\mathscr{P}_{PESS} w = r$

1: Solve $\widehat{X} v_1 = r_2 + sC^T\Lambda_3^{-1} r_3$ to find v_1;
2: Compute $v = r_1 - sB^T v_1$;
3: Solve $\widetilde{A} w_1 = v$ to find w_1;
4: Solve $\widehat{X} v_2 = sBw_1$ to find v_2;
5: Compute $w_2 = v_1 + v_2$;
6: Solve $\Lambda_3 w_3 = r_3 - sCw_2$ for w_3.

two linear subsystems with coefficient matrices $\widehat{X}$ and $\widetilde{A}$. Since $\Lambda_1, \Lambda_2, \Lambda_3, \widehat{X}$ and $\widetilde{A}$ are SPD and $s > 0$, we can apply the Cholesky factorization to solve these linear subsystems exactly or inexactly by the preconditioned conjugate gradient method.

2.3. Convergence Analysis of the PESS Iterative Method

In this section, we investigate the convergence of the PESS iterative method (2.2.2). To achieve this aim, the following definition and lemmas are crucial.

Definition 2.3.1. *A matrix $\mathcal{A}$ is called positive stable if $\Re(\lambda) > 0$ for all $\lambda \in \sigma(\mathcal{A})$.*

Lemma 2.3.1. *[37] Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices. Then, the saddle point matrix $\mathcal{A}$ is positive stable.*

Lemma 2.3.2. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices. Then $\Sigma^{-1}\mathcal{A}$ (or $\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}$) is positive stable.*

Proof. Since $\Sigma^{-\frac{1}{2}} = \text{diag}(\Lambda_1^{-\frac{1}{2}}, \Lambda_2^{-\frac{1}{2}}, \Lambda_3^{-\frac{1}{2}})$, we obtain

$$\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}} = \begin{bmatrix} \Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}} & \Lambda_1^{-\frac{1}{2}}B^T\Lambda_2^{-\frac{1}{2}} & \mathbf{0} \\ -\Lambda_2^{-\frac{1}{2}}B\Lambda_1^{-\frac{1}{2}} & \mathbf{0} & \Lambda_2^{-\frac{1}{2}}C^T\Lambda_3^{-\frac{1}{2}} \\ \mathbf{0} & \Lambda_3^{-\frac{1}{2}}C\Lambda_2^{-\frac{1}{2}} & \mathbf{0} \end{bmatrix}. \tag{2.3.1}$$

Given that the matrices Λ_1, Λ_2 and Λ_3 are SPD, then $\Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}}$ is SPD. Furthermore, the matrices $\Lambda_2^{-\frac{1}{2}}B\Lambda_1^{-\frac{1}{2}}$ and $\Lambda_3^{-\frac{1}{2}}C\Lambda_2^{-\frac{1}{2}}$ are of full row rank. Consequently, the block structure of $\mathcal{A}$ and the matrix $\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}$ are identical. Then by Lemma 2.3.1, $\mathcal{A}$ is positive stable implies that the matrix $\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}$ and consequently, $\Sigma^{-1}\mathcal{A}$ is positive stable. ■

By Lemma 1.2.1, the stationary iterative method (2.2.2) converges to the exact solution of the DSPP (2.1.1) for any initial guess vector if and only if $|\vartheta(\mathcal{T})| < 1$, where $\mathcal{T}$ is the iteration matrix. Now, we establish an if and only if condition that precisely determines the convergence of the PESS iterative method.

Theorem 2.3.3. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix, and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices with $s > 0$. Then, the PESS iterative method (2.2.2) converges to the unique solution of the DSPP (2.1.1) if and only if*

$$(2s-1)|\mu|^2 + 2\Re(\mu) > 0, \ \forall\, \mu \in \sigma(\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}). \tag{2.3.2}$$

Proof. From (2.2.2), we have

$$\mathcal{T} = \mathscr{P}_{PESS}^{-1}\mathcal{Q}_{PESS} = (\Sigma + s\mathcal{A})^{-1}(\Sigma - (1-s)\mathcal{A}), \tag{2.3.3}$$

where $\Sigma = \text{diag}(\Lambda_1, \Lambda_2, \Lambda_3)$. Given that Λ_1, Λ_2 and Λ_3 are SPD, then

$$\mathcal{T} = \Sigma^{-\frac{1}{2}}(I + s\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}})^{-1}(I - (1-s)\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}})\Sigma^{\frac{1}{2}}. \tag{2.3.4}$$

Thus, (2.3.4) shows that the iteration matrix $\mathcal{T}$ is similar to $\widetilde{\mathcal{T}}$, where

$$\widetilde{\mathcal{T}} = (I + s\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}})^{-1}(I - (1-s)\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}).$$

Let θ and μ be the eigenvalues of the matrices $\widetilde{\mathcal{T}}$ and $\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}$, respectively. Then, it is easy to show that

$$\theta = \frac{1-(1-s)\mu}{1+s\mu}.$$

Since $\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}$ is nonsingular then $\mu \neq 0$ and we have

$$|\theta| = \frac{|1-(1-s)\mu|}{|1+s\mu|} = \sqrt{\frac{(1-(1-s)\Re(\mu))^2 + (1-s)^2\Im(\mu)^2}{(1+s\Re(\mu))^2 + s^2\Im(\mu)^2}}. \tag{2.3.5}$$

From (2.3.5), it is clear that $|\theta| < 1$ if and only if

$$(1-(1-s)\Re(\mu))^2 + (1-s)^2\Im(\mu)^2 < (1+s\Re(\mu))^2 + s^2\Im(\mu)^2.$$

This implies $|\theta| < 1$ if and only if $(2s-1)|\mu|^2 + 2\Re(\mu) > 0$. Since, $\vartheta(\mathcal{T}) = \vartheta(\widetilde{\mathcal{T}})$, the proof is conclusive. ■

Using the condition in Theorem 2.3.3, next, we present two sufficient conditions that ensure the convergence of the PESS iterative method. The first one is presented as follows.

Corollary 2.3.1. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices and if $s \geq \frac{1}{2}$, then $\vartheta(\mathcal{T}) < 1$, i.e., the PESS iterative method always converges to the unique solution of the DSPP (2.1.1).*

Proof. By Lemma 2.3.2, $\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}}$ is positive stable, implies that $\Re(\mu) > 0$ and $|\mu| > 0$. Thus, the inequality (2.3.2) holds if $s \geq \frac{1}{2}$. This completes the proof. ■

Remark 2.3.4. *By applying Corollary 2.3.1, the unconditional convergence of the existing preconditioners, namely $\mathscr{P}_{SS}$, $\mathscr{P}_{GSS}$ and $\mathscr{P}_{EGSS}$ can be obtained from $\mathscr{P}_{PESS}$ by substituting $s = 1/2$.*

Next, we present a stronger sufficient condition to Corollary 2.3.1 for the convergence of the PESS iterative method.

Corollary 2.3.2. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices, if*

$$s > \max\left\{\frac{1}{2}\left(1 - \frac{\lambda_{\min}(\Sigma^{-\frac{1}{2}}(\mathcal{A}+\mathcal{A}^T)\Sigma^{-\frac{1}{2}})}{\vartheta(\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}})^2}\right), 0\right\},$$

then the PESS iterative method is convergent for any initial guess.

Proof. Note that, as $\Re(\mu) > 0$, the condition in (2.3.2) holds if and only if $\frac{1}{2} - \frac{\Re(\mu)}{|\mu|^2} < s$. Let $\mathbf{p}$ be an eigenvector corresponding to μ, then we have

$$\Re(\mu) = \frac{\mathbf{p}^H(\Sigma^{-\frac{1}{2}}(\mathcal{A}+\mathcal{A}^T)\Sigma^{-\frac{1}{2}})\mathbf{p}}{2\mathbf{p}^H\mathbf{p}} \geq \frac{1}{2}\lambda_{\min}(\Sigma^{-\frac{1}{2}}(\mathcal{A}+\mathcal{A}^T)\Sigma^{-\frac{1}{2}}).$$

Since $|\mu| \leq \vartheta(\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}})$, we obtain the following:

$$\frac{1}{2} - \frac{\lambda_{\min}(\Sigma^{-\frac{1}{2}}(\mathcal{A}+\mathcal{A}^T)\Sigma^{-\frac{1}{2}})}{2\vartheta(\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}})^2} \geq \frac{1}{2} - \frac{\Re(\mu)}{|\mu|^2}. \tag{2.3.6}$$

If we apply $s > 0$, then we get the desired result from (2.3.6). ■

The sufficient condition on Corollary 2.3.2 is difficult to find for large $\mathcal{A}$ due to the involvement of computation of $\lambda_{\min}(\Sigma^{-\frac{1}{2}}(\mathcal{A}+\mathcal{A}^T)\Sigma^{-\frac{1}{2}})$ and $\vartheta(\Sigma^{-\frac{1}{2}}\mathcal{A}\Sigma^{-\frac{1}{2}})^2$. Thus, we consider $s \geq \frac{1}{2}$ for practical implementation.

2.4. Spectral Distribution of the PESS Preconditioned Matrix

Solving the DSPP (2.1.1) is equivalent to solving the following preconditioned system of linear equations:

$$\mathscr{P}_{PESS}^{-1}\mathcal{A}\mathbf{u} = \mathscr{P}_{PESS}^{-1}\mathbf{d}, \tag{2.4.1}$$

where $\mathscr{P}_{PESS}$ serves as a preconditioner for the preconditioned GMRES (PGMRES) method. Since preconditioned matrices with clustered spectrum frequently culminate in rapid convergence for GMRES, see, for example, [26, 122], the spectral distribution of the preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ requires careful attention. As an immediate consequence of Corollary 2.3.1, the following clustering property for the eigenvalues of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ can be established.

Theorem 2.4.1. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ are full row rank matrices. Suppose that λ is an eigenvalue of the PESS preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$. Then, for $s \geq \frac{1}{2}$, we have*

$$|\lambda - 1| < 1,$$

i.e., the spectrum of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ entirely contained in a disk with centered at $(1, 0)$ and radius strictly smaller than 1.

Proof. Assume that λ and θ are the eigenvalues of the PESS preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and the iteration matrix $\mathcal{T}$, respectively. On the other hand, from the matrix splitting (2.2.1), we have

$$\mathscr{P}_{PESS}^{-1}\mathcal{A} = \mathscr{P}_{PESS}^{-1}(\mathscr{P}_{PESS} - \mathcal{Q}_{PESS}) = I - \mathcal{T}.$$

Then, we get $\lambda = 1 - \theta$. Now, from Corollary 2.3.1, it holds that $|\theta| < 1$ for $s \geq \frac{1}{2}$. Hence, it follows that, $|\lambda - 1| < 1$ for $s \geq \frac{1}{2}$, which completes the proof. ■

Let $(\lambda, \mathbf{p} = [u^T, v^T, w^T]^T)$ be an eigenpair of the preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$. Then, we have $\mathcal{A}\mathbf{p} = \lambda\mathscr{P}_{PESS}\mathbf{p}$, which can be written as

$$\begin{cases} Au + B^T v = \lambda(\Lambda_1 + sA)u + s\lambda B^T v, & \text{(2.4.2a)} \\ -Bu - C^T w = -s\lambda Bu + \lambda\Lambda_2 v - s\lambda C^T w, & \text{(2.4.2b)} \\ Cv = s\lambda Cv + \lambda\Lambda_3 w. & \text{(2.4.2c)} \end{cases}$$

Remark 2.4.2. *Notice that from (2.4.2a)-(2.4.2c), we get $\lambda \neq 1/s$, otherwise , $\mathbf{p} = [u^T, v^T, w^T]^T = \mathbf{0} \in \mathbb{R}^{n+m+p}$, which is impossible as $\mathbf{p}$ is an eigenvector.*

Proposition 2.4.3. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices and $s > 0$. Let $(\lambda, \mathbf{p} = [u^T, v^T, w^T]^T)$ be an eigenpair of the preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$. Then, the following holds:*

(1) $u \neq \mathbf{0}$,
(2) *when λ is real, $\lambda > 0$,*
(3) $\Re(\mu) > 0$, *where $\mu = \lambda/(1 - s\lambda)$.*

Proof. (1) Let $u = \mathbf{0}$. Then by (2.4.2a), we get $(1 - s\lambda)B^T v = \mathbf{0}$, which shows that $v = \mathbf{0}$, as B has full row rank and $\lambda \neq 1/s$. Furthermore, when combined with (2.4.2c) leads to $w = \mathbf{0}$, as Λ_3 is SPD. Combining the above facts, it follows that $\mathbf{p} = [u^T, v^T, w^T]^T = \mathbf{0}$, which is impossible as $\mathbf{p}$ is an eigenvector. Hence, $u \neq \mathbf{0}$.
(2) Let $\mathbf{p} = [u^T, v^T, w^T]^T$ be an eigenvector corresponding to the real eigenvalue λ. Then, we have $\mathscr{P}_{PESS}^{-1}\mathcal{A}\mathbf{p} = \lambda\mathbf{p}$, or $\mathcal{A}\mathbf{p} = \lambda\mathscr{P}_{PESS}\,\mathbf{p}$. Consequently, λ can be written as

$$\begin{aligned} \lambda &= \frac{\mathbf{p}^T\mathcal{A}\mathbf{p} + \mathbf{p}^T\mathcal{A}^T\mathbf{p}}{2(\mathbf{p}^T\mathscr{P}_{PESS}\mathbf{p} + \mathbf{p}^T\mathscr{P}_{PESS}^T\mathbf{p})} \\ &= \frac{u^T Au}{2(su^T Au + u^T\Lambda_1 u + v^T\Lambda_2 v + w^T\Lambda_3 w)} > 0. \end{aligned}$$

The last inequality follows from the assumptions that A, Λ_1, Λ_2 and Λ_3 are SPD matrices.
(3) The system of linear equations (2.4.2a)-(2.4.2c) can be reformulated as

$$\begin{cases} Au + B^T v = (\lambda/(1 - s\lambda))\Lambda_1 u, \\ -Bu - C^T w = (\lambda/(1 - s\lambda))\Lambda_2 v, \\ Cv = (\lambda/(1 - s\lambda))\Lambda_3 w. \end{cases} \quad \text{(2.4.3)}$$

The system of linear equations in (2.4.3) can also be expressed as $\mathcal{A}\mathbf{p} = \mu\Sigma\mathbf{p}$, or $\Sigma^{-1}\mathcal{A}\mathbf{p} = \mu\mathbf{p}$, where $\mu = \lambda/(1 - s\lambda)$. By Lemma (2.3.2), we have $\Sigma^{-1}\mathcal{A}$ is positive stable, which implies $\Re(\mu) > 0$, hence the proof is completed. ■

Remark 2.4.4. *From Proposition 2.4.3, it follows that when λ is real, then $\mu > 0$. Furthermore, the values of λ lies in $(0, 1/s)$ for all $s > 0$.*

In the following theorem, we provide sharper bounds for real eigenvalues of the PESS preconditioned matrix. Before that, we introduce the following notation:

$$\xi_{\max} := \lambda_{\max}(\Lambda_1^{-1}A),\ \xi_{\min} := \lambda_{\min}(\Lambda_1^{-1}A),\ \eta_{\max} := \lambda_{\max}(\Lambda_2^{-1}B\Lambda_1^{-1}B^T), \tag{2.4.4}$$

$$\eta_{\min} := \lambda_{\min}(\Lambda_2^{-1}B\Lambda_1^{-1}B^T) \text{ and } \theta_{\max} := \lambda_{\max}(\Lambda_2^{-1}C^T\Lambda_3^{-1}C). \tag{2.4.5}$$

Theorem 2.4.5. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices and $s > 0$. Suppose λ is a real eigenvalue of the preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$. Then*

$$\lambda \in \left(0, \frac{\xi_{\max}}{1 + s\xi_{\max}}\right]. \tag{2.4.6}$$

Proof. Let $\mathbf{p} = [u^T, v^T, w^T]^T$ be an eigenvector corresponding to the real eigenvalue λ. Then, the system of linear equations in (2.4.2a)-(2.4.2c) are satisfied. First, we assume that $v = \mathbf{0}$. Subsequently, (2.4.2a) and (2.4.2c) are simplified to

$$Au = \lambda(\Lambda_1 + sA)u \quad \text{and} \quad \lambda\Lambda_3 w = \mathbf{0}, \tag{2.4.7}$$

respectively. The second equation in (2.4.7) gives $w = \mathbf{0}$, as Λ_3 is SPD. Premultiplying by u^T to the first equation of (2.4.7), we get

$$\lambda = \frac{u^TAu}{u^T\Lambda_1 u + su^TAu} = \frac{t}{1+st},$$

where $t = \dfrac{u^TAu}{u^T\Lambda_1 u} > 0$. Now, for any nonzero vector $u \in \mathbb{R}^n$, we have

$$\lambda_{\min}(\Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}}) \le \frac{u^TAu}{u^T\Lambda_1 u} = \frac{(\Lambda_1^{\frac{1}{2}}u)^T\Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}}\Lambda_1^{\frac{1}{2}}u}{(\Lambda_1^{\frac{1}{2}}u)^T\Lambda_1^{\frac{1}{2}}u} \le \lambda_{\max}(\Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}}). \tag{2.4.8}$$

Since $\Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}}$ is similar to $\Lambda_1^{-1}A$ and the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(t) := t/(1 + st),$$

where $s > 0$, is monotonically increasing in $t > 0$, we obtain the following bound for λ :

$$\frac{\xi_{\min}}{1 + s\xi_{\min}} \le \lambda \le \frac{\xi_{\max}}{1 + s\xi_{\max}}. \tag{2.4.9}$$

Next, consider $v \neq \mathbf{0}$ but $Cv = \mathbf{0}$. Then (2.4.2c) implies $w = \mathbf{0}$ and from (2.4.2b), we get

$$v = -\frac{1}{\mu}\Lambda_2^{-1}Bu. \tag{2.4.10}$$

Substituting (2.4.10) in (2.4.2a) yields

$$Au = \lambda(\Lambda_1 + sA)u + \frac{(1-s\lambda)}{\mu}B^T\Lambda_2^{-1}Bu. \tag{2.4.11}$$

Premultiplying both sides of (2.4.11) by u^T, we obtain

$$\lambda u^T Au = \lambda^2 u^T\Lambda_1 u + s\lambda^2 u^T Au + (1-s\lambda)^2 u^T B^T\Lambda_2^{-1}Bu, \tag{2.4.12}$$

which is further equivalent to

$$\lambda^2 - \lambda\frac{2sq+t}{s^2q+st+1} + \frac{q}{s^2q+st+1} = 0, \tag{2.4.13}$$

where $q = \dfrac{u^T B^T\Lambda_2^{-1}Bu}{u^T\Lambda_1 u}$. By solving (2.4.13), we obtain the following real solutions for λ:

$$\lambda^{\pm} = \frac{2sq+t \pm \sqrt{t^2-4q}}{2(s^2q+st+1)}, \tag{2.4.14}$$

where $t^2 - 4q \geq 0$. Consider the functions $\Phi_1, \Phi_2 : \mathbb{R}\times\mathbb{R}\times\mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$\Phi_1(t,q,s) = \frac{2sq+t+\sqrt{t^2-4q}}{2(s^2q+st+1)},$$
$$\Phi_2(t,q,s) = \frac{2sq+t-\sqrt{t^2-4q}}{2(s^2q+st+1)},$$

respectively, where $t, s > 0$, $q \geq 0$ and $t^2 > 4q$. Then Φ_1 and Φ_2 are strictly monotonically increasing in the argument $t > 0$ and $q \geq 0$, and decreasing in the argument $q \geq 0$ and $t > 0$, respectively. Now, using (2.4.8), (2.4.14) and monotonicity of the functions Φ_1 and Φ_2, we get the following bounds:

$$\lambda^+ = \Phi_1(t,q,s) < \Phi_1(\lambda_{\max}(\Lambda_1^{-1}A), 0, s), \tag{2.4.15}$$

$$\Phi_2(\lambda_{\max}(\Lambda_1^{-1}A), 0, s) < \lambda^- = \Phi_2(t,q,s). \tag{2.4.16}$$

Combining (2.4.15) and (2.4.16), we get the following bounds for λ:

$$0 < \lambda^{\pm} < \frac{\xi_{\max}}{1+s\xi_{\max}}. \tag{2.4.17}$$

Next, consider the case $Cv \neq \mathbf{0}$. Then from (2.4.2c), we get $w = \frac{1}{\mu}\Lambda_3^{-1}Cv$. Substituting the value of w in (2.4.2b), we obtain

$$(1-s\lambda)Bu + \frac{(1-s\lambda)}{\mu}C^T\Lambda_3^{-1}Cv = -\lambda\Lambda_2 v. \tag{2.4.18}$$

Now, we assume that

$$\lambda > \frac{\xi_{\max}}{1+s\xi_{\max}}. \tag{2.4.19}$$

By Remark 2.4.4, we get $1 - s\lambda > 0$ and the above assertion yields $\mu > \xi_{\max}$. Hence, the matrix $\mu\Lambda_1 - A$ is nonsingular and (2.4.2a) gives

$$u = (\mu\Lambda_1 - A)^{-1} B^T v. \tag{2.4.20}$$

Substituting (2.4.20) into (2.4.18), we obtain

$$(1 - s\lambda)B (\mu\Lambda_1 - A)^{-1} B^T v + \frac{(1 - s\lambda)}{\mu} C^T \Lambda_3^{-1} C v = -\lambda\Lambda_2 v. \tag{2.4.21}$$

Note that, $v \neq \mathbf{0}$. Thus, premultiplying by v^T on the both sides of (2.4.21) leads to the following identity:

$$(1 - s\lambda)v^T B (\mu\Lambda_1 - A)^{-1} B^T v + \lambda v^T \Lambda_2 v = -\frac{(1 - s\lambda)}{\mu} v^T C^T \Lambda_3^{-1} C v. \tag{2.4.22}$$

On the other hand, $\mu\Lambda_1 - A \succ \Lambda_1(\mu - \lambda_{\max}(\Lambda^{-1}A))I \succ \mathbf{0}$ and hence is a SPD matrix. Consequently, the matrix $B (\mu\Lambda_1 - A)^{-1} B^T$ is also a SPD matrix and this implies $v^T B (\mu\Lambda_1 - A)^{-1} B^T v > 0$ for all $v \neq \mathbf{0}$. This leads to a contradiction to (2.4.22) as $v^T C^T \Lambda_3^{-1} C v \geq 0$. Hence, the assumption (2.4.19) is not true and we get

$$\lambda \leq \frac{\xi_{\max}}{1 + s\xi_{\max}}. \tag{2.4.23}$$

Again, we have $\lambda > 0$ from (1) of Proposition 2.4.3. Combining the above together with (2.4.9) and (2.4.17), we obtain the desired bounds in (2.4.6) for λ. Hence, the proof is concluded. ■

Since the preconditioners $\mathscr{P}_{SS}$ and $\mathscr{P}_{EGSS}$ are special cases of the PESS preconditioner, next, we obtain refined bounds for real eigenvalues of the SS and EGSS preconditioned matrices from Theorem 2.4.5.

Corollary 2.4.1. *Let $\mathscr{P}_{SS}$ be defined as in (2.1.5) with $\alpha > 0$ and let λ be an real eigenvalue of the SS preconditioned matrix $\mathscr{P}_{SS}^{-1}\mathcal{A}$. Then*

$$\lambda \in \left(0, \frac{2\kappa_{\max}}{\alpha + \kappa_{\max}}\right], \tag{2.4.24}$$

where $\kappa_{\max} = \lambda_{\max}(A)$.

Proof. Since $\mathscr{P}_{SS}$ is a special case of the PESS preconditioner $\mathscr{P}_{PESS}$ for $s = 1/2$ as discussed in Table 2.2.1, the bounds in (2.4.24) are obtained by substituting $\Lambda_1 = \frac{1}{2}\alpha I$ and $s = 1/2$ in Theorem 2.4.5. ■

In the next result, we discuss bounds for the real eigenvalues of the EGSS preconditioned matrix $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$.

Corollary 2.4.2. *Let $\mathscr{P}_{EGSS}$ be defined as in (2.1.6) with $\alpha > 0$ and let λ be an real eigenvalue of the EGSS preconditioned matrix $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$. Then*

$$\lambda \in \left(0, \frac{2\tilde{\kappa}_{\max}}{\alpha + \tilde{\kappa}_{\max}}\right], \tag{2.4.25}$$

where $\tilde{\kappa}_{\max} = \lambda_{\max}(P^{-1}A)$.

Proof. Since $\mathscr{P}_{EGSS}$ is a special case of the PESS preconditioner $\mathscr{P}_{PESS}$ for $s = 1/2$ as discussed in Table 2.2.1 and the bounds in (2.4.25) are obtained by substituting $\Lambda_1 = \frac{1}{2}\alpha P$ and $s = 1/2$ in Theorem 2.4.5. ■

The following result shows the bounds when λ is a non-real eigenvalue.

Theorem 2.4.6. *Let $A \in \mathbb{R}^{n\times n}$ be a SPD matrix and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices. Let $s > 0$ and λ be a non-real eigenvalue of the preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ with $\Im(\lambda) \neq 0$. Suppose $\mathbf{p} = [u^T, v^T, w^T]^T$ is the eigenvector corresponding to λ. Then the following holds:*

(1) *If $Cv = 0$, then λ satisfies*

$$\frac{\xi_{\min}}{2 + s\xi_{\min}} \leq |\lambda| \leq \sqrt{\frac{\eta_{\max}}{1 + s\xi_{\min} + s^2\eta_{\max}}}.$$

(2) *If $Cv \neq 0$, then real and imaginary part of $\lambda/(1 - s\lambda)$ satisfies*

$$\frac{\xi_{\min}\,\eta_{\min}}{2\,(\xi_{\max}^2 + \eta_{\max} + \theta_{\max})} \leq \Re\left(\frac{\lambda}{1 - s\lambda}\right) \leq \frac{\xi_{\max}}{2} \text{ and } \left|\Im\left(\frac{\lambda}{1 - s\lambda}\right)\right| \leq \sqrt{\eta_{\max} + \theta_{\max}},$$

where $\xi_{\max}, \xi_{\min}, \eta_{\max}, \eta_{\min}$ and $\theta_{\max}$ are defined as in (2.4.4) and (2.4.5).

Proof. (1) Let λ be an eigenvalue of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ with $\Im(\lambda) \neq 0$ and the corresponding eigenvector is $\mathbf{p} = [u^T, v^T, w^T]^T$. Then, the system of linear equations in (2.4.2a)-(2.4.2c) holds. We assert that, $v \neq \mathbf{0}$, otherwise from (2.4.2a), we get $Au = \lambda(\Lambda_1 + sA)u$, which further implies $(\Lambda_1 + sA)^{-\frac{1}{2}}A(\Lambda_1 + sA)^{-\frac{1}{2}}\bar{u} = \lambda\bar{u}$, where $\bar{u} = (\Lambda_1 + sA)^{\frac{1}{2}}u$. Since $(\Lambda_1 + sA)^{-\frac{1}{2}}A(\Lambda_1 + sA)^{-\frac{1}{2}}$ is a SPD matrix and by Lemma 2.4.3, $u \neq \mathbf{0}$ and thus we have $\bar{u} \neq \mathbf{0}$. This implies λ is real, leading to a contradiction to $\Im(\lambda) \neq 0$.

Initially, we consider the case $Cv = \mathbf{0}$. The aforementioned discussion suggests that $Au \neq \lambda(\Lambda_1 + sA)u$ for any $u \neq \mathbf{0}$. Consequently, $(\mu\Lambda_1 - A)$ is nonsingular. Again, since $Cv = \mathbf{0}$, from (2.4.2c) and following a similar method as in (2.4.10)-(2.4.12), we obtain quadratic equation (2.4.13) in λ, which has complex solutions:

$$\lambda^{\pm i} = \frac{2sq + t \pm i\sqrt{4q - t^2}}{2(s^2q + st + 1)} \tag{2.4.26}$$

for $4q > t^2$. From (2.4.26), we get

$$|\lambda^{\pm i}|^2 = \frac{q}{s^2q + st + 1}. \tag{2.4.27}$$

Now, consider the function $\Psi_1 : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$\Psi_1(t, q, s) = \frac{q}{s^2q + st + 1},$$

where $t, q, s > 0$. Then Ψ_1 is monotonically increasing in the argument $q > 0$ while decreasing in the argument $t > 0$. Due to the fact that $q \leq \lambda_{\max}(\Lambda_1^{-1}B^T\Lambda_2^{-1}B) = \eta_{\max}$, we get the following bounds:

$$|\lambda^{\pm i}|^2 = \Psi_1(t, q, s) \leq \Psi_1(\xi_{\min}, \eta_{\max}, s) = \frac{\eta_{\max}}{1 + s\xi_{\min} + s^2\eta_{\max}}. \tag{2.4.28}$$

Again, considering the condition $q > (t^2/4)$, (2.4.27) adheres to the inequality $(t/(2 + st)) \leq |\lambda^{\pm i}|$. Notably, the function $\Psi_2 : \mathbb{R} \longrightarrow \mathbb{R}$ defined by $\Psi_2(t) = t/(2 + st)$, where $t, s > 0$, is monotonically increasing in the argument $t > 0$, we obtain the following bound:

$$|\lambda^{\pm i}| = \Psi_2(t) \geq \Psi_2(\xi_{\min}) = \frac{\xi_{\min}}{2 + s\xi_{\min}}. \tag{2.4.29}$$

Hence, combining (2.4.28) and (2.4.29), proof for the part (1) follows.

(2) Next, consider the case $Cv \neq \mathbf{0}$, then (2.4.2c) yields $w = \frac{1}{\mu}\Lambda_3^{-1}Cv$, where $\mu = \lambda/(1 - s\lambda)$. Substituting this in (2.4.2b), we get

$$Bu + \frac{1}{\mu}C^T\Lambda_3^{-1}Cv = -\mu\Lambda_2 v. \tag{2.4.30}$$

On the other side, nonsingularity of $(\mu\Lambda_1 - A)$ and (2.4.2a) enables us the following identity:

$$u = (\mu\Lambda_1 - A)^{-1}B^T v. \tag{2.4.31}$$

Putting (2.4.31) on (2.4.30), we get

$$B(\mu\Lambda_1 - A)^{-1}B^T v + \frac{1}{\mu}C^T\Lambda_3^{-1}Cv = -\mu\Lambda_2 v. \tag{2.4.32}$$

As $v \neq \mathbf{0}$, premultiplying by v^H on the both sides of (2.4.32) yields

$$v^H B(\mu\Lambda_1 - A)^{-1}B^T v + \frac{1}{\mu}v^H C^T\Lambda_3^{-1}Cv = -\mu v^H\Lambda_2 v. \tag{2.4.33}$$

Consider the following eigenvalue decomposition of $\Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}}$:

$$\Lambda_1^{-\frac{1}{2}}A\Lambda_1^{-\frac{1}{2}} = V\mathcal{D}V^T,$$

where $\mathcal{D} = \operatorname{diag}(\theta_i) \in \mathbb{R}^{n\times n}$, $\theta_i \in \sigma(\Lambda_1^{-\frac{1}{2}} A \Lambda_1^{-\frac{1}{2}})$ with $\theta_i > 0$, for $i = 1, 2, \ldots, n$, and $V \in \mathbb{R}^{n\times n}$ is an orthonormal matrix. Then, (2.4.33) can be rewritten as

$$v^H B\Lambda_1^{-\frac{1}{2}} V(\mu I - \mathcal{D})^{-1} V^T \Lambda_1^{-\frac{1}{2}} B^T v + \frac{1}{\mu} v^H C^T \Lambda_3^{-1} C v = -\mu v^H \Lambda_2 v. \tag{2.4.34}$$

Since $(\mu I - \mathcal{D})^{-1} = \Theta_1 - \boldsymbol{i}\Im(\mu)\Theta_2$, where $\boldsymbol{i}$ denotes the imaginary unit and

$$\Theta_1 = \operatorname{diag}\left(\frac{\Re(\mu) - \theta_i}{(\Re(\mu) - \theta_i)^2 + \Im(\mu)^2}\right), \ \Theta_2 = \operatorname{diag}\left(\frac{1}{(\Re(\mu) - \theta_i)^2 + \Im(\mu)^2}\right),$$

from (2.4.34) and the fact $\Im(\mu) \neq 0$, we get

$$v^H B\Lambda_1^{-\frac{1}{2}} V\Theta_1 V^T \Lambda_1^{-\frac{1}{2}} B^T v + \frac{\Re(\mu)}{|\mu|^2} v^H C^T \Lambda_3^{-1} C v = -\Re(\mu) v^H \Lambda_2 v, \tag{2.4.35}$$

$$v^H B\Lambda_1^{-\frac{1}{2}} V\Theta_2 V^T \Lambda_1^{-\frac{1}{2}} B^T v + \frac{1}{|\mu|^2} v^H C^T \Lambda_3^{-1} C v = v^H \Lambda_2 v. \tag{2.4.36}$$

Observed that, $\frac{1}{\Im(\mu)^2} I \succeq \Theta_2$, then from (2.4.36), we obtain

$$1 \leq \frac{1}{\Im(\mu)^2}\left(\frac{v^H B\Lambda_1^{-1} B^T v}{v^H \Lambda_2 v} + \frac{v^H C^T \Lambda_3^{-1} C v}{v^H \Lambda_2 v}\right).$$

Therefore, we get

$$|\Im(\mu)| \leq \sqrt{\eta_{\max} + \theta_{\max}}. \tag{2.4.37}$$

Furthermore, notice that $\Theta_1 \succeq (\Re(\mu) - \xi_{\max})\,\Theta_2$, then from (2.4.35), we deduce

$$-\Re(\mu) \geq (\Re(\mu) - \xi_{\max}) \frac{v^H B\Lambda_1^{-\frac{1}{2}} V\Theta_2 V^T \Lambda_1^{-\frac{1}{2}} B^T v}{v^H \Lambda_2 v} + \frac{\Re(\mu)}{|\mu|^2} \frac{v^H C^T \Lambda_3^{-1} C v}{v^H \Lambda_2 v}. \tag{2.4.38}$$

Using (2.4.36) to (2.4.38), we get

$$2\Re(\mu) \leq \xi_{\max} - \frac{\xi_{\max}}{|\mu|^2} \frac{v^H C^T \Lambda_3^{-1} C v}{v^H \Lambda_2 v} \leq \xi_{\max}. \tag{2.4.39}$$

Hence, (2.4.39) yields the following bound:

$$\Re(\mu) \leq \frac{\xi_{\max}}{2}. \tag{2.4.40}$$

On the other side, from (2.4.35) and (2.4.36), we deduce that

$$2\Re(\mu) = \frac{v^H B\Lambda_1^{-\frac{1}{2}} V(\Re(\mu)\Theta_2 - \Theta_1) V^T \Lambda_1^{-\frac{1}{2}} B^T v}{v^H \Lambda_2 v} \geq \xi_{\min} \frac{v^H B\Lambda_1^{-\frac{1}{2}} V \mathcal{X} V^T \Lambda_1^{-\frac{1}{2}} B^T v}{v^H \Lambda_2 v}, \tag{2.4.41}$$

where $\mathcal{X} = \operatorname{diag}\left(\frac{1}{(\Re(\mu) - \theta_i)^2 + \Im(\mu)^2}\right)$.

Note that

$$\frac{1}{(\Re(\mu) - \theta_i)^2 + \Im(\mu)^2} \geq \frac{1}{\max\limits_i (\Re(\mu) - \theta_i)^2 + \Im(\mu)^2}.$$

Now, using Proposition 2.4.3 and (2.4.40), we get

$$-\theta_i < \Re(\mu) - \theta_i \leq \frac{\xi_{\max}}{2} - \theta_i,$$

which further yields $(\Re(\mu) - \theta_i)^2 \leq \max\left\{\left(\frac{\xi_{\max}}{2} - \theta_i\right)^2, \theta_i^2\right\}$. Hence, we get

$$\begin{aligned}\max_i (\Re(\mu) - \theta_i)^2 &= \max_i \{(\Re(\mu) - \xi_{\min})^2, (\Re(\mu) - \xi_{\max})^2\} \\ &\leq \max_i \left\{\left(\frac{\xi_{\max}}{2} - \xi_{\min}\right)^2, \xi_{\max}^2\right\} = \xi_{\max}^2. \end{aligned} \tag{2.4.42}$$

Therefore, by (2.4.42) and from the bound in (2.4.37), we obtain

$$\mathcal{X} \succeq \frac{I}{\xi_{\max}^2 + \eta_{\max} + \theta_{\max}}. \tag{2.4.43}$$

Combining (2.4.41) and (2.4.43) leads to the following bounds:

$$2\Re(\mu) \geq \frac{\xi_{\min}}{\xi_{\max}^2 + \eta_{\max} + \theta_{\max}} \frac{v^H B \Lambda_1^{-1} B^T v}{v^H \Lambda_2 v} \geq \frac{\xi_{\min}\eta_{\min}}{\xi_{\max}^2 + \eta_{\max} + \theta_{\max}}. \tag{2.4.44}$$

Hence, the proof of part (2) follows by merging the two inequalities of (2.4.37) and (2.4.44). ■

Remark 2.4.7. *From the bounds in* (2) *of Theorem 2.4.6, we obtain*

$$\frac{1}{|\tau|^2}\left(\frac{\xi_{\min}\eta_{\min}}{2(\xi_{\max}^2 + \eta_{\max} + \theta_{\max})} + \frac{1}{s}\right) \leq \frac{1}{s} - \Re(\lambda) \leq \frac{1}{|\tau|^2}\left(\frac{\xi_{\max}}{2} + \frac{1}{s}\right)$$

and

$$|\Im(\lambda)| \leq \frac{1}{s\tau\bar{\tau}}\sqrt{\eta_{\max} + \theta_{\max}},$$

where $\tau = s\mu + 1$. *Thus, the real and imaginary parts of the eigenvalues of the preconditioned matrix* $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ *cluster better as* s *grows. Therefore, the PESS preconditioner can accelerate the rate of convergence of the Krylov subspace method, like GMRES.*

Utilizing the established bounds in Theorem 2.4.6, we derive the subsequent estimations for non-real eigenvalues of SS and EGSS preconditioned matrices.

Corollary 2.4.3. *Let* $\mathscr{P}_{SS}$ *be defined as in* (2.1.5) *with* $\alpha > 0$ *and let* λ *be a non-real eigenvalue of the SS preconditioned matrix* $\mathscr{P}_{SS}^{-1}\mathcal{A}$ *with* $\Im(\lambda) \neq 0$ *and* $\mathbf{p} = [u^T, v^T, w^T]^T$ *is the corresponding eigenvector. Then,*

(1) *If* $Cv = \mathbf{0}$, λ *satisfies*

$$\frac{2\kappa_{\min}}{2\alpha + \kappa_{\min}} \leq |\lambda| \leq \sqrt{\frac{4\tau_{\max}}{\alpha^2 + \alpha\kappa_{\min} + \tau_{\max}}}.$$

(2) *If* $Cv \neq \mathbf{0}$, *the real and imaginary part of* $\lambda/(2-\lambda)$ *satisfies*

$$\frac{\kappa_{\min}\,\tau_{\min}}{2\alpha\,(\kappa_{\max}^2+\tau_{\max}+\beta_{\max})} \leq \Re\left(\frac{\lambda}{2-\lambda}\right) \leq \frac{\kappa_{\max}}{2\alpha} \text{ and } \left|\Im\left(\frac{\lambda}{2-\lambda}\right)\right| \leq \frac{\sqrt{\tau_{\max}+\beta_{\max}}}{\alpha},$$

where $\kappa_{\min}=\lambda_{\min}(A)$, $\tau_{\max}=\lambda_{\max}(BB^T)$, $\tau_{\min}=\lambda_{\min}(BB^T)$ *and* $\beta_{\max}=\lambda_{\max}(C^TC)$.

Proof. Since $\mathscr{P}_{SS}$ is a special case of $\mathscr{P}_{PESS}$ for $s=1/2$ as discussed in Table 2.2.1, then desired bounds will be obtained by setting $\Lambda_1=\Lambda_2=\Lambda_3=\frac{1}{2}\alpha I$ and $s=1/2$ into Theorem 2.4.6. ■

Corollary 2.4.4. *Let* $\mathscr{P}_{EGSS}$ *be defined as in* (2.1.6) *and let* P,Q *and* W *are SPD matrices with* $\alpha,\beta,\gamma>0$. *Let* λ *be a non-real eigenvalue of the EGSS preconditioned matrix* $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$ *with* $\Im(\lambda)\neq 0$ *and* $\mathbf{p}=[u^T,v^T,w^T]^T$ *is the corresponding eigenvector. Then*

(1) *If* $Cv=\mathbf{0}$, λ *satisfies*

$$\frac{2\tilde{\kappa}_{\min}}{2\alpha+\tilde{\kappa}_{\min}} \leq |\lambda| \leq \sqrt{\frac{4\tilde{\tau}_{\max}}{\alpha\beta+\beta\tilde{\kappa}_{\min}+\tilde{\tau}_{\max}}}.$$

(2) *If* $Cv\neq\mathbf{0}$, *the real and imaginary part of* $\lambda/(2-\lambda)$ *satisfies*

$$\frac{\tilde{\kappa}_{\min}\,\tilde{\tau}_{\min}}{2\left(\beta\tilde{\kappa}_{\max}^2+\alpha\tilde{\tau}_{\max}+(\alpha^2/\gamma)\tilde{\beta}_{\max}\right)} \leq \Re\left(\frac{\lambda}{2-\lambda}\right) \leq \frac{\tilde{\kappa}_{\max}}{2\alpha} \text{ and}$$

$$\left|\Im\left(\frac{\lambda}{2-\lambda}\right)\right| \leq \sqrt{\frac{1}{\alpha\beta}\tilde{\tau}_{\max}+\frac{1}{\beta\gamma}\tilde{\beta}_{\max}},$$

where $\tilde{\kappa}_{\min}=\lambda_{\min}(P^{-1}A)$, $\tilde{\tau}_{\max}=\lambda_{\max}(Q^{-1}BP^{-1}B^T)$, $\tilde{\tau}_{\min}=\lambda_{\min}(Q^{-1}BP^{-1}B^T)$ *and* $\tilde{\beta}_{\max}=\lambda_{\max}(C^TC)$.

Proof. Since $\mathscr{P}_{EGSS}$ is a special case of $\mathscr{P}_{PESS}$ for $s=1/2$ as discussed in Table 2.2.1, then the desired bounds will be obtained by setting $\Lambda_1=\frac{1}{2}\alpha P$, $\Lambda_2=\frac{1}{2}\beta Q$, $\Lambda_3=\frac{1}{2}\gamma W$ and $s=1/2$ in Theorem 2.4.6. ■

2.5. Local PESS (LPESS) Preconditioner

To enhance the efficiency of the PESS preconditioner, in this section, we propose a relaxed version of the PESS preconditioner by incorporating the concept of RSS preconditioner [37]. By omitting the term Λ_1 from the $(1,1)$-block of $\mathscr{P}_{PESS}$, we present the

local PESS (LPESS) preconditioner, denoted as $\mathscr{P}_{LPESS}$, defined as follows:

$$\mathscr{P}_{LPESS} := \begin{bmatrix} sA & sB^T & \mathbf{0} \\ -sB & \Lambda_2 & -sC^T \\ \mathbf{0} & sC & \Lambda_3 \end{bmatrix}. \tag{2.5.1}$$

The implementation of the LPESS preconditioner is similar to the Algorithm 2.2.1. However, there is one modification in step 3, i.e., we need to solve a linear subsystem $(sA + s^2B^T\widehat{X}^{-1}B)w_1 = v$ instead of $\widetilde{A}w_1 = v$.

To illustrate the efficiency of the LPESS preconditioner, we study the spectral distribution of the preconditioned matrix $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$. To achieve this, we consider the following decomposition of the matrices $\mathcal{A}$ and $\mathscr{P}_{LPESS}$ as follows:

$$\mathcal{A} = \mathfrak{L}\mathfrak{D}\mathfrak{U} \text{ and } \mathscr{P}_{LPESS} = \mathfrak{L}\widetilde{\mathfrak{D}}\mathfrak{U}, \tag{2.5.2}$$

where

$$\mathfrak{L} = \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ -BA^{-1} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix}, \ \mathfrak{U} = \begin{bmatrix} I & A^{-1}B^T & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix},$$

$$\mathfrak{D} = \begin{bmatrix} A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & Q & -C^T \\ \mathbf{0} & C & \mathbf{0} \end{bmatrix}, \ \widetilde{\mathfrak{D}} = \begin{bmatrix} sA & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Lambda_2 + sQ & -sC^T \\ \mathbf{0} & sC & \Lambda_3 \end{bmatrix}$$

and $Q = BA^{-1}B^T$. Using the decomposition in (2.5.2), we have

$$\mathscr{P}_{LPESS}^{-1}\mathcal{A} = \mathfrak{U}^{-1}\begin{bmatrix} s^{-1}I & \mathbf{0} \\ \mathbf{0} & M^{-1}K \end{bmatrix}\mathfrak{U}, \tag{2.5.3}$$

where $M = \begin{bmatrix} \Lambda_2 + sQ & -sC^T \\ sC & \Lambda_3 \end{bmatrix}$ and $K = \begin{bmatrix} Q & -C^T \\ C & \mathbf{0} \end{bmatrix}$. Since, $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ is similar to $\begin{bmatrix} s^{-1}I & \mathbf{0} \\ \mathbf{0} & M^{-1}K \end{bmatrix}$, $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ has eigenvalue $1/s$ with multiplicity n and the remaining eigenvalues satisfies the generalized eigenvalue problem $Kp = \lambda Mp$.

Theorem 2.5.1. *Let $A \in \mathbb{R}^{n\times n}, \Lambda_2 \in \mathbb{R}^{m\times m}$ and $\Lambda_3 \in \mathbb{R}^{p\times p}$ be SPD matrices and let $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ be full row rank matrices. Assume that $s > 0$, then LPESS preconditioned matrix $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ has n repeated eigenvalues equal to $1/s$. The remaining $m + p$ eigenvalues satisfies the following:*

(1) *real eigenvalues located in the interval*

$$\Big[\min\Big\{\frac{\vartheta_{\min}}{1+s\vartheta_{\min}}, \frac{\tilde{\theta}_{\min}}{\vartheta_{\max}+s\tilde{\theta}_{\min}}\Big\}, \frac{\vartheta_{\max}}{1+s\vartheta_{\max}}\Big].$$

(2) *If λ is any non-real eigenvalue (i.e., $\Im(\lambda) \neq 0$), then*

$$\frac{\vartheta_{\min}}{2+s\vartheta_{\min}} \leq |\lambda| \leq \sqrt{\frac{\tilde{\theta}_{\max}}{1+s\vartheta_{\min}+s^2\tilde{\theta}_{\max}}} \text{ and } \frac{1}{s(1+s\sqrt{\tilde{\theta}_{\max}})} \leq |\lambda - \frac{1}{s}| \leq \frac{2}{s(2+s\vartheta_{\min})},$$

where $\vartheta_{\max} := \lambda_{\max}(\Lambda_2^{-1}Q)$, $\vartheta_{\min} := \lambda_{\min}(\Lambda_2^{-1}Q)$ $\tilde{\theta}_{\max} := \lambda_{\max}(\Lambda_3^{-1}C\Lambda_2^{-1}C^T)$ *and* $\tilde{\theta}_{\min} := \lambda_{\min}(\Lambda_3^{-1}C\Lambda_2^{-1}C^T)$.

Proof. Observe that,

$$M^{-1}K = \mathcal{F}^{-1}\widetilde{M}^{-1}\widetilde{K}\mathcal{F}, \tag{2.5.4}$$

where $\mathcal{F} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & -I \end{bmatrix}$, $\widetilde{M} = \begin{bmatrix} \Lambda_2 + sQ & sC^T \\ -sC & \Lambda_3 \end{bmatrix}$ and $\widetilde{K} = \begin{bmatrix} Q & C^T \\ -C & \mathbf{0} \end{bmatrix}$. Hence, $M^{-1}K$ is similar to $\widetilde{M}^{-1}\widetilde{K}$. Now, $\widetilde{M}^{-1}\widetilde{K}$ is in the form of the preconditioned matrix discussed in [135]. Therefore, by applying the Corollary 4.1 and Theorem 4.3 of [135], we obtain the desired bounds in (1) and (2). ■

2.6. The Strategy of Parameter Selection

It is worth noting that the efficiency of the proposed PESS preconditioner depends on the selection of involved SPD matrices Λ_1, Λ_2, Λ_3, and a positive real parameter $s > 0$. In this section, motivated by [135, 79, 80], we discuss a practical way to choose the parameter s and the SPD matrices for the effectiveness of the proposed preconditioners.

It is well acknowledged that the optimal parameter of the PESS iteration method is obtained when $\vartheta(\mathcal{T})$ is minimized [79]. For achieving this, similar to the approach in [80], we first define a function $\varphi(s) = \|\mathcal{Q}_{PESS}\|_F^2$ depending on the parameter s. Our aim is to minimize $\varphi(s)$. Then, after some straightforward calculations, we obtain

$$\begin{aligned}\varphi(s) = \|\mathcal{Q}_{PESS}\|_F^2 &= \mathtt{tr}(\mathcal{Q}_{PESS}^T\mathcal{Q}_{PESS}) \\ &= \|\Lambda_1\|_F^2 + \|\Lambda_2\|_F^2 + \|\Lambda_3\|_F^2 + (s-1)^2\|A\|_F^2 + 2(s-1)\mathtt{tr}(\Lambda_1 A) \\ &\quad + 2(s-1)^2\|B\|_F^2) + 2(s-1)^2\|C\|_F^2).\end{aligned}$$

Now we choose the parameter s and SPD matrices Λ_1, Λ_2 and Λ_3 to make $\varphi(s)$ as small as possible. Since $\|A\|_F^2, \|B\|_F^2, \|C\|_F^2$ and $\mathtt{tr}(\Lambda_1 A)$ are positive, we can select $s = 1$, then

$\varphi(s) = \|\Lambda_1\|_F^2 + \|\Lambda_2\|_F^2 + \|\Lambda_3\|_F^2$. Thus if we choose $\|\Lambda_1\|_F, \|\Lambda_2\|_F, \|\Lambda_3\|_F \to 0$, we have $\varphi(s) \to 0$ and consequently, $\mathcal{Q}_{PESS} \to \mathbf{0}$.

On the other hand, motivated by [144, 124], we discuss another strategy for choosing the parameter s. Notice that in Algorithm 2.2.1, we need to solve two linear system with coefficient matrices $\widehat{X} = \Lambda_2 + s^2 C^T \Lambda_3^{-1} C$ and $\widetilde{A} = \Lambda_1 + sA + s^2 B^T \widehat{X} B$. Similar to [144, 124], we choose s and $\|\Lambda_2\|_2$ as follows:

$$s = \sqrt{\frac{\|\Lambda_2\|_2}{\|C^T\Lambda_3^{-1}C\|_2}} \quad \text{and} \quad \|\Lambda_2\|_2 = \frac{\|B\|_2^4}{4\|C^T\Lambda_3^{-1}C\|_2\|A\|_2^2}, \tag{2.6.1}$$

which balance the matrices Λ_2 and $C^T\Lambda_3^{-1}C$ in $\widehat{X}$ and the matrices A and $C^T\Lambda_3^{-1}C$ in $\widetilde{A}$. In this case, we denote s by s_{est} and $\|\Lambda_2\|_2$ by β_{est}. Numerical results are presented in Section 2.7 to demonstrate the effectiveness of $\mathscr{P}_{PESS}$ for the above choices of the parameters.

In the sequel, we can rewrite the PESS preconditioner as $\mathscr{P}_{PESS} = s\widetilde{\mathscr{P}}_{PESS}$, where

$$\widetilde{\mathscr{P}}_{PESS} = \begin{bmatrix} \frac{1}{s}\Lambda_1 + A & B^T & \mathbf{0} \\ -B & \frac{1}{s}\Lambda_2 & -C^T \\ \mathbf{0} & C & \frac{1}{s}\Lambda_3 \end{bmatrix}.$$

Since the prefactor s has not much effect on the performance of PESS preconditioner, investigating the optimal parameters of $\mathscr{P}_{PESS}$ and $\widetilde{\mathscr{P}}_{PESS}$ are equivalent. A general criterion for a preconditioner to perform efficiently is that it should closely approximate the coefficient matrix $\mathcal{A}$ [26]. Consequently, the difference $\widetilde{\mathscr{P}}_{PESS} - \mathcal{A} = \frac{1}{s}\Sigma$ approaches zero matrix as s tends to positive infinity for fixed Λ_1, Λ_2 and Λ_3. Thus, the preconditioner PESS shows enhance efficiency for large values of s. However, s can not be too large as the coefficient matrix $\widetilde{A}$ of the linear subsystem in step 3 of Algorithm 2.2.1 becomes very-ill conditioned. Similar investigations also hold for LPESS preconditioner. Nevertheless, in Figure 2.7.9, we show the adaptability of the PESS and LPESS preconditioners by varying the parameter s.

2.7. Numerical Experiments

In this section, we conduct a few numerical experiments to showcase the superiority and efficiency of the proposed PESS and LPESS preconditioners over the existing preconditioners to enhance the convergence speed of the Krylov subspace iterative method to solve DSPPs. Our study involves a comparative analysis among the GMRES method and the PGMRES method employing the proposed preconditioners $\mathscr{P}_{PESS}$ and $\mathscr{P}_{LPESS}$ and

the existing baseline preconditioners $\mathscr{P}_{BD}$, $\mathscr{P}_{IBD}$, $\mathscr{P}_{MAPSS}$, $\mathscr{P}_{SL}$, $\mathscr{P}_{SS}$, $\mathscr{P}_{RSS}$, $\mathscr{P}_{EGSS}$ and $\mathscr{P}_{RPGSS}$. The numerical results are reported from the aspect of iteration counts (abbreviated as "IT") and elapsed CPU times in seconds (abbreviated as "CPU"). Each subsystem involving $\widehat{X}$ and $\widetilde{A}$ featured in **Algorithm 2.2.1** are precisely solved by applying the Cholesky factorization of the coefficient matrices. For all iterative method, the initial guess vector is $\mathbf{u}_0 = \mathbf{0} \in \mathbb{R}^{n+m+p}$ and the termination criterion is

$$\texttt{RES} := \frac{\|\mathcal{A}\mathbf{u}^{k+1} - \widehat{\mathbf{d}}\|_2}{\|\widehat{\mathbf{d}}\|_2} < 10^{-6}.$$

The vector $\mathbf{d} \in \mathbb{R}^{n+m+p}$ is chosen so that the exact solution of the system (2.1.1) is $\mathbf{u}_* = [1, 1, \ldots, 1]^T \in \mathbb{R}^{n+m+p}$. All numerical tests are run in MATLAB (version R2023a) on a Windows 11 operating system with Intel(R) Core(TM) i7-10700 CPU, 2.90GHz, 16 GB memory.

Example 2.7.1. Problem formulation: We consider the DSPP (2.1.1) taken from [75] with $A = \begin{bmatrix} I \otimes G + G \otimes I & \mathbf{0} \\ \mathbf{0} & I \otimes G + G \otimes I \end{bmatrix} \in \mathbb{R}^{2l^2 \times 2l^2}$, $B = \begin{bmatrix} I \otimes F & F \otimes I \end{bmatrix} \in \mathbb{R}^{l^2 \times 2l^2}$, $C = E \otimes F \in \mathbb{R}^{l^2 \times l^2}$, where $G = \frac{1}{(l+1)^2}\,\texttt{tridiag}(-1, 2, -1) \in \mathbb{R}^{l \times l}$, $F = \frac{1}{l+1}\,\texttt{tridiag}(0, 1, -1) \in \mathbb{R}^{l \times l}$ and $E = \mathrm{diag}(1, l+1, \ldots, l^2 - l + 1) \in \mathbb{R}^{l \times l}$. For this problem, the size of the matrix $\mathcal{A}$ is $4l^2$.

Parameter selection: Following [75], selection of $\widehat{A}$ and $\widehat{S}$ (SPD approximations of A and S, respectively) for IBD preconditioner are done as follows:

$$\widehat{A} = LL^T, \quad \widehat{S} = \mathrm{diag}(B\widehat{A}^{-1}B^T),$$

where L is the incomplete Cholesky factor of A produced by the Matlab function:

ichol(A, struct('type', 'ict', 'droptol', 1e-8, 'michol', 'off')).

For the preconditioner $\mathscr{P}_{MAPSS}$, we take $\alpha = \sqrt[4]{\frac{\mathtt{tr}(BB^TC^TC)}{m}}$ and $\beta = 10^{-4}$ as in [43]. We consider the parameter choices for the preconditioners $\mathscr{P}_{SS}$, $\mathscr{P}_{RSS}$, $\mathscr{P}_{EGSS}$, $\mathscr{P}_{RPGSS}$, $\mathscr{P}_{PESS}$ and $\mathscr{P}_{LPESS}$ in the following two cases.

- In Case I: $\alpha = 0.1$ for $\mathscr{P}_{SS}$ and $\mathscr{P}_{RSS}$; $\alpha = 0.1, \beta = 1, \gamma = 0.001$ and $P = I, Q = I, W = I$ for $\mathscr{P}_{EGSS}$ and $\mathscr{P}_{RPGSS}$; and $\Lambda_1 = I, \Lambda_2 = I, \Lambda_3 = 0.001I$ for $\mathscr{P}_{PESS}$ and $\mathscr{P}_{LPESS}$.

- In Case II: $\alpha = 1$ for $\mathscr{P}_{SS}$ and $\mathscr{P}_{RSS}$; $\alpha = 1, \beta = 1, \gamma = 0.001$ and $P = A, Q = I, W = CC^T$ for $\mathscr{P}_{EGSS}$ and $\mathscr{P}_{RPGSS}$; and $\Lambda_1 = A, \Lambda_2 = I, \Lambda_3 = 0.001CC^T$ for $\mathscr{P}_{PESS}$ and $\mathscr{P}_{LPESS}$.

The parameter selection in Case II is made as in [156].

Table 2.7.1: Numerical results of GMRES, BD, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS and LPESS PGMRES methods for Example 2.7.1.

method	l	16	32	48	64	80	128
	size($\mathcal{A}$)	1024	4096	9216	16384	25600	65536
GMRES	IT	865	3094	6542	$--$	$--$	$--$
	CPU	0.6607	101.7659	989.0242	2116.1245	3755.9251	8694.5762
	RES	$8.2852e-07$	$9.9189e-07$	$9.8389e-07$	$1.0813e-03$	$1.9862e-03$	0.0037
BD	IT	4	4	4	4	4	4
	CPU	**0.0611**	2.3135	17.0971	105.4514	337.4322	1355.6953
	RES	$1.2728e-13$	$1.7577e-13$	$6.6653e-13$	$6.0211e-12$	$2.6610e-12$	$6.4572e-07$
IBD	IT	22	22	21	21	21	27
	CPU	**0.0702**	1.1711	9.9850	44.2197	141.1108	3064.8487
	RES	$4.0221e-07$	$4.6059e-07$	$9.1711e-07$	$7.8518e-07$	$6.8357e-07$	$9.8667e-07$
MAPSS	IT	5	5	6	6	6	7
	CPU	0.21299	1.2038	8.0537	36.5074	116.7495	514.1132
	RES	$4.6434e-07$	$9.0780e-07$	$3.8651e-07$	$3.8350e-07$	$4.3826e-07$	$2.4983e-07$
SL	IT	6	6	5	5	5	4
	CPU	0.19624	1.0819	6.3741	30.6976	102.0981	481.7656
	RES	$2.5612e-08$	$7.4194e-08$	$9.4761e-08$	$3.7303e-09$	$1.7982e-09$	$6.1568e-08$
				Case I			
SS	IT	4	4	4	4	4	4
	CPU	0.2415	1.3375	6.7059	33.6902	111.3118	439.1221
	RES	$7.7528e-08$	$5.5120e-08$	$4.5134e-08$	$3.9157e-08$	$3.5073e-08$	$2.7836e-08$
RSS	IT	4	4	4	4	4	4
	CPU	0.2776	1.0467	6.5207	32.4428	110.4587	432.6755
	RES	$8.0898e-08$	$6.0033e-08$	$4.9839e-08$	$4.3510e-08$	$3.9097e-08$	$3.1111e-08$
EGSS	IT	4	4	4	4	4	4
	CPU	0.3061	1.2771	6.9552	32.8960	119.7763	435.0198
	RES	$5.9583e-10$	$4.4128e-10$	$3.6626e-10$	$3.2009e-10$	$2.8807e-10$	$1.9145e-08$
RPGSS	IT	4	4	4	4	4	3
	CPU	0.2249	1.1312	7.0322	35.0140	130.2593	371.4522
	RES	$5.9497e-10$	$4.4042e-10$	$3.6494e-10$	$3.1838e-10$	$2.8603e-10$	$9.9326e-07$
PESS†	IT	**2**	**2**	**2**	**2**	**2**	**2**
	CPU	0.2285	**0.9854**	**4.8795**	**23.3687**	**79.9355**	**283.8561**
$s = 12$	RES	$3.1630e-07$	$2.3239e-07$	$1.9486e-07$	$1.7260e-07$	$1.5754e-07$	$1.3135e-07$

Table 2.7.1: Numerical results of GMRES, BD, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS, and LPESS PGMRES methods for Example 2.7.1 (continued).

method	l	16	32	48	64	80	128
	IT	**2**	**2**	**2**	**2**	**2**	**2**
LPESS†	CPU	0.2834	**0.7989**	**4.5599**	**21.1664**	**72.8376**	**268.3114**
$s = 12$	RES	$3.1180e-07$	$2.2463e-07$	$1.8481e-07$	$1.6071e-07$	$1.4411e-07$	$1.1441e-07$
				Case II			
	IT	7	7	7	7	7	7
SS	CPU	0.2780	1.6189	10.6707	54.4493	184.6025	703.1082
	RES	$6.7967e-07$	$4.9738e-07$	$4.1383e-07$	$3.6189e-07$	$3.2556e-07$	$2.5956e-07$
	IT	7	7	7	7	7	7
RSS	CPU	0.3248	1.4644	10.4985	52.4559	177.1895	690.1611
	RES	$5.2397e-07$	$3.6832e-07$	$2.7532e-07$	$2.1869e-07$	$1.8240e-07$	$1.2720e-07$
	IT	5	5	4	4	4	4
EGSS	CPU	0.2588	1.3792	6.7546	34.8358	118.2298	439.7907
	RES	$6.8491e-08$	$1.5239e-07$	$9.7099e-07$	$6.3466e-07$	$4.8159e-07$	$4.1376e-07$
	IT	4	4	4	4	4	3
RPGSS	CPU	0.2445	1.2987	8.1982	28.5129	104.9127	397.4766
	RES	$5.2642e-08$	$1.1366e-07$	$7.1116e-08$	$8.2780e-07$	$5.6396e-07$	$2.5131e-07$
	IT	3	3	3	3	3	3
PESS†	CPU	0.2971	1.1597	6.6682	31.2670	103.8616	371.1214
$s = 12$	RES	$7.4100e-08$	$7.4642e-08$	$7.3327e-08$	$6.9803e-08$	$6.1920e-08$	$4.0967e-08$
	IT	3	3	3	3	3	3
LPESS†	CPU	0.2226	0.9905	6.1339	30.0975	98.6443	359.8141
$s = 12$	RES	$1.1683e-09$	$2.3215e-09$	$3.1540e-09$	$3.3907e-09$	$3.0271e-09$	$1.6141e-09$

Here † represents the proposed preconditioners. The boldface represents the top two results.

$--$ indicates that the iteration method does not converge within the prescribed IT.

Results for experimentally found optimal parameter: The optimal value (experimentally) for s in the range $[10, 20]$ is determined to be 12 for minimal CPU times. The test problems generated for the values of $l = 16, 32, 48, 64, 80, 128$, and their numerical results are reported in Table 2.7.1.

Results using parameters selection strategy in Section 2.6: To demonstrate the effectiveness of the proposed preconditioners PESS and LPESS using the parameters discussed in Section 2.6, we present the numerical results by choosing $s = 1$, $\Lambda_1 = 0.01I, \Lambda_2 = 0.1I, \Lambda_3 = 0.001I$ (denoted by PESS-I and LPESS-I) and $s = s_{est}$, $\Lambda_1 = A$,

Table 2.7.2: Numerical results of PESS-I, LPESS-I, PESS-II and LPESS-II PGMRES methods for Example 2.7.1.

method	l	16	32	48	64	80	128
	size($\mathcal{A}$)	1024	4096	9216	16384	25600	65536
PESS-I	IT	2	2	2	2	2	2
($s=1$, $\Lambda_1=0.01I$,	CPU	0.2517	0.8373	4.6626	21.8133	77.9420	308.1418
$\Lambda_2=0.1I$, $\Lambda_3=0.001I$)	RES	$4.4970e-07$	$3.8763e-07$	$3.6261e-07$	$3.4880e-07$	$3.3983e-07$	$7.7830e-07$
LPESS-I	IT	2	2	2	2	2	2
($s=1$, $\Lambda_2=0.1I$,	CPU	0.3070	0.8184	4.7244	21.6534	82.3687	288.9701
$\Lambda_3=0.001I$)	RES	$3.1116e-07$	$2.2410e-07$	$1.8435e-07$	$1.6030e-07$	$1.4375e-07$	$1.1441e-07$
PESS-II	IT	3	3	3	3	3	3
($s=s_{est}$, $\Lambda_1=A$,	CPU	0.2037	1.0533	5.8415	29.3914	104.9663	424.2004
$\Lambda_2=\beta_{est}I$, $\Lambda_3=10^{-4}CC^T$)	RES	$2.2013e-08$	$3.4850e-08$	$4.3689e-08$	$5.0705e-08$	$5.8005e-08$	$7.2906e-08$
LPESS-II	IT	3	3	3	3	3	3
($s=s_{est}$, $\Lambda_2=\beta_{est}I$,	CPU	0.1922	1.0496	4.6314	24.2092	102.8465	389.1076
$\Lambda_3=10^{-4}CC^T$)	RES	$1.8126e-09$	$1.0207e-13$	$8.0620e-15$	$9.1601e-14$	$5.7292e-13$	$1.1950e-07$

$\Lambda_2=\beta_{est}I$ and $\Lambda_3=10^{-4}C^TC$ (denoted by PESS-II and LPESS-II) for the proposed preconditioner. These results are summarized in Table 2.7.2.

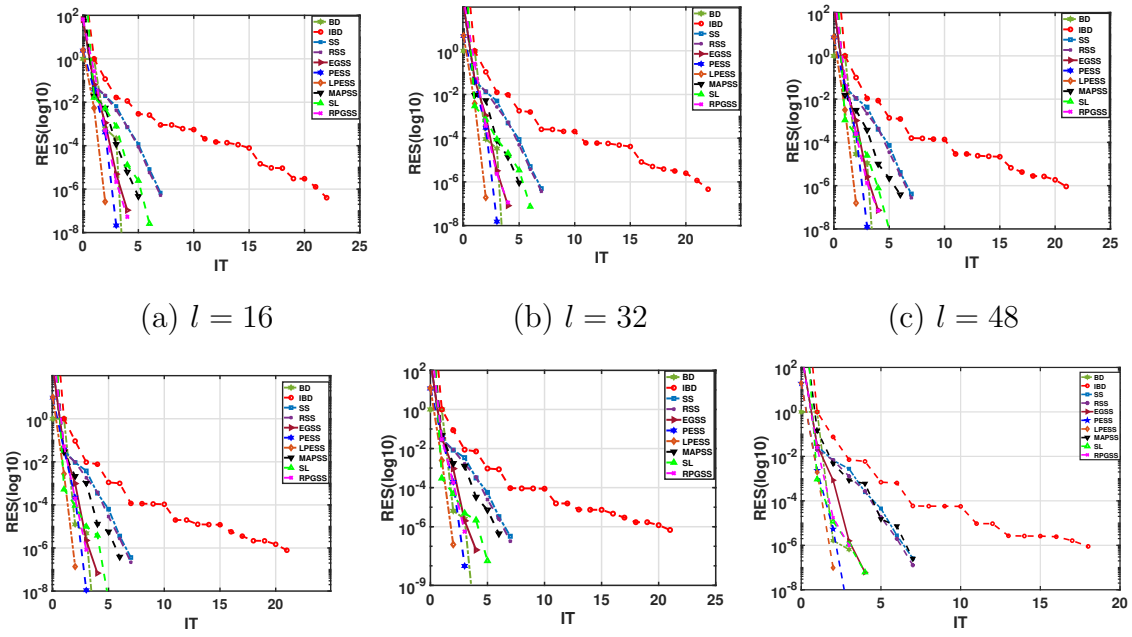
Convergence curves: Figure 2.7.1 illustrates convergence curves pertaining to preconditioners BD, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS and LPESS ($s=12$) for Case II with $l=16,32,48,64,80$ and 128. These curves depict the relationship between the relative residue (RES) at each iteration step and IT counts.

Spectral distributions: To further illustrate the superiority of the PESS preconditioner, spectral distributions of $\mathcal{A}$ and the preconditioned matrices $\mathscr{P}_{BD}^{-1}\mathcal{A}$, $\mathscr{P}_{IBD}^{-1}\mathcal{A}$, $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$, $\mathscr{P}_{SL}^{-1}\mathcal{A}$, $\mathscr{P}_{SS}^{-1}\mathcal{A}$, $\mathscr{P}_{RSS}^{-1}\mathcal{A}$, $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{RPGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for the Case II with $l=16$ and $s=13$ are displayed in Figure 2.7.2.

Spectral bounds: Furthermore, we draw the eigenvalue bounds of Theorems 2.4.1 and 2.4.5 for the PESS preconditioned matrix. The $|\lambda-1|=1$ of Theorem 2.4.1 is drawn by the green unit circle in Figure 2.7.3(a) and the points in blue color are the bounds of Theorem 2.4.5. To draw the eigenvalues bounds in Theorem 2.5.1 for LPESS preconditioned matrix, we define the following circles:

$$C_1:=\left\{\lambda\in\mathbb{C}:|\lambda|=\sqrt{\frac{\tilde{\theta}_{\max}}{1+s\vartheta_{\min}+s^2\tilde{\theta}_{\max}}}\right\},\ C_2:=\left\{\lambda\in\mathbb{C}:|\lambda|=\frac{\vartheta_{\min}}{2+s\vartheta_{\min}}\right\},$$

$$C_3:=\left\{\lambda\in\mathbb{C}:\left|\lambda-\frac{1}{s}\right|=\frac{\vartheta_{\min}}{2+s\vartheta_{\min}}\right\}\ \text{and}\ C_4:=\left\{\lambda\in\mathbb{C}:\left|\lambda-\frac{1}{s}\right|=\frac{1}{s(1+s\sqrt{\tilde{\theta}_{\max}})}\right\}.$$



(a) $l = 16$ (b) $l = 32$ (c) $l = 48$

(d) $l = 64$ (e) $l = 80$ (f) $l = 128$

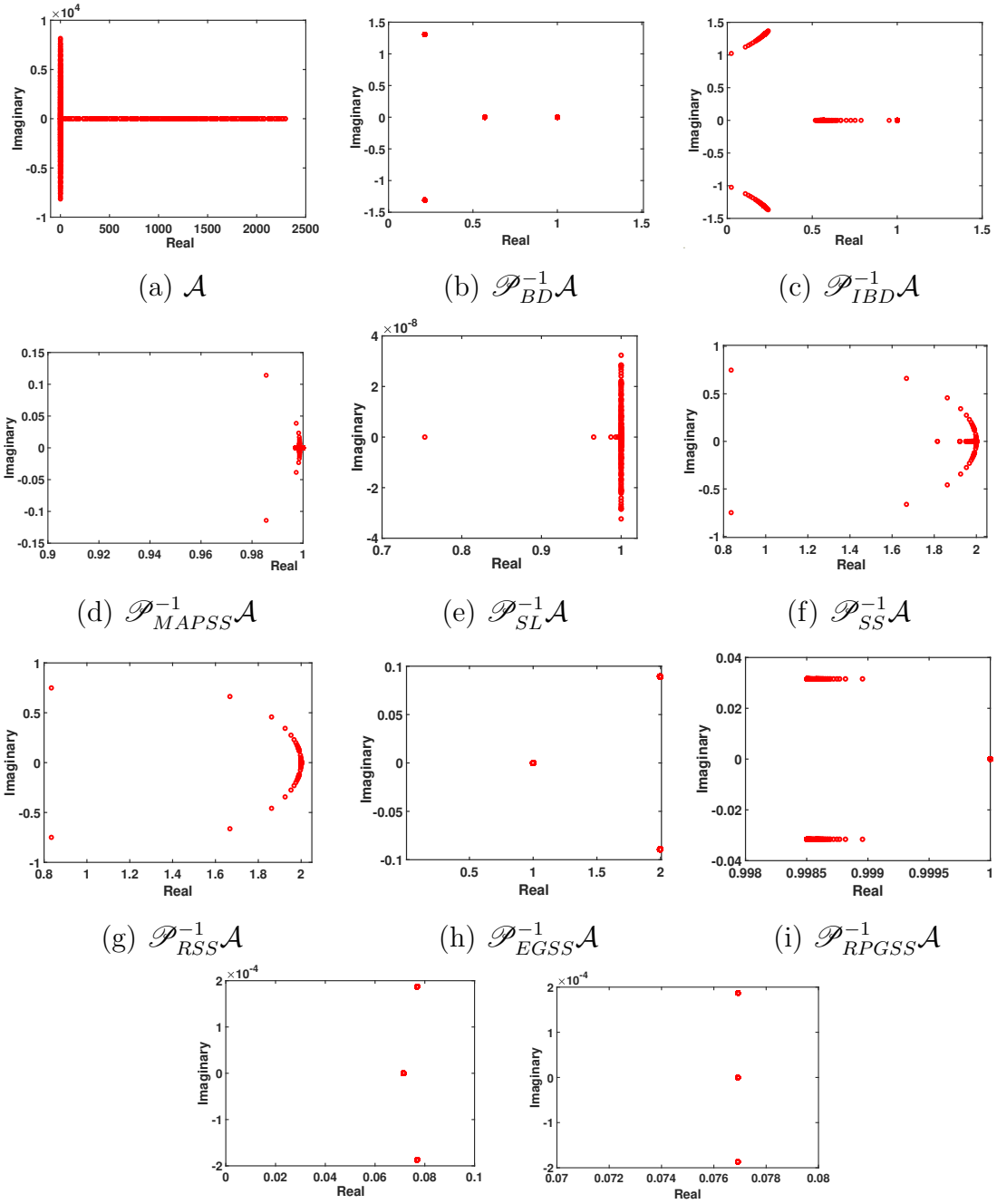
Figure 2.7.1: Convergence curves for IT versus RES of PGMRES methods employing BD, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS and LPESS ($s = 12$) preconditioners in Case II for Example 2.7.1.

In Figure 2.7.3(b), C_1 is in red color, C_2 is in blue color, C_3 is in black color and C_4 is in green color. Additionally, the eigenvalue bounds in Theorems 2.4.5, 2.4.6 and 2.5.1 are also presented in Table 2.7.3.

CN analysis: To investigate the robustness of the proposed PESS preconditioner, we measure the CN of the $\mathscr{P}_{PESS}^{-1}\mathcal{A}$, which is for any nonsingular matrix A is defined by $\kappa(A) := \|A^{-1}\|_2\|A\|_2$. In Figure 2.7.4, the influence of the parameter s on the CN of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ for Case II with $l = 32$ is depicted. The parameter s is chosen from the interval $[5, 50]$ with step size one.

Discussions: The data presented in Tables 2.7.1 and 2.7.2 and Figures 2.7.1-2.7.4 allow us to make the following noteworthy observations:

- From Table 2.7.1, it is observed that the GMRES has a very slow convergence speed. In both Cases I and II, our proposed PESS and LPESS preconditioners outperform all other compared existing preconditioners from the aspects of IT and CPU times. For example, in Case I with $l = 80$, our proposed PESS preconditioner is 76%, 43%, 36%, 22%, 28%, 27%, 33% and 39% more efficient than the existing BD, IBD,



(a) $\mathcal{A}$ (b) $\mathscr{P}_{BD}^{-1}\mathcal{A}$ (c) $\mathscr{P}_{IBD}^{-1}\mathcal{A}$

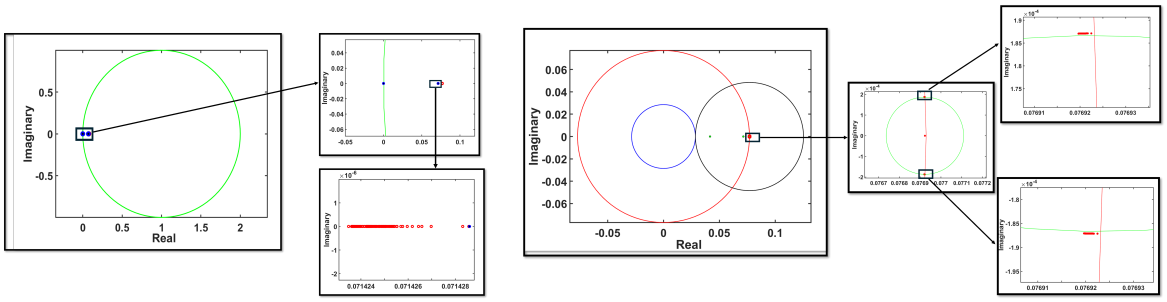
(d) $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$ (e) $\mathscr{P}_{SL}^{-1}\mathcal{A}$ (f) $\mathscr{P}_{SS}^{-1}\mathcal{A}$

(g) $\mathscr{P}_{RSS}^{-1}\mathcal{A}$ (h) $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$ (i) $\mathscr{P}_{RPGSS}^{-1}\mathcal{A}$

(j) $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ with $s = 13$ (k) $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ with $s = 13$

Figure 2.7.2: Spectral distributions of $\mathcal{A}, \mathscr{P}_{BD}^{-1}\mathcal{A}$, $\mathscr{P}_{IBD}^{-1}\mathcal{A}$, $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$, $\mathscr{P}_{SL}^{-1}\mathcal{A}$, $\mathscr{P}_{SS}^{-1}\mathcal{A}$, $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{RPGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $l = 16$ for Example 2.7.1.

MAPSS, SL, SS, RSS, EGSS and RPGSS preconditioner, respectively. Moreover, LPESS preconditioners perform approximately 78%, 48%, 38%, 29%, 35%, 34%, 39% and 44% faster than BD, IBD, MAPSS, SL, SS, RSS, EGSS and RPGSS



(a) $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ with $s = 13$ (b) $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ with $s = 13$

Figure 2.7.3: Spectral bounds for $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $l = 16$ for Example 2.7.1.

Table 2.7.3: Spectral bounds for $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for case II with $l = 16$ for Example 2.7.1.

	Real eigenvalue λ	Non-real eigenvalue λ
	Bounds of Theorem 2.4.5	Bounds of Theorem 2.4.6
$\mathscr{P}_{PESS}^{-1}\mathcal{A}$	contained in the interval $(0, 0.071429]$	$0.0667 \leq \lvert\lambda\rvert \leq 0.0739$ $4.258 \times 10^{-5} \leq \Re(\lambda/(1 - s\lambda)) \leq 0.5$ $\lvert\Im(\lambda/(1 - s\lambda))\rvert \leq 31.6386$
	Bounds of Theorem 2.5.1	Bounds of Theorem 2.5.1
$\mathscr{P}_{LPESS}^{-1}\mathcal{A}$	contained in the interval $(0.0416, 0.0769]$	$0.0285 \leq \lvert\lambda\rvert \leq 0.0769$ $1.8666 \times 10^{-4} \leq \lvert\lambda - \frac{1}{s}\rvert \leq 0.0438$

preconditioner, respectively. Similar patterns are noticed for $l = 16, 32, 48, 64$ and 128.

- Comparing the results of Tables 2.7.1 and 2.7.2, we observe that in both cases, PESS and LPESS preconditioners outperform the existing baseline preconditioners and IT are the same as in the experimentally found optimal parameters. This shows that the parameter selection strategy in Section 2.6 is effective.
- In Figure 2.7.1, it is evident that the PESS and LPESS preconditioners have a faster convergence speed than the other baseline preconditioners when applied to the PGMRES method for all $l = 16, 32, 48, 64, 80$ and 128.

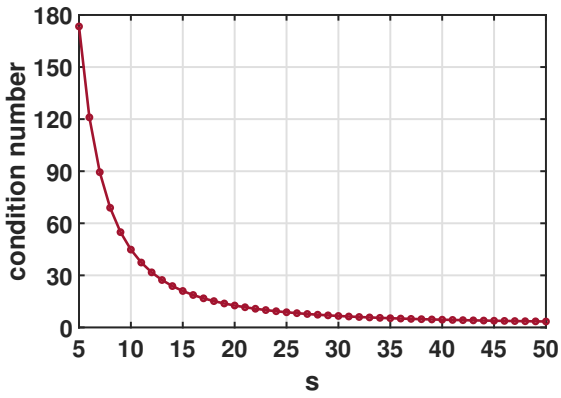


Figure 2.7.4: Parameter s vs CN of the preconditioned matrix $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ for Case II with $l = 32$ for Example 2.7.1.

- According to Figure 2.7.2, the spectrum of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LESS}^{-1}\mathcal{A}$ have better clustering properties than the baseline preconditioned matrices, consequently improves the computational efficiency of our proposed PESS and LPESS PGMRES methods. Moreover, the real eigenvalues of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ are contained in the interval $(0, 0.071429]$ and for $\mathscr{P}_{SS}^{-1}\mathcal{A}$ and $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$ are in $(0, 1.9991]$ and $(0, 1]$, respectively, which are consistent with bounds of Theorem 2.4.5, Corollaries 2.3.1 and 2.3.2, respectively. For non-real eigenvalues of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$, obtained bounds are consistent with Theorem 2.4.6. Moreover, from Figure 2.7.3(b), we observe that eigenvalues of $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ are contained in $C_2 \leq \lambda \leq C_1 \cap C_4 \leq \lambda \leq C_4$, which shows the consistency of the bounds in Theorem 2.5.1.
- For $l = 32$, the computed CNs of $\mathcal{A}$, $\mathscr{P}_{BD}^{-1}\mathcal{A}$, $\mathscr{P}_{IBD}^{-1}\mathcal{A}$, $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$ and $\mathscr{P}_{SL}^{-1}\mathcal{A}$ are $5.4289e{+}04$, $9.5567e{+}09$, $9.5124e{+}09$, $7.2548e{+}05$ and $4.2852e{+}09$, respectively, which are very large. Whereas Figure 2.7.4 illustrates a decreasing trend in the CN of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ (for instance, when $s = 50$, $\kappa(\mathscr{P}_{PESS}^{-1}\mathcal{A}) = 3.4221$) with increasing values of s. This observation highlights that $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ emerges as well-conditioned, consequently, the solution obtained by the proposed PESS PGMRES method is more reliable and robust.

The above discussions affirm that the proposed PESS and LPESS preconditioners are efficient, robust and better well-conditioned with respect to the baseline preconditioners.

Sensitivity analysis of the solution by employing the proposed PESS preconditioner: To study the sensitivity of the solution obtained by employing the proposed

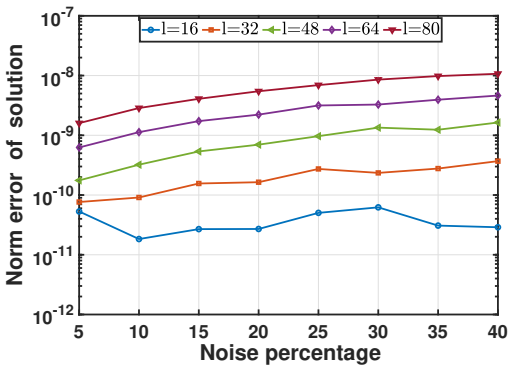


Figure 2.7.5: Relationship of norm error of solution with increasing noise percentage, employing proposed $\mathscr{P}_{PESS}$ for Case II with $s = 12$ for $l = 16, 32, 48, 64$ and 80 for Example 2.7.1.

PESS PGMRES method to small perturbations on the input data matrices. In the following, we consider the perturbed counterpart $(\mathcal{A} + \Delta\mathcal{A})\tilde{\mathbf{u}} = \mathbf{d}$ of the DSPP (2.1.1), where the perturbation matrix $\Delta\mathcal{A}$ has same structure as of $\mathcal{A}$. We construct the perturbation matrix $\Delta\mathcal{A}$ by adding noise to the block matrices B and C of $\mathcal{A}$ as follows:

$$\Delta B = 10^{-4} * N_P * std(B). * \texttt{randn}(m, n) \quad \text{and} \quad \Delta C = 10^{-4} * N_P * std(C). * \texttt{randn}(p, m),$$

where N_P is the noise percentage and $std(B)$ and $std(C)$ are the standard deviation of B and C, respectively. We perform the numerical test for $l = 16, 32, 48, 64$ and 80 for Example 2.7.1 using the PESS PGMRES method (Case II, $s = 12$). The calculated norm error $\|\tilde{\mathbf{u}} - \mathbf{u}\|_2$ among the solution of the perturbed system and the original system with increasing N_P from 5% to 40% with step size 5% are displayed in the Figure 2.7.5. We noticed that with growing N_P, the norm error of the solution remains consistently less than 10^{-8}, which demonstrates the robustness of the proposed PESS preconditioner and the solution $\mathbf{u}$ is insensitive to small perturbation on $\mathcal{A}$.

Example 2.7.2. Problem formulation: In this example, we consider the DSPP (2.1.1), where block matrices A and B originate from the two dimensional Stokes equation namely "leaky" lid-driven cavity problem [59], in a square domain $\Xi = \{(x, y) \mid 0 \leq x \leq 1, 0 \leq y \leq 1\}$, which is defined as follows:

$$\begin{aligned} -\Delta\mathbf{u} + \nabla\boldsymbol{p} &= \mathbf{0}, \quad \text{in} \quad \Xi, \\ \nabla \cdot \mathbf{u} &= 0, \quad \text{in} \quad \Xi. \end{aligned} \tag{2.7.1}$$

Table 2.7.4: Numerical results of GMRES, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS, and LPESS PGMRES methods for Example 2.7.2.

method	**h**	1/8	1/16	1/32	1/64	1/128
	size($\mathcal{A}$)	288	1088	4224	16640	66048
GMRES	IT	158	548	2048	$--$	$--$
	CPU	0.2325	0.6960	124.7595	4545.5072	9396.6958
	RES	$8.0113e-07$	$8.8023e-07$	$9.0887e-07$	$4.0481e-06$	$8.3021e-06$
IBD	IT	34	35	25	23	21
	CPU	**0.0147**	**0.1166**	1.9070	32.6213	1664.6991
	RES	$9.4111e-07$	$6.9932e-07$	$9.2544e-07$	$9.3340e-07$	$7.3550e-07$
MAPSS	IT	11	7	4	4	4
	CPU	0.2061	0.3453	1.8717	26.6600	482.0390
	RES	$5.2642e-08$	$1.1366e-07$	$7.1116e-08$	$8.2780e-07$	$2.1189e-09$
SL	IT	6	6	5	5	4
	CPU	0.2130	0.2929	1.8697	27.1124	516.3012
	RES	$6.6518e-07$	$1.1953e-08$	$9.5638e-08$	$2.1528e-11$	$2.9479e-10$
Case I						
SS	IT	5	6	9	11	11
	CPU	0.3047	0.4422	2.4279	51.5825	1245.2019
	RES	$4.5523e-07$	$7.9331e-07$	$2.5307e-07$	$4.5327e-07$	$9.8497e-07$
RSS	IT	**4**	**4**	5	5	6
	CPU	0.2044	0.3436	1.6047	27.6128	720.4234
	RES	$3.4853e-08$	$9.2155e-07$	$1.7469e-08$	$2.7261e-07$	$3.2621e-09$
EGSS	IT	7	9	11	15	5
	CPU	0.2389	0.4051	4.4353	52.7804	744.5983
	RES	$1.6861e-07$	$4.0517e-08$	$1.3223e-07$	$1.774e-07$	$9.1207e-07$
RPGSS	IT	4	4	4	5	5
	CPU	0.2105	0.3883	2.2981	28.1748	544.8828
	RES	$2.8700e-09$	$5.1131e-08$	$3.3646e-07$	$8.8663e-09$	$1.4981e-07$
PESS†	IT	4	4	5	5	4
$s = 30$	CPU	0.2068	0.3480	1.7633	25.2920	484.2796
	RES	$2.3499e-08$	$4.2761e-07$	$2.1247e-08$	$2.0987e-07$	$4.3859e-07$
LPESS†	IT	**3**	**3**	**4**	**3**	4
$s = 30$	CPU	**0.1990**	**0.2578**	**1.4778**	**17.9730**	457.7446
	RES	$1.6606e-07$	$8.3803e-07$	$4.9289e-10$	$5.2658e-07$	$3.3191e-08$
Case II						
SS	IT	29	38	60	52	55
	CPU	0.2999	1.1144	13.4580	213.2381	4980.7387
	RES	$7.3560e-07$	$9.5790e-07$	$9.2515e-07$	$9.6425e-07$	$9.6948e-07$

Table 2.7.4: Numerical results of GMRES, IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS, and LPESS PGMRES methods for Example 2.7.2 (continued).

method	**h**	1/8	1/16	1/32	1/64	1/128
	IT	12	13	14	12	12
RSS	CPU	0.3797	0.6019	3.4491	50.3476	1361.8902
	RES	$3.2456e-07$	$5.3086e-07$	$6.6600e-07$	$4.2834e-07$	$1.2733e-07$
	IT	9	9	9	7	4
EGSS	CPU	0.2459	0.4180	2.4320	27.9359	419.1294
	RES	$7.9001e-09$	$8.6623e-09$	$7.4644e-09$	$1.2457e-07$	$7.6469e-08$
	IT	6	7	7	5	4
RPGSS	CPU	0.2154	0.4096	1.8893	23.2802	423.0121
	RES	$5.2642e-08$	$1.1366e-07$	$7.1116e-08$	$8.2780e-07$	$2.1189e-09$
	IT	5	5	5	5	**3**
PESS†	CPU	0.2091	0.4184	1.8118	25.8048	**365.3128**
$s=26$	RES	$2.3801e-08$	$1.1965e-08$	$8.4923e-09$	$5.9356e-08$	$4.2039e-07$
	IT	4	5	**5**	**5**	**3**
LPESS†	CPU	0.2070	0.3662	**1.5330**	**23.5997**	**359.1330**
$s=26$	RES	$2.4250e-07$	$8.6182e-12$	$5.3930e-11$	$8.1253e-10$	$5.8345e-07$

Here † represents the proposed preconditioners. The boldface represents the top two results.
$--$ indicates that the iteration method does not converge within the prescribed IT.

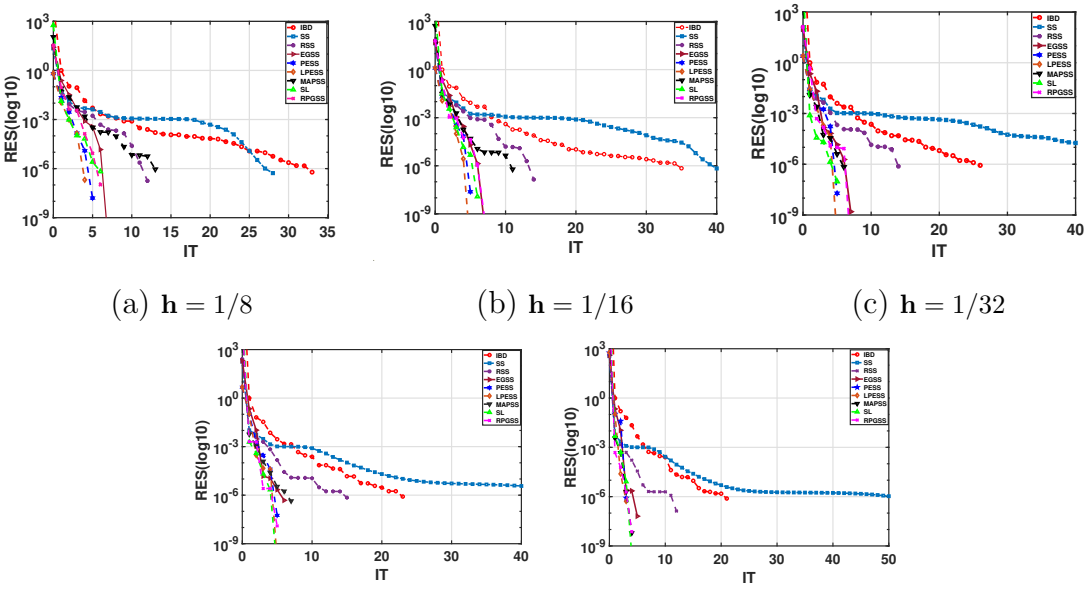
A Dirichlet no-flow condition is applied on the side and bottom boundaries, and the nonzero horizontal velocity on the lid is $\{y=1; -1 \le x \le 1 | \mathbf{u}_x = 1\}$. Here, $\mathbf{u}$ and $\boldsymbol{p}$ refer to the velocity vector field and the pressure scalar field, respectively.

To generate the matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times n}$ of the system (2.1.1), the discretization task of the Stokes equation (2.7.1) is accomplished by the IFISS software developed by Elman et al. [59]. Following [59] with grid parameters $\mathbf{h} = 1/8, 1/16, 1/32, 1/64, 1/128$, we get $n = 2(1+1/\mathbf{h})^{\mathbf{2}}$ and $m = 1/\mathbf{h}^2$. To make the system of equations in (2.1.1) not too ill-conditioned and not too sparse, we construct the block matrix $C = \begin{bmatrix}\Pi & \texttt{randn}(p, m-p)\end{bmatrix}$, where $\Pi = \text{diag}(1, 3, 5, \ldots, 2p-1)$ and $p = m-2$. Additionally, to ensure the positive definiteness of the matrix A, we add $0.001I$ with A.

In Table 2.7.4, we list numerical results produced using the GMRES method and PGMRES method with the preconditioners $\mathscr{P}_{IBD}$, $\mathscr{P}_{MAPSS}$, $\mathscr{P}_{SL}$, $\mathscr{P}_{SS}$, $\mathscr{P}_{RSS}$, $\mathscr{P}_{EGSS}$, $\mathscr{P}_{RPGSS}$, $\mathscr{P}_{PESS}$ and $\mathscr{P}_{LPESS}$ for different grid parameter values of $\mathbf{h}$.

Table 2.7.5: Numerical results of PESS-I, LPESS-I, PESS-II and LPESS-II PGMRES methods for Example 2.7.2.

method	**h**	1/8	1/16	1/32	1/64	1/128
	size($\mathcal{A}$)	288	1088	4224	16640	66048
PESS-I	IT	5	5	6	6	5
($s=1$, $\Lambda_1=0.001I$,	CPU	0.30867	0.46688	2.48254	46.52887	577.9420
$\Lambda_2=0.1I$, $\Lambda_3=0.001I$)	RES	$1.4125e-08$	$1.3379e-07$	$3.8821e-08$	$2.6061e-07$	$9.7383e-07$
LPESS-I	IT	4	4	4	5	3
($s=1$, $\Lambda_2=0.1I$,	CPU	0.1829	0.3319	1.2774	19.1090	349.6728
$\Lambda_3=0.001I$)	RES	$3.5507e-09$	$3.4257e-08$	$7.0375e-07$	$2.3073e-09$	$7.1453e-07$
PESS-II	IT	3	4	4	4	4
($s=s_{est}$, $\Lambda_1=A$,	CPU	0.1847	0.3954	1.3215	17.22840	485.0177
$\Lambda_2=\beta_{est}I$, $\Lambda_3=10^{-4}CC^T$)	RES	$9.6100e-07$	$5.6520e-07$	$7.8750e-08$	$3.9903e-08$	$6.9694e-09$
LPESS-II	IT	6	6	5	5	3
($s=s_{est}$, $\Lambda_2=\beta_{est}I$,	CPU	0.1656	0.3777	1.4341	19.0658	377.2001
$\Lambda_3=10^{-4}CC^T$)	RES	$2.5227e-07$	$5.1622e-07$	$4.8924e-08$	$3.2751e-08$	$7.9950e-07$

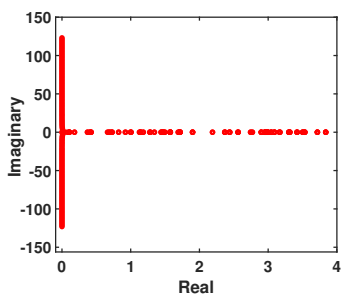


(a) $\mathbf{h}=1/8$ (b) $\mathbf{h}=1/16$ (c) $\mathbf{h}=1/32$

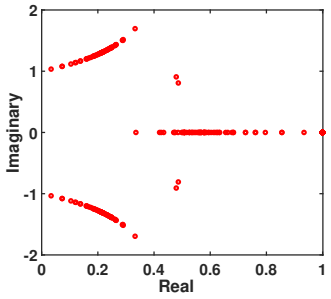
(d) $\mathbf{h}=1/64$ (e) $\mathbf{h}=1/128$

Figure 2.7.6: Convergence curves for IT versus RES of the PGMRES methods by employing IBD, MAPSS, SL, SS, RSS, EGSS, RPGSS, PESS and LPESS preconditioners in Case II for Example 2.7.2.

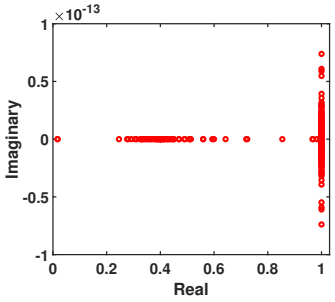
Parameter selection: Parameter choices for the proposed preconditioners are made in two cases as in Example 2.7.1. However, in Case I, we take $\alpha=0.01, \beta=0.1$, $\Lambda_1=0.01I$



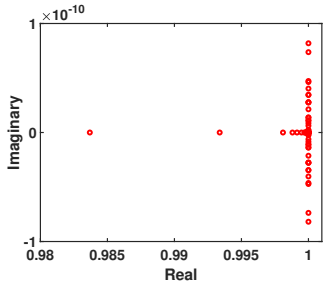
(a) $\mathcal{A}$



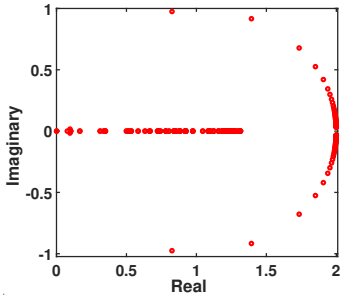
(b) $\mathscr{P}_{IBD}^{-1}\mathcal{A}$.



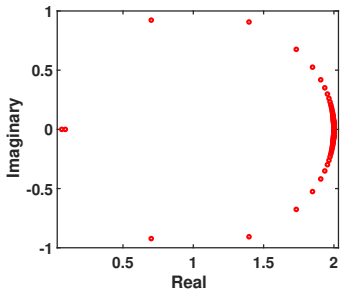
(c) $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$



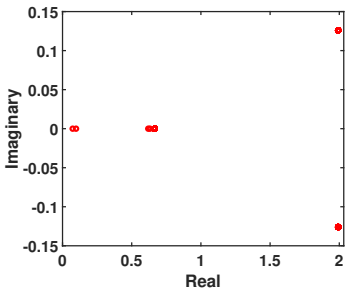
(d) $\mathscr{P}_{SL}^{-1}\mathcal{A}$



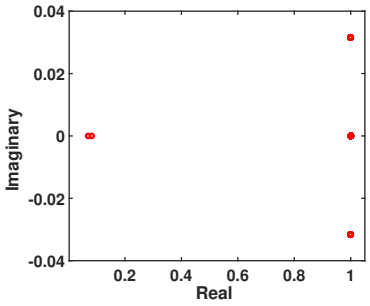
(e) $\mathscr{P}_{SS}^{-1}\mathcal{A}$



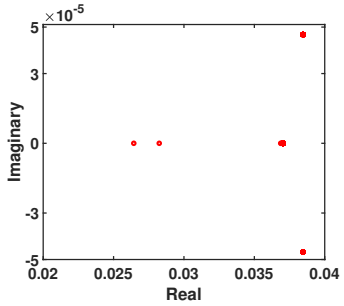
(f) $\mathscr{P}_{RSS}^{-1}\mathcal{A}$



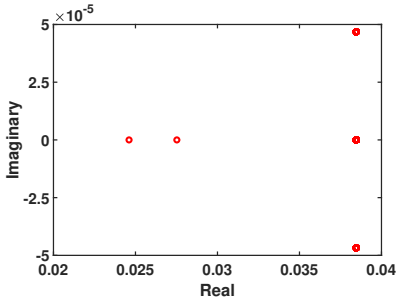
(g) $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$



(h) $\mathscr{P}_{RPGSS}^{-1}\mathcal{A}$

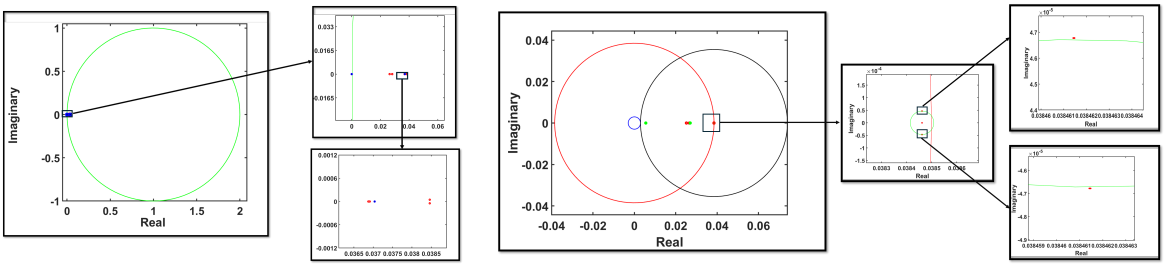


(i) $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ with $s = 26$



(j) $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ with $s = 26$

Figure 2.7.7: Spectral distributions of $\mathcal{A}$, $\mathscr{P}_{IBD}^{-1}\mathcal{A}$, $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$, $\mathscr{P}_{SL}^{-1}\mathcal{A}$, $\mathscr{P}_{SS}^{-1}\mathcal{A}$, $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{RPGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $\mathbf{h} = 1/8$.



(a) $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ with $s = 26$ (b) $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ with $s = 26$

Figure 2.7.8: Spectral bounds for $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $\mathbf{h} = 1/8$ for Example 2.7.2.

and $\Lambda_2 = 0.1I$. The parameter α in Case I is taken as in [37] and the parameters in Case II are taken as in [156]. For the IBD preconditioner, the matrices $\widehat{A}$ and $\widehat{S}$ and for MAPSS preconditioner α and β are constructed as in Example 2.7.1.

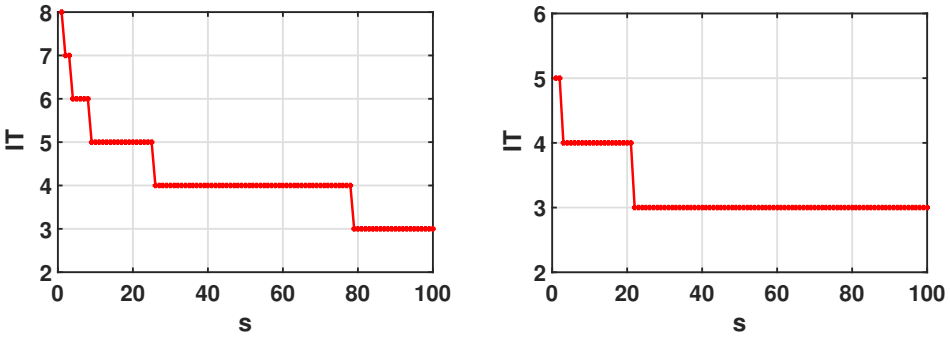
Results for experimentally found optimal parameter: In the interval $[10, 30]$, the empirical optimal choice for s is found to be 30 for Case I and 26 for Case II. Table 2.7.4 shows that the GMRES method has a very slow convergence speed and also does not converge for $\mathbf{h} = 1/64, 1/128$ within 7000 iterations. The proposed preconditioners require almost five times fewer iterations compared to the IBD preconditioner for convergence. Moreover, in both cases, the PESS and LPESS preconditioners outperform the MAPSS, SL, SS, RSS, EGSS and RPGSS preconditioners in terms of IT and CPU times. Notably, for the PESS preconditioner, the IT remains constant in both cases, whereas for the SS and EGSS preconditioners, the IT increases as the size of $\mathcal{A}$ increases. Furthermore, in Case I with $\mathbf{h} = 1/64$, our proposed LPESS preconditioner is approximately 45%, 33%, 34%, 65%, 35%, 66% and 36% more efficient than the existing IBD, MAPSS, SL, SS, RSS, EGSS and RPGSS preconditioners, respectively. Similar patterns are observed for other values of $\mathbf{h}$.

Results using parameters selection strategy in Section 2.6: To demonstrate the effectiveness of the proposed PESS and LPESS preconditioners using the parameters discussed in Section 2.6, we present the numerical test results for PESS-I, LPESS-I, PESS-II and LPESS-II in Table 2.7.5 as in Example 2.7.1. Comparing the results in Tables 2.7.4 and 2.7.5, we observe that the parameter selection strategy described in Section 2.6 is effective.

Convergence curves: The convergence curves in Figure 2.7.6 demonstrate the rapid convergence of the proposed PESS and LPESS PGMRES methods compared to the IBD, MAPSS, SL, SS, RSS, EGSS and RPGSS in terms of RES versus IT counts.

Spectral distributions: Figure 2.7.7 illustrates the spectral distributions of the original matrix $\mathcal{A}$, and the preconditioned matrices $\mathscr{P}_{IBD}^{-1}\mathcal{A}$, $\mathscr{P}_{MAPSS}^{-1}\mathcal{A}$, $\mathscr{P}_{SL}^{-1}\mathcal{A}$, $\mathscr{P}_{SS}^{-1}\mathcal{A}$, $\mathscr{P}_{RSS}^{-1}\mathcal{A}$, $\mathscr{P}_{EGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{PRGSS}^{-1}\mathcal{A}$, $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ for Case II with $\mathbf{h} = 1/8$. According to Figure 2.7.7, eigenvalues of $\mathscr{P}_{PESS}^{-1}\mathcal{A}$ and $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ demonstrate a superior clustering compared to the other preconditioned matrices, which leads to a favorable convergence speed for the proposed PESS and LPESS PGMRES methods.

Spectral bounds: In Figure 2.7.8, we draw the spectral bounds of Theorems 2.4.1, 2.4.5 and 2.5.1 for PESS and LPESS preconditioned matrices. In Figure 2.7.8(a), $|\lambda - 1| = 1$ of Theorem 2.4.1 is drawn by the green unit circle, while the points in blue indicate the bounds from Theorem 2.4.5. Moreover, we draw the circles C_1 (in red), C_2 (in blue), C_3 (in black) and C_4 (in green) in Figure 2.7.8(b). We can observe that the eigenvalues of the preconditioned matrix $\mathscr{P}_{LPESS}^{-1}\mathcal{A}$ lie in the intersection of the annulus of the circles C_1 and C_2, and the annulus of the circles C_3 and C_4. .



(a) PESS PGMRES method (b) LPESS PGMRES method

Figure 2.7.9: Characteristic curves for IT of the proposed PESS (left) and LPESS (right) PGMRES methods by varying s in the interval $[1, 100]$ with step size 1 with $\mathbf{h} = 1/16$ in Case I for Example 2.7.2.

CN analysis: Furthermore, the system (2.1.1) exhibits ill-conditioning nature with $\kappa(\mathcal{A}) = 5.0701e + 05$. While for Case I with $\mathbf{h} = 1/16$, $\kappa(\mathscr{P}_{PESS}^{-1}\mathcal{A}) = 1.3353$ and $\kappa(\mathscr{P}_{LPESS}^{-1}\mathcal{A}) = 1.2056$, indicating that the proposed PESS and LPESS preconditioned systems are well-conditioned, ensuring an efficient and robust solution.

Relationship between s and convergence speed: In addition, to demonstrate the relationship of the parameter s and the convergence speed of the PESS and LPESS preconditioners, we plot graphs of IT counts by varying the parameters s in the interval $[1, 100]$ with step size one in Figure 2.7.9. We consider $\mathbf{h} = 1/16$ and other choices for Λ_1, Λ_2 and Λ_3 as in Case I. Figure 2.7.9 shows that, with the increasing value of s, decreasing trend in the IT counts for both the PESS and LPESS preconditioner. Moreover, for $s > 22$, using LPESS preconditioner, we can solve this DSPP only in 3 iterations.

2.8. Summary

In this chapter, we proposed the PESS iterative method and corresponding PESS preconditioner and its relaxed variant LPESS preconditioner to solve the DSPP (2.1.1). For the convergence of the proposed PESS iterative method, necessary and sufficient criteria are derived. Moreover, we estimated the spectral bounds of the proposed PESS and LPESS preconditioned matrices. This empowers us to derive spectral bounds for SS and EGSS preconditioned matrices. Numerous experimental analyses are performed to demonstrate the effectiveness of our proposed PESS and LPESS preconditioners. The key observations are as follows: (i) the proposed PESS and LPESS preconditioners are found to outperform the existing baseline preconditioners in terms of IT and CPU times. (ii) The proposed preconditioners significantly reduces the CN of $\mathcal{A}$, consequently showing their proficiency in solving DSPPs. (iii) The proposed PESS and LPESS preconditioned matrices have better clustered spectral distribution than the baseline preconditioned matrices. (iv) Sensitivity analysis conducted by introducing different percentages of noise on the system (2.1.1) showcases the robustness of the proposed PESS preconditioner.

CHAPTER 3

A Class of Generalized Shift-Splitting Preconditioners for Double Saddle Point Problems*

In this chapter, we propose a generalized shift-splitting (GSS) preconditioner, along with its two relaxed variants, to solve the DSPP by considering $F = B$, $G = C$, and $D = \mathbf{0}$. The convergence of the associated GSS iterative method is analyzed, and sufficient conditions for its convergence are established. Spectral analyses are performed to derive sharp bounds for the eigenvalues of the preconditioned matrices. Numerical experiments based on examples arising from the PDE-constrained optimization problems demonstrate the effectiveness and robustness of the proposed preconditioners compared with existing state-of-the-art preconditioners.

3.1. Background

Suppose n, m and p are given positive integers with $n \geq m \geq p$. Then, we consider the DSPP in the following form [33]:

$$\mathfrak{B}\widehat{\boldsymbol{w}} := \begin{bmatrix} A & \mathbf{0} & B^T \\ \mathbf{0} & E & C \\ -B & -C^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{z} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} p \\ q \\ r \end{bmatrix} =: \widetilde{\mathbf{d}}, \tag{3.1.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times m}$ and $E \in \mathbb{R}^{p\times p}$. Further, $p \in \mathbb{R}^n$, $q \in \mathbb{R}^p$ and $r \in \mathbb{R}^m$ are known vectors and $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{z} \in \mathbb{R}^p$ are unknown vectors to be determined. In this section, we consider that the matrices A and E can be both symmetric or nonsymmetric.

The DSPP (3.1.1) is frequently encountered in a wide range of scientific and computational disciplines. Notable areas of application include quadratic programming problems [71], EILS problems [30], PDE-constrained optimization problem [115], and so on.

Owing to the broad applicability of the DSPP (3.1.1), this chapter primarily concentrates on its numerical solution. Nevertheless, for the large and sparse nature of the double

* S. S. Ahmad and **P. Khatun**, "A class of generalized shift-splitting preconditioners for double saddle point problems." *Revision submitted in Applied Mathematics and Computation.*

saddle point matrix $\mathfrak{B}$, iterative methods are generally preferred over direct approaches [122]. In this chapter, we develop robust and efficient preconditioners to enhance the convergence of Krylov subspace methods, such as GMRES, for solving the DSPP (3.1.1).

To leverage the full block structure of $\mathfrak{B}$, various preconditioners have been studied in the literature to solve the DSPP (3.1.1). When $E = \mathbf{0}$ in (3.1.1), BD and block tridiagonal-type preconditioners [1, 75], SS-type preconditioners [37], Uzawa methods [74, 76], etc. have been explored. When $E \neq \mathbf{0}$, BD preconditioners for the DSPP (3.1.1) have been investigated in [33]. By splitting the coefficient matrix $\mathfrak{B}$ as $\mathfrak{B} = \mathfrak{B}_1 + \mathfrak{B}_2$, where

$$\mathfrak{B}_1 = \begin{bmatrix} A & \mathbf{0} & B^T \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -B & \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathfrak{B}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & E & C \\ \mathbf{0} & -C^T & \mathbf{0} \end{bmatrix}. \tag{3.1.2}$$

Benzi and Guo [23] proposed the dimensional spitting (DS) preconditioner $\mathscr{P}_{\text{DS}}$ and a relaxed dimensional factorization (RDF) preconditioner $\mathscr{P}_{\text{RDF}}$ [27]. These are given as follows:

$$\mathscr{P}_{\text{DS}} = \frac{1}{\alpha} \begin{bmatrix} \alpha I + A & \mathbf{0} & B^T \\ \mathbf{0} & \alpha I & \mathbf{0} \\ -B & \mathbf{0} & \alpha I \end{bmatrix} \begin{bmatrix} \alpha I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \alpha I + E & C \\ \mathbf{0} & -C^T & \alpha I \end{bmatrix}, \tag{3.1.3}$$

$$\mathscr{P}_{\text{RDF}} = \frac{1}{\alpha} \begin{bmatrix} A & \mathbf{0} & B^T \\ \mathbf{0} & \alpha I & \mathbf{0} \\ -B & \mathbf{0} & \alpha I \end{bmatrix} \begin{bmatrix} \alpha I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & E & C \\ \mathbf{0} & -C^T & \alpha I \end{bmatrix}. \tag{3.1.4}$$

For more on DS-based preconditioners, refer to [70, 154].

For the DSPP (3.1.1) arising from PDE-constrained optimization problem, Rees et al. [115] introduced BD preconditioner $\mathscr{P}_{\text{E}}$ and constrained preconditioner $\mathscr{P}_{\text{C}}$. Later, a block triangular (BT) preconditioner is developed in [114]. In [161], the authors introduced two types of preconditioners for solving the DSPP: a block-counter-diagonal preconditioner, denoted as $\mathscr{P}_{\text{BCD}}$, and a block-counter-tridiagonal preconditioner, denoted as $\mathscr{P}_{\text{BCT}}$. By using different approximations of Schur complement, preconditioners in BD and BT formats are constructed in [107, 108, 109]. For more research on solving DSPP in the context of PDE-constrained optimization problems, refer to [61, 83, 104].

In recent years, several SS-type preconditioners have been developed for GSPP and DSPP (2.1.1) with nonsymmetric coefficient matrix, demonstrating promising efficiency; see, for example, [5, 37, 40, 41, 124]. However, despite their potential for high efficiency, SS-type preconditioners have not yet been explored specifically for DSPP (3.1.1).

Motivated by this gap, in this chapter, we introduce a generalized shift-splitting (GSS) iterative method along with its associated GSS preconditioner for solving DSPP of the form (3.1.1). Our approach extends the concept of SS to the coefficient matrix $\mathfrak{B}$, aiming to enhance computational efficiency and convergence. Moreover, we derive sufficient criteria for the convergence of the GSS iterative method. The main contributions of the chapter are summarized as follows:

- A novel GSS iterative method and corresponding GSS preconditioner are introduced by implementing the SS approach for the coefficient matrix $\mathfrak{B}$ to solve DSPP (3.1.1).
- Convergence analysis of the proposed GSS iterative method is carried out, yielding sufficient conditions for its convergence.
- To further enhance the effectiveness of the GSS preconditioner, two relaxed variants, termed RGSS-I and RGSS-II are proposed, and the spectral bounds of the RGSS-I and RGSS-II preconditioned matrices are thoroughly investigated.
- Finally, numerical experiments are conducted for the DSPP arising from the PDE-constrained optimization problem to demonstrate the effectiveness of the proposed preconditioners.

The outline of the rest of the chapter is as follows. Section 3.2 investigates the solvability conditions of the DSPP (3.1.1) and properties of the coefficient matrix $\mathfrak{B}$. The GSS iterative method and associated preconditioner are proposed in Section 3.3. Section 3.4 investigates the convergence criteria for the proposed GSS iterative method. Two relaxed variants of the proposed GSS preconditioner are presented in Section 3.5, and the spectral analyses of the corresponding preconditioned matrices are performed. Section 3.6 deals with the parameter selection strategy of the proposed preconditioners. Experimental results, analyses and discussions of the proposed and existing preconditioners are discussed in Section 3.7. At the end, Section 3.8 includes some concluding statements.

3.2. Solvability Conditions and Properties of the DSPP

In this section, we provide the solvability conditions on the block matrices A, B, C and E for the system (3.1.1) and a few important properties of the matrix $\mathfrak{B}$. The nonsymmetric coefficient matrix $\mathfrak{B}$ possesses the following desirable properties, which are crucial in the theoretical analysis of the iterative method and preconditioners designed to solve DSPP (3.1.1). Before that, we define $A_H := \frac{A+A^T}{2}$ and $E_H := \frac{E+E^T}{2}$.

Proposition 3.2.1. *Let $A \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{p\times p}$ with A_H and E_H be positive semidefinite. If B has full row rank, then*

(i) $\mathfrak{B}$ is semipositive real: $\boldsymbol{u}^T\mathfrak{B}\boldsymbol{u} \geq 0$ for all $\boldsymbol{u} \in \mathbb{R}^{n+p+m}$.

(ii) $\mathfrak{B}$ is positive semistable: the real part of all eigenvalues of $\mathfrak{B}$ is nonnegative.

Proof. (i) Let $\boldsymbol{u} = [x^T,\, y^T\, z^T]^T \in \mathbb{R}^{n+p+m}$. Then $\boldsymbol{u}^T\mathfrak{B}\boldsymbol{u} = x^TAx + y^TEy$. Therefore, we have $\boldsymbol{u}^T\mathfrak{B}\boldsymbol{u} + \boldsymbol{u}^T\mathfrak{B}^T\boldsymbol{u} = 2(x^TA_Hx + y^TE_Hy)$ and hence, $\boldsymbol{u}^T\mathfrak{B}\boldsymbol{u} = x^TA_Hx + y^TE_Hy \geq 0$, as A_H and E_H are positive semidefinite. Thus, $\mathfrak{B}$ is semipositive real.
(ii) Let λ be an eigenvalue of $\mathcal{B}$ and $\boldsymbol{u} = [u^T, v^T, w^T]^T \in \mathbb{R}^{n+p+m}$ is the corresponding eigenvector. Then $\boldsymbol{u}^*\mathfrak{B}\boldsymbol{u} = \lambda\|\boldsymbol{u}\|_2$ and $(\boldsymbol{u}^*\mathfrak{B}\boldsymbol{u})^* = \bar{\lambda}\|\boldsymbol{u}\|_2$. Thus

$$\begin{aligned}\Re(\lambda) &= \frac{\boldsymbol{u}^*(\mathfrak{B}+\mathfrak{B}^T)\boldsymbol{u}}{2\|\boldsymbol{u}\|_2}\\ &= \frac{\Re(\boldsymbol{u})^T(\mathfrak{B}+\mathfrak{B}^T)\Re(\boldsymbol{u})^T + \Im(\boldsymbol{u})^T(\mathfrak{B}+\mathfrak{B}^T)\Im(\boldsymbol{u})^T}{2\|\boldsymbol{u}\|_2}.\end{aligned}$$

Then, using (i), we have $\Re(\lambda) \geq 0$. ■

Proposition 3.2.2. *Let $A \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{p\times p}$ be nonsingular matrices with A_H and E_H positive definite. If B has full row rank, then the double saddle point matrix $\mathfrak{B}$ is nonsingular.*

Proof. Let B has full row rank, and $\widehat{\boldsymbol{w}} = [\boldsymbol{x}^T, \boldsymbol{z}^T, \boldsymbol{y}^T]^T \in \mathbb{R}^{n+p+m}$ be such that $\mathfrak{B}\widehat{\boldsymbol{w}} = \boldsymbol{0}$. Then, we have

$$\begin{cases} A\boldsymbol{x} + B^T\boldsymbol{y} &= \boldsymbol{0},\\ E\boldsymbol{z} + C\boldsymbol{y} &= \boldsymbol{0},\\ -B\boldsymbol{x} - C^T\boldsymbol{z} &= \boldsymbol{0}. \end{cases} \tag{3.2.1}$$

We first assert that $\boldsymbol{x} = \boldsymbol{0}$. Then from the first equation in (3.2.1), we obtain $B^T\boldsymbol{y} = \boldsymbol{0}$. Since, B has full row rank, this implies $\boldsymbol{y} = \boldsymbol{0}$. Thus the second equation of (3.2.1) gives $\boldsymbol{z} = \boldsymbol{0}$ as E is nonsingular. This implies $\widehat{\boldsymbol{w}} = \boldsymbol{0}$. Next, we assert $\boldsymbol{y} = \boldsymbol{0}$. Then, from the first and second equation of (3.2.1), we find that $\boldsymbol{x} = \boldsymbol{0}$ and $\boldsymbol{z} = \boldsymbol{0}$, respectively, as A and E are nonsingular. Hence, $\widehat{\boldsymbol{w}} = \boldsymbol{0}$. Now, we assume that $\boldsymbol{x} \neq \boldsymbol{0}$ and $\boldsymbol{y} \neq \boldsymbol{0}$. Then multiplying by $\boldsymbol{x}^T$ from the left side of the first equation of (3.2.1), we obtain

$$\boldsymbol{x}^TA\boldsymbol{x} + \boldsymbol{x}^TB^T\boldsymbol{y} = 0. \tag{3.2.2}$$

Again, multiplying third equation of (3.2.1) by $\boldsymbol{y}^T$ from the left, we get

$$-\boldsymbol{y}^TB\boldsymbol{x} - \boldsymbol{y}^TC^T\boldsymbol{z} = 0. \tag{3.2.3}$$

Substituting (3.2.2) and $E\boldsymbol{z} = -C\boldsymbol{y}$ on (3.2.3), we have $\boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{z}^T E\boldsymbol{z} = 0$, and therefore it must be $\boldsymbol{x}^T A\boldsymbol{x} = 0$ and $\boldsymbol{z}^T E\boldsymbol{z} = 0$, as both the quantities are nonnegative. However, $\boldsymbol{x}^T A\boldsymbol{x} = \boldsymbol{x}^T A_H\boldsymbol{x} = 0$ and $\boldsymbol{z}^T E\boldsymbol{z} = \boldsymbol{z}^T E_H\boldsymbol{z} = 0$, which implies $\boldsymbol{x} = \boldsymbol{0}$ and $\boldsymbol{z} = \boldsymbol{0}$, since A_H and E_H are positive definite matrices. Thus, $\widehat{\boldsymbol{w}} = \boldsymbol{0}$, and hence, $\mathfrak{B}$ is nonsingular. ■

Proposition 3.2.3. *Let $A \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{p\times p}$ with A_H and E_H positive definite matrices. If B and C are of full row rank, then the matrix $\mathfrak{B}$ is positive stable, i.e., $\lambda > 0$ for all $\lambda \in \sigma(\mathfrak{B})$.*

Proof. Suppose λ is an eigenvalue of $\mathfrak{B}$ and $\boldsymbol{w} = [u^T, v^T, w^T]^T \in \mathbb{R}^{n+p+m}$ is the corresponding eigenvector. Then, we have $\mathfrak{B}\boldsymbol{w} = \lambda\boldsymbol{w}$, which leads to following three linear system of equations:

$$\begin{cases} Au + B^T w &= \lambda u, \\ Ev + Cw &= \lambda v, \\ -Bu - C^T v &= \lambda w. \end{cases} \tag{3.2.4}$$

Premultiplying $\mathfrak{B}\boldsymbol{w} = \lambda\boldsymbol{w}$ by $\boldsymbol{w}^H$, we get

$$\lambda\|\boldsymbol{w}\|_2^2 = u^H Au + v^H Ev + 2\boldsymbol{i}\Im(u^H B^T w + v^H Cw). \tag{3.2.5}$$

On the other hand, from $\boldsymbol{w}^H\mathfrak{B}^T\boldsymbol{w} = \bar{\lambda}\|\boldsymbol{w}\|_2^2$, we get

$$\bar{\lambda}\|\boldsymbol{w}\|_2^2 = u^H A^T u + v^H E^T v - 2\boldsymbol{i}\Im(u^H B^T w + v^H Cw). \tag{3.2.6}$$

By adding (3.2.5) and (3.2.6), we obtain

$$\begin{aligned} (\lambda + \bar{\lambda})\|\boldsymbol{w}\|_2^2 &= u^H(A + A^T)u + v^H(E + E^T)v \\ &= \Re(u)^T(A + A^T)\Re(u) + \Im(u)^T(A + A^T)\Im(u) \\ &\quad + \Re(v)^T(E + E^T)\Re(v) + \Im(v)^T(E + E^T)\Im(v). \end{aligned} \tag{3.2.7}$$

Therefore, from (3.2.7), we obtain

$$\Re(\lambda) = \frac{\Re(u)^T A_H\Re(u) + \Im(u)^T A_H\Im(u) + \Re(v)^T E_H\Re(v) + \Im(v)^T E_H\Im(v)}{\|\boldsymbol{w}\|^2}. \tag{3.2.8}$$

Since A_H and E_H are positive definite matrices, from (3.2.8), we get $\Re(\lambda) \geq 0$ and $\Re(\lambda) = 0$ if and only if $u = \boldsymbol{0}$ and $v = \boldsymbol{0}$.

Next, we show that the vectors u and v can not be zero simultaneously. First, assume that $u = \boldsymbol{0}$. From the first equation in (3.2.4), it follows that $B^T w = \boldsymbol{0}$, which leads to $w = \boldsymbol{0}$, as B has full row rank. Substituting $u = \boldsymbol{0}$ and $w = \boldsymbol{0}$ into the third equation of (3.2.4) yields $v = \boldsymbol{0}$. As a result, we obtain $\boldsymbol{w} = \boldsymbol{0}$, which contradicts the assumption

that $\boldsymbol{w}$ is an eigenvector. Therefore, we conclude that $u \neq \mathbf{0}$, which in turn implies that $\Re(\lambda) > 0$. This completes the proof. ■

To ensure the properties stated in Propositions 3.2.2-3.2.3 are satisfied, throughout the chapter, we assume that A and E are nonsingular matrices, where A_H and E_H are SPD.

3.3. Proposed Generalized Shift-Splitting (GSS) Iterative Method and Preconditioner

This section proposes a GSS iterative method to solve the DSPP. Let $\alpha, \beta, \tau, \boldsymbol{\omega}$ be positive real numbers and $P \in \mathbb{R}^{n\times n}, Q \in \mathbb{R}^{p\times p}$, and $R \in \mathbb{R}^{m\times m}$ be SPD matrices, then $\mathfrak{B}$ admits the following matrix splitting:

$$\mathfrak{B} = (\Theta + \boldsymbol{\omega}\mathfrak{B}) - (\Theta - (1-\boldsymbol{\omega})\mathfrak{B}) =: \mathscr{P}_{\text{GSS}} - \mathcal{N}_{\text{GSS}}, \tag{3.3.1}$$

where

$$\mathscr{P}_{\text{GSS}} = \begin{bmatrix} \alpha P + \boldsymbol{\omega} A & \mathbf{0} & \boldsymbol{\omega} B^T \\ \mathbf{0} & \beta Q + \boldsymbol{\omega} E & \boldsymbol{\omega} C \\ -\boldsymbol{\omega} B & -\boldsymbol{\omega} C^T & \tau R \end{bmatrix}, \tag{3.3.2}$$

$$\mathcal{N}_{\text{GSS}} = \begin{bmatrix} \alpha P - (1-\boldsymbol{\omega})A & \mathbf{0} & (1-\boldsymbol{\omega})B^T \\ \mathbf{0} & \beta Q - (1-\boldsymbol{\omega})E & (1-\boldsymbol{\omega})C \\ -(1-\boldsymbol{\omega})B & -(1-\boldsymbol{\omega})C^T & \tau R \end{bmatrix}, \tag{3.3.3}$$

and $\Theta = \begin{bmatrix} \alpha P & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tau R \end{bmatrix}$.

The special matrix splitting in equation (3.3.1) introduces a novel iteration scheme, known as the GSS iterative method, for solving the DSPP.

Method 3.3.1. (**GSS iterative method**). Given the initial guess vector $\widehat{\boldsymbol{w}}_0 = \left[\boldsymbol{x}_0{}^T, \boldsymbol{z}_0{}^T, \boldsymbol{y}_0{}^T\right]^T$, positive real numbers α, β, τ and $\boldsymbol{\omega}$, and SPD matrices $P \in \mathbb{R}^{n\times n}, Q \in \mathbb{R}^{p\times p}$ and $R \in \mathbb{R}^{m\times m}$, until the stopping criterion is satisfied, compute

$$\widehat{\boldsymbol{w}}_{k+1} = \mathcal{G}\widehat{\boldsymbol{w}}_k + \mathbf{d}, \qquad k = 0, 1, 2, \ldots, \tag{3.3.4}$$

where $\widehat{\boldsymbol{w}}_k = \left[\boldsymbol{x}_k{}^T, \boldsymbol{z}_k{}^T, \boldsymbol{y}_k{}^T\right]^T \in \mathbb{R}^{n+p+m}$, $\mathcal{G} = \mathscr{P}_{\text{GSS}}^{-1}\mathcal{N}_{\text{GSS}}$ is the iteration matrix and $\mathbf{d} = \mathscr{P}_{\text{GSS}}^{-1}\widetilde{\mathbf{d}} \in \mathbb{R}^{n+p+m}$.

The matrix splitting in (3.3.1) induces a preconditioner, denoted as $\mathscr{P}_{\text{GSS}}$, which can be utilized to speed up the convergence rate of the Krylov subspace methods, like GMRES. This preconditioner is referred to as the GSS preconditioner.

At each step of the GSS iterative method or GSS PGMRES method, we are required to solve a system of linear equations in the following form:

$$\mathscr{P}_{\text{GSS}}\boldsymbol{z} = r, \tag{3.3.5}$$

where $\boldsymbol{z} = [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \boldsymbol{z}_3^T]^T \in \mathbb{R}^{n+p+m}$ and $\boldsymbol{r} = [\boldsymbol{r}_1^T, \boldsymbol{r}_2^T, \boldsymbol{r}_3^T]^T \in \mathbb{R}^{n+p+m}$. However, $\mathscr{P}_{\text{GSS}}$ admits the following decomposition:

$$\mathscr{P}_{\text{GSS}} = \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ -\boldsymbol{\omega} B(\alpha P + \boldsymbol{\omega} A)^{-1} & -\boldsymbol{\omega} C^T(\beta Q + \boldsymbol{\omega} E)^{-1} & I \end{bmatrix} \begin{bmatrix} \alpha P + \boldsymbol{\omega} A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta Q + \boldsymbol{\omega} E & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widehat{R} \end{bmatrix} \tag{3.3.6}$$

$$\cdot \begin{bmatrix} I & \mathbf{0} & \boldsymbol{\omega}(\alpha P + \boldsymbol{\omega} A)^{-1} B^T \\ \mathbf{0} & I & \boldsymbol{\omega}(\beta Q + \boldsymbol{\omega} E)^{-1} C \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix},$$

where $\widehat{R} = \tau R + \boldsymbol{\omega}^2 B(\alpha P + \boldsymbol{\omega} A)^{-1} B^T + \boldsymbol{\omega}^2 C^T(\beta Q + \boldsymbol{\omega} E)^{-1} C$. In the following, we present the algorithmic implementations of the GSS preconditioner designed to accelerate the GMRES method.

Algorithm 3.3.1 Solving $\mathscr{P}_{\text{GSS}}\boldsymbol{z} = \boldsymbol{r}$

Input: The matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times m}$, $E \in \mathbb{R}^{p\times p}$, $\boldsymbol{r} \in \mathbb{R}^{n+p+m}$, positive parameters $\alpha, \beta, \tau, \boldsymbol{\omega}$, SPD matrices $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{p\times p}$ and $R \in \mathbb{R}^{m\times m}$.

Output: Solution vector $\boldsymbol{z} = [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \boldsymbol{z}_3^T]^T \in \mathbb{R}^{n+p+m}$.

Steps:

1 : Solve $(\alpha P + \boldsymbol{\omega} A)t_1 = \boldsymbol{r}_1$ to find t_1.

2 : Solve $(\beta Q + \boldsymbol{\omega} E)t_2 = \boldsymbol{r}_2$ to find t_2.

3 : Solve $\widehat{R}\boldsymbol{z}_3 = \boldsymbol{r}_3 + \boldsymbol{\omega} B t_1 + \boldsymbol{\omega} C^T t_2$ to obtain $\boldsymbol{z}_3$.

4 : Solve $(\alpha P + \boldsymbol{\omega} A)\boldsymbol{z}_1 = \boldsymbol{r}_1 - B^T\boldsymbol{z}_3$ to find $\boldsymbol{z}_1$.

5 : Solve $(\beta Q + \boldsymbol{\omega} E)\boldsymbol{z}_2 = \boldsymbol{r}_2 - \boldsymbol{\omega} C\boldsymbol{z}_3$ to obtain $\boldsymbol{z}_2$.

Remark 3.3.1. *Algorithm 3.3.1 necessitates solving two linear subsystems having the coefficient matrix* $(\alpha P+\omega A)$*, two subsystems having the coefficient matrix* $(\beta Q+\boldsymbol{\omega} E)$*, and one subsystem with the coefficient matrix* $\widehat{R}$*. We can use LU factorization to solve them*

efficiently. Moreover, when A and E are SPD matrices, we have the flexibility to employ exact solvers, such as Cholesky factorization, and inexact solvers, like the preconditioned conjugate gradient method, to solve them efficiently. Nevertheless, to solve steps 1 *and* 4*, only one Cholesky or LU factorization of $\alpha P + \boldsymbol{\omega} A$ and to solve steps* 2 *and* 5*, only one Cholesky or LU factorization of $\beta Q + \boldsymbol{\omega} E$ is needed to perform.*

Remark 3.3.2. *From Algorithm 3.3.1, observe that the most tedious task to implement the GSS preconditioner is to solve the linear subsystem in step* 3*. To avoid the direct construction of the matrices $B(\alpha P + \boldsymbol{\omega} A)^{-1} B^T$ and $C^T(\beta Q + \boldsymbol{\omega} E)^{-1} C$, we modify the decomposition in (3.3.6) in the following way:*

$$\widetilde{\mathscr{P}}_{\text{GSS}} := \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ -\boldsymbol{\omega} B(\alpha P + \boldsymbol{\omega} A)^{-1} & -\boldsymbol{\omega} C^T(\beta Q + \boldsymbol{\omega} E)^{-1} & I \end{bmatrix} \begin{bmatrix} \alpha P + \boldsymbol{\omega} A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta Q + \boldsymbol{\omega} E & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \widetilde{P} + \widetilde{Q} \end{bmatrix} \cdot \begin{bmatrix} I & \mathbf{0} & \boldsymbol{\omega}(\alpha P + \boldsymbol{\omega} A)^{-1} B^T \\ \mathbf{0} & I & \boldsymbol{\omega}(\beta Q + \boldsymbol{\omega} E)^{-1} C \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix}, \tag{3.3.7}$$

where $\widetilde{P}$ and $\widetilde{Q}$ are (efficient and economical) approximations of the matrices $B(\alpha P + \boldsymbol{\omega} A)^{-1} B^T$ and $C^T(\beta Q + \boldsymbol{\omega} E)^{-1} C$, respectively. Then, the step 4 *in Algorithm 3.3.1 changes to the linear subsystem $(\widetilde{P} + \widetilde{Q})\boldsymbol{z}_3 = \boldsymbol{r}_3 + \boldsymbol{\omega} B t_1 + \boldsymbol{\omega} C^T t_2$. With suitably chosen $\widetilde{P}$ and $\widetilde{Q}$ (see for Example* 1*), this subsystem is much easier to implement than step* 3 *of Algorithm 3.3.1. We denote this inexact version of the GSS preconditioner as $\widetilde{P}_{GSS}$.*

3.4. Convergence Analysis of the GSS Iterative Method

The purpose of this section is to investigate the convergence behavior of the proposed GSS iterative method. To achieve this, the following result plays a crucial role.

Lemma 3.4.1. *Let $A \in \mathbb{R}^{n \times n}$ and $E \in \mathbb{R}^{p \times p}$ with A_H and E_H positive definite matrices, $B \in \mathbb{R}^{m \times n}$ and $C \in \mathbb{R}^{p \times m}$ be full row matrices. Then, the matrix $\Theta^{-1}\mathfrak{B}$ is positive stable.*

Proof. Since Θ is SPD, $\Theta^{-1}\mathfrak{B}$ is similar to the matrix $\Theta^{-\frac{1}{2}}\mathfrak{B}\Theta^{-\frac{1}{2}}$. By computation, we find that the block structure of the matrix $\Theta^{-\frac{1}{2}}\mathfrak{B}\Theta^{-\frac{1}{2}}$ is identical to that of $\mathfrak{B}$. Therefore, using Proposition 3.2.3, it follows that $\Theta^{-\frac{1}{2}}\mathfrak{B}\Theta^{-\frac{1}{2}}$, and hence, $\Theta^{-1}\mathfrak{B}$ is positive stable. ■

As noted in Lemma 1.2.1, a stationary iterative method in the form (3.3.4) converges if and only if the iteration matrix has the spectral radius strictly less than one. The following result discusses the convergence of the GSS iterative method (3.3.4).

Theorem 3.4.2. *Assume that $A \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{p\times p}$, where A_H and E_H are SPD matrices, $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ are full row matrices. Let $\alpha, \beta, \tau, \boldsymbol{\omega} > 0$, $P \in \mathbb{R}^{n\times n}, Q \in \mathbb{R}^{p\times p}$ and $R \in \mathbb{R}^{m\times m}$ be positive definite matrices. Then, the GSS iterative method converges to the unique solution of the DSPP (3.1.1) if*

$$\boldsymbol{\omega} \geq \max\left\{\frac{1}{2} - \frac{\lambda_{\min}(\widetilde{\Theta})}{\vartheta(\Theta^{-1}\mathfrak{B})^2}, 0\right\},$$

where $\widetilde{\Theta} = \dfrac{\Theta^{-1}\mathfrak{B} + \mathfrak{B}^T\Theta^{-1}}{2}$.

Proof. The iteration matrix of the GSS iterative method (3.3.4) is

$$\mathcal{G} = \mathscr{P}_{\text{GSS}}^{-1}\mathcal{N}_{\text{GSS}} = (\Theta + \boldsymbol{\omega}\mathfrak{B})^{-1}(\Theta(1-\boldsymbol{\omega})\mathfrak{B}) \tag{3.4.1}$$

$$= (I + \boldsymbol{\omega}\Theta^{-1}\mathfrak{B})^{-1}(I - (1-\boldsymbol{\omega})\Theta^{-1}\mathfrak{B}). \tag{3.4.2}$$

Let λ be an eigenvalue of $\mathcal{G}$. Then

$$(I + \boldsymbol{\omega}\Theta^{-1}\mathfrak{B})^{-1}\left(I - (1-\boldsymbol{\omega})\Theta^{-1}\mathfrak{B}\right)\mathbf{x} = \lambda\mathbf{x}, \tag{3.4.3}$$

where $\mathbf{x} \in \mathbb{R}^{n+p+m}$ is the corresponding eigenvector. From (3.4.3), we write

$$(I - (1-\boldsymbol{\omega})\Theta^{-1}\mathfrak{B})\mathbf{x} = \lambda(I + \boldsymbol{\omega}\Theta^{-1}\mathfrak{B})\mathbf{x} \tag{3.4.4}$$

$$\implies ((1-\boldsymbol{\omega}) + \lambda\boldsymbol{\omega})\Theta^{-1}\mathfrak{B}\mathbf{x} = (1-\lambda)\mathbf{x}. \tag{3.4.5}$$

Note that $\lambda \neq 1$, otherwise (3.4.5) reduces to $\Theta^{-1}\mathfrak{B}\mathbf{x} = \mathbf{0}$, which implies that $\mathbf{x} = \mathbf{0}$. On the other hand, $(1-\boldsymbol{\omega}) + \lambda\boldsymbol{\omega} \neq 0$, otherwise $(1-\lambda)\mathbf{x} = \mathbf{0}$. This gives $\mathbf{x} = \mathbf{0}$, which contradicts the assumption that $\mathbf{x}$ is an eigenvector. Therefore, from (3.4.5) we get

$$\Theta^{-1}\mathfrak{B}\mathbf{x} = \frac{1-\lambda}{(1-\boldsymbol{\omega}) + \lambda\boldsymbol{\omega}}\mathbf{x}. \tag{3.4.6}$$

Thus $\theta := \frac{1-\lambda}{(1-\boldsymbol{\omega})+\lambda\boldsymbol{\omega}}$ is an eigenvalue of $\Theta^{-1}\mathfrak{B}$. Further, we can write

$$\lambda = \frac{1 - (1-\boldsymbol{\omega})\theta}{1 + \boldsymbol{\omega}\theta}.$$

Therefore, $|\lambda| < 1$ if and only if $|1 - (1-\boldsymbol{\omega})\theta| < |1 + \boldsymbol{\omega}\theta|$, i.e.,

$$(1 - (1-\boldsymbol{\omega})\Re(\theta))^2 + (1-\boldsymbol{\omega})^2\Im(\theta)^2 < (1 + \boldsymbol{\omega}\Re(\theta))^2 + \boldsymbol{\omega}^2\Im(\theta). \tag{3.4.7}$$

Consequently, it follows from (3.4.7) that the iterative method (3.3.4) is convergent if

$$2\Re(\theta) + (2\boldsymbol{\omega} - 1)|\theta|^2 > 0. \tag{3.4.8}$$

By Lemma 3.4.1, we have $\Re(\theta) > 0$, and this implies $\frac{\Re(\theta)}{|\theta|^2} > 0$. From (3.4.8), we get $\boldsymbol{\omega} > \frac{1}{2} - \frac{\Re(\theta)}{|\theta|^2}$.

Next, assume that $\boldsymbol{w}$ is the eigenvector corresponding to the eigenvalue θ. Then $\Theta^{-1}\mathfrak{B}\boldsymbol{w} = \theta\boldsymbol{w}$. Multiplying by $\boldsymbol{w}^H$ from the left side, we have $\boldsymbol{w}^H\Theta^{-1}\mathfrak{B}\boldsymbol{w} = \theta\boldsymbol{w}^H\boldsymbol{w}$, and taking conjugate transpose gives $\boldsymbol{w}^H\mathfrak{B}^T\Theta^{-1}\boldsymbol{w} = \bar{\theta}\boldsymbol{w}^H\boldsymbol{w}$. Hence,

$$\Re(\theta) = \frac{\boldsymbol{w}^H(\Theta^{-1}B + B^T\Theta^{-1})\boldsymbol{w}}{2\boldsymbol{w}^H\boldsymbol{w}} \geq \lambda_{\min}(\widetilde{\Theta}).$$

Again $|\theta| \leq \vartheta(\Theta)$ gives $\frac{1}{2} - \frac{\Re(\theta)}{|\theta|^2} \leq \frac{1}{2} - \frac{\lambda_{\min}(\widetilde{\Theta})}{\vartheta(\Theta)^2}$. Since $\boldsymbol{\omega} > 0$, the GSS iterative method is convergent if $\boldsymbol{\omega} > \max\left\{\frac{1}{2} - \frac{\lambda_{\min}(\widetilde{\Theta})}{\vartheta(\Theta)^2}, 0\right\}$. ■

Remark 3.4.3. *Note that if $\boldsymbol{\omega} \geq \frac{1}{2}$, then the condition (3.4.7) holds. This shows that the GSS iterative method (3.3.4) is convergent for any initial guess vector if $\boldsymbol{\omega} \geq \frac{1}{2}$.*

Notably, solving the DSPP (3.1.1) is same as to solve the preconditioned linear system $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}\widehat{\boldsymbol{w}} = \mathscr{P}_{\text{GSS}}^{-1}\widetilde{\mathbf{d}}$. Hence, as an immediate consequence of Theorem 3.4.2 and Remark 3.4.3, we have the following results regarding the clustering properties of the spectrum of the preconditioned matrix $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$.

Theorem 3.4.4. *Assume that $A \in \mathbb{R}^{n\times n}$, $E \in \mathbb{R}^{p\times p}$ are nonsingular matrices with A_H and E_H are positive definite, $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ are full row rank matrices. Let $\mathscr{P}_{\text{GSS}}$ be defined as in (3.3.2) and λ be an eigenvalue of the preconditioned matrix $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$. If $\boldsymbol{\omega} \geq 1/2$, then λ satisfies the following:*

$$|\lambda - 1| < 1,$$

i.e., all eigenvalues of the preconditioned matrix $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$ are entirely contained in a circle centered at $(1, 0)$ with radius strictly less one.

Proof. The proof follows immediately from the identity $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B} = \mathscr{P}_{\text{GSS}}^{-1}(\mathscr{P}_{\text{GSS}} - \mathcal{N}_{\text{GSS}}) = I - \mathcal{G}$. ■

3.5. Two Relaxed Variants of GSS Preconditioner

To enhance efficiency, this section introduces two relaxed variants of the GSS preconditioner. By removing αP from $(1, 1)$ block and βQ from $(2, 2)$ block of $\mathscr{P}_{\text{GSS}}$, we obtain the following two relaxed GSS (RGSS) preconditioners:

$$\mathscr{P}_{\text{RGSS-I}} = \begin{bmatrix} \boldsymbol{\omega}A & \mathbf{0} & \boldsymbol{\omega}B^T \\ \mathbf{0} & \beta Q + \boldsymbol{\omega}E & \boldsymbol{\omega}C \\ -\boldsymbol{\omega}B & -\boldsymbol{\omega}C^T & \tau R \end{bmatrix} \tag{3.5.1}$$

and

$$\mathscr{P}_{\text{RGSS-II}} = \begin{bmatrix} \boldsymbol{\omega} A & \mathbf{0} & \boldsymbol{\omega} B^T \\ \mathbf{0} & \boldsymbol{\omega} E & \boldsymbol{\omega} C \\ -\boldsymbol{\omega} B & -\boldsymbol{\omega} C^T & \tau R \end{bmatrix}. \tag{3.5.2}$$

In the implementation, at each step of RGSS preconditioners in conjunction with GMRES, we are required to solve the linear systems of the following forms:

$$\mathscr{P}_{\text{RGSS-I}} \boldsymbol{w} = \boldsymbol{r} \quad \text{or} \quad \mathscr{P}_{\text{RGSS-II}} \boldsymbol{w} = \boldsymbol{r}. \tag{3.5.3}$$

Set $\mathcal{R}_1 = \tau R + \boldsymbol{\omega} B A^{-1} B^T + \boldsymbol{\omega}^2 C^T (\beta Q + \boldsymbol{\omega} E)^{-1} C$ and $\mathcal{R}_2 = \tau R + \boldsymbol{\omega} B A^{-1} B^T + \boldsymbol{\omega} C^T E^{-1} C$, then

$$\mathscr{P}_{\text{RGSS-I}} = \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ -BA^{-1} & -\boldsymbol{\omega} C^T (\beta Q + \boldsymbol{\omega} E)^{-1} & I \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega} A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta Q + \boldsymbol{\omega} E & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{R}_1 \end{bmatrix} \begin{bmatrix} I & \mathbf{0} & A^{-1} B^T \\ \mathbf{0} & I & \boldsymbol{\omega} (\beta Q + \boldsymbol{\omega} E)^{-1} C \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix}, \tag{3.5.4}$$

$$\mathscr{P}_{\text{RGSS-II}} = \begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ -BA^{-1} & -C^T E^{-1} & I \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega} A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\omega} E & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{R}_2 \end{bmatrix} \begin{bmatrix} I & \mathbf{0} & A^{-1} B^T \\ \mathbf{0} & I & E^{-1} C \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix}. \tag{3.5.5}$$

By applying a similar methodology as in Algorithm 3.3.1, we derive Algorithms 3.5.1 and **??** for implementing the RGSS preconditioners.

Algorithm 3.5.1 Solving $\mathscr{P}_{\text{RGSS-I}} \boldsymbol{w} = \boldsymbol{r}$

Input: The matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{p \times m}$, $E \in \mathbb{R}^{p \times p}$, $\boldsymbol{r} \in \mathbb{R}^{n+p+m}$, positive parameters $\beta, \tau, \boldsymbol{\omega}$, and SPD matrices $Q \in \mathbb{R}^{p \times p}$ and $R \in \mathbb{R}^{m \times m}$.

Output: Solution vector $\boldsymbol{w} = [\boldsymbol{w}_1^T, \boldsymbol{w}_2^T, \boldsymbol{w}_3^T]^T \in \mathbb{R}^{n+p+m}$.

Steps:

1 : Solve $At_1 = r_1/\boldsymbol{\omega}$ to find t_1.

2 : Solve $(\beta Q + \boldsymbol{\omega} E) t_2 = r_2$ to find t_2.

3 : Solve $\mathcal{R}_1 \boldsymbol{w}_3 = r_3 + \boldsymbol{\omega} B t_1 + \boldsymbol{\omega} C^T t_2$ to obtain $\boldsymbol{w}_3$.

4 : Solve $A\boldsymbol{w}_1 = \frac{1}{\boldsymbol{\omega}}(r_1 - B^T \boldsymbol{w}_3)$ to find $\boldsymbol{w}_1$.

5 : Solve $(\beta Q + \boldsymbol{\omega} E)\boldsymbol{w}_2 = r_2 - \boldsymbol{\omega} C \boldsymbol{w}_3$ to obtain $\boldsymbol{w}_2$.

Algorithm 3.5.2 Solving $\mathscr{P}_{\text{RGSS-II}}\boldsymbol{w} = \boldsymbol{r}$

Input: The matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times m}$, $E \in \mathbb{R}^{p\times p}$, $\boldsymbol{r} \in \mathbb{R}^{n+p+m}$, positive parameters $\tau, \boldsymbol{\omega}$ and the SPD matrix $R \in \mathbb{R}^{m\times m}$.

Output: Solution vector $\boldsymbol{w} = [\boldsymbol{w}_1^T, \boldsymbol{w}_2^T, \boldsymbol{w}_3^T]^T \in \mathbb{R}^{n+p+m}$.

Steps:

1 : Solve $At_1 = r_1/\boldsymbol{\omega}$ to find t_1.

2 : Solve $Et_2 = r_2/\boldsymbol{\omega}$ to find t_2.

3 : Solve $\mathcal{R}_2\boldsymbol{w}_3 = r_3 + \boldsymbol{\omega}Bt_1 + \boldsymbol{\omega}C^Tt_2$ to obtain $\boldsymbol{w}_3$.

4 : Solve $A\boldsymbol{w}_1 = \frac{1}{\boldsymbol{\omega}}(r_1 - B^T\boldsymbol{w}_3)$ to find $\boldsymbol{w}_1$.

5 : Solve $E\boldsymbol{w}_2 = \frac{1}{\boldsymbol{\omega}}(r_2 - \boldsymbol{\omega}C\boldsymbol{w}_3)$ to obtain $\boldsymbol{w}_2$.

Remark 3.5.1. *A key challenge in implementing the RGSS-I and RGSS-II preconditioners lies in solving the linear subsystems associated with the coefficient matrices $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. As noted in Remark 3.3.2, to avoid the direct computation $BA^{-1}B^T$ and $C^T(\beta Q + \boldsymbol{\omega}E)^{-1}C$ in Algorithm 3.5.1, and $BA^{-1}B^T$ and $C^TE^{-1}C$ in Algorithm 3.5.2, we can use approximate versions of these terms. Following a similar technique as in decomposition (3.3.7), let $\widetilde{P}_1$, $\widetilde{Q}_1$ and $\widetilde{Q}_2$ be efficient and economical approximations of the matrices $BA^{-1}B^T$, $C^T(\beta Q+\boldsymbol{\omega}E)^{-1}C$, and $C^TE^{-1}C$, respectively. With these approximations, step 4 in Algorithms 3.5.1 and 3.5.2 transforms into solving the linear subsystems $(\widetilde{P}_1 + \widetilde{Q}_1)\boldsymbol{z}_3 = \boldsymbol{r}_3 + \boldsymbol{\omega}Bt_1 + \boldsymbol{\omega}C^Tt_2$ and $(\widetilde{P}_1 + \widetilde{Q}_2)\boldsymbol{z}_3 = \boldsymbol{r}_3 + \boldsymbol{\omega}Bt_1 + \boldsymbol{\omega}C^Tt_2$, respectively. By appropriately selecting $\widetilde{P}_1$ $\widetilde{Q}_1$, and $\widetilde{Q}_2$ (see for Example 1), these subsystems become significantly easier to implement compared to step 3 of Algorithms 3.5.1 and 3.5.2. The resulting inexact preconditioners are denoted by $\widetilde{\mathscr{P}}_{RGSS\text{-}I}$ and $\widetilde{\mathscr{P}}_{RGSS\text{-}II}$, respectively.*

Next, we investigate the spectral properties of the preconditioned matrices $\mathscr{P}^{-1}_{\text{RGSS-I}}\mathfrak{B}$ and $\mathscr{P}^{-1}_{\text{RGSS-II}}\mathfrak{B}$ by considering A and E are SPD matrices.

Theorem 3.5.2. *Assume that A and E are SPD matrices, B and C have full row rank, and let Q and R be SPD matrices. Then, the preconditioned matrix $\mathscr{P}^{-1}_{\text{RGSS-I}}\mathfrak{B}$ has $\dfrac{1}{\boldsymbol{\omega}}$ as the eigenvalue with multiplicity n. Further, let μ be an eigenvalue among the remaining $m + p$ eigenvalues with the corresponding eigenvector $[u^T, v^T]^T$ such that $\|\sqrt{\beta}Q^{1/2}u^T, \sqrt{\tau}R^{1/2}v^T]^T\|_2 = 1$. Then*

(1) if $\Im(\frac{\mu}{1-\boldsymbol{\omega}\mu}) \neq 0$, we have $\beta\|Q^{1/2}u\|_2^2 = \frac{1}{2} = \tau\|R^{1/2}v\|_2^2$ and

$$\Re\left(\frac{\mu}{1-\boldsymbol{\omega}\mu}\right) = \frac{1}{2}\left(\frac{u^HEu}{\beta u^HQu} + \frac{v^HSv}{\tau v^HRv}\right).$$

Thus, it satisfies the following bounds:

$$\begin{cases} \frac{1}{2}\left(\frac{\lambda_{\min}(Q^{-1}E)}{\beta}+\frac{\lambda_{\min}(R^{-1}S)}{\tau}\right) \le \Re\left(\frac{\mu}{1-\boldsymbol{\omega}\mu}\right) \le \frac{1}{2}\left(\frac{\lambda_{\max}(Q^{-1}E)}{\beta}+\frac{\lambda_{\max}(R^{-1}S)}{\tau}\right), \\ |\Im\left(\frac{\mu}{1-\boldsymbol{\omega}\mu}\right)| \le \sigma_{\max}\left(\frac{1}{\sqrt{\beta\tau}}R^{-\frac{1}{2}}CQ^{-\frac{1}{2}}\right). \end{cases} \tag{3.5.6}$$

(2) If $\Im(\frac{\mu}{1-\boldsymbol{\omega}\mu})=0$, *we have*

$$\frac{\mu}{1-\boldsymbol{\omega}\mu}=\frac{u^HEu+v^HSv}{\beta u^HQu+\tau v^HRv}$$

and it holds that:

$$2\min\left\{\frac{\lambda_{\min}(Q^{-1}E)}{\beta},\frac{\lambda_{\min}(R^{-1}S)}{\tau}\right\} \le \frac{\mu}{1-\boldsymbol{\omega}\mu} \le \max\left\{\frac{\lambda_{\max}(Q^{-1}E)}{\beta},\frac{\lambda_{\max}(R^{-1}S)}{\tau}\right\}. \tag{3.5.7}$$

Proof. Let $S=BA^{-1}B^T$. Then, the matrix $\mathfrak{B}$ and the preconditioner $\mathscr{P}_{\text{RGSS-I}}$ admits the following decompositions:

$$\mathfrak{B}=\mathbf{LEU} \quad \text{and} \quad \mathscr{P}_{\text{RGSS-I}}=\mathbf{L}\widetilde{\mathbf{E}}\mathbf{U}, \tag{3.5.8}$$

where

$$\mathbf{L}=\begin{bmatrix} I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} \\ -BA^{-1} & \mathbf{0} & I \end{bmatrix}, \mathbf{E}=\begin{bmatrix} A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & E & C \\ \mathbf{0} & -C^T & S \end{bmatrix}, \widetilde{\mathbf{E}}=\begin{bmatrix} \boldsymbol{\omega}A & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\omega}E+\beta Q & \boldsymbol{\omega}C \\ \mathbf{0} & -\boldsymbol{\omega}C^T & S+\tau R \end{bmatrix},$$

$$\text{and} \quad \mathbf{U}=\begin{bmatrix} I & \mathbf{0} & A^{-1}B^T \\ \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I \end{bmatrix}.$$

Using decompositions in (3.5.8), we obtain

$$\mathscr{P}^{-1}_{\text{RGSS-I}}\mathfrak{B}=\mathbf{U}^{-1}\widetilde{\mathbf{E}}^{-1}\mathbf{E}\mathbf{U}. \tag{3.5.9}$$

Therefore, $\mathscr{P}^{-1}_{\text{RGSS-I}}\mathfrak{B}$ is similar to $\widetilde{\mathbf{E}}^{-1}\mathbf{E}$, which is given by

$$\widetilde{\mathbf{E}}^{-1}\mathbf{E}=\begin{bmatrix} \boldsymbol{\omega}^{-1}I & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^{-1}\mathbf{A} \end{bmatrix}, \tag{3.5.10}$$

where $\mathbf{M}=\begin{bmatrix} \boldsymbol{\omega}E+\beta Q & \boldsymbol{\omega}C \\ -\boldsymbol{\omega}C^T & \boldsymbol{\omega}S+\tau R \end{bmatrix}$ and $\mathbf{A}=\begin{bmatrix} E & C \\ -C^T & S \end{bmatrix}$. Hence, $\mathscr{P}^{-1}_{\text{RGSS-I}}\mathfrak{B}$ has an eigenvalue $\dfrac{1}{\boldsymbol{\omega}}$ with multiplicity at least n and while the remaining eigenvalues are those of the preconditioned matrix $\mathbf{M}^{-1}\mathbf{A}$. Consider $\mathbf{I}=\begin{bmatrix} \mathbf{0} & I \\ -I & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{p+m}$, then $\mathbf{M}^{-1}\mathbf{A}=$

$\mathbf{I}\left(\mathbf{I}^{-1}\mathbf{M}\mathbf{I}\right)^{-1}\left(\mathbf{I}^{-1}\mathbf{A}\mathbf{I}\right)\mathbf{I}^{-1}$. Consequently, $\mathbf{M}^{-1}\mathbf{A}$ is similar to $\widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{A}}$, where

$$\widetilde{\mathbf{M}} := \mathbf{I}^{-1}\mathbf{M}\mathbf{I} = \begin{bmatrix} \boldsymbol{\omega}E + \beta Q & \boldsymbol{\omega}C^T \\ -\boldsymbol{\omega}C & \boldsymbol{\omega}S + \tau R \end{bmatrix} \text{ and } \widetilde{\mathbf{A}} := \mathbf{I}^{-1}\mathbf{A}\mathbf{I} = \begin{bmatrix} E & C^T \\ -C & S \end{bmatrix}. \quad (3.5.11)$$

Let μ be an eigenvalue of the matrix $\widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{A}}$ with the corresponding eigenvalue $[u^T, v^T]^T$. Assert that $\mu = \frac{1}{\boldsymbol{\omega}}$, then

$$\begin{bmatrix} \beta Q & \mathbf{0} \\ \mathbf{0} & \tau R \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \mathbf{0}, \quad (3.5.12)$$

which gives $[u^T, v^T]^T = \mathbf{0}$. This contradicts to the assumption that $[u^T, v^T]^T$ is an eigenvector, and hence $\mu \neq \frac{1}{\boldsymbol{\omega}}$. Since, μ is an eigenvalue of $\widetilde{\mathbf{M}}^{-1}\widetilde{\mathbf{A}}$, we have

$$\begin{aligned}
&\mu \begin{bmatrix} \boldsymbol{\omega}E + \beta Q & \boldsymbol{\omega}C^T \\ -\boldsymbol{\omega}C & \boldsymbol{\omega}S + \tau R \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} E & C^T \\ -C & S \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \\
&\Longrightarrow \frac{\mu}{1-\boldsymbol{\omega}\mu} \begin{bmatrix} \beta Q & \mathbf{0} \\ \mathbf{0} & \tau R \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} E & C^T \\ -C & S \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \\
&\Longrightarrow \frac{\mu}{1-\boldsymbol{\omega}\mu} \begin{bmatrix} \sqrt{\beta}Q^{\frac{1}{2}}u \\ \sqrt{\tau}R^{\frac{1}{2}}v \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\beta}}Q^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{\beta}}R^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} E & C^T \\ -C & S \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\beta}}Q^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{\tau}}R^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \sqrt{\beta}Q^{\frac{1}{2}}u \\ \sqrt{\tau}R^{\frac{1}{2}}v \end{bmatrix} \\
&\Longrightarrow \frac{\mu}{1-\boldsymbol{\omega}\mu} \begin{bmatrix} \widetilde{u} \\ \widetilde{v} \end{bmatrix} = \begin{bmatrix} \widetilde{E} & \widetilde{C}^T \\ -\widetilde{C} & \widetilde{S} \end{bmatrix} \begin{bmatrix} \widetilde{u} \\ \widetilde{v} \end{bmatrix},
\end{aligned}$$

where $\widetilde{E} = \frac{1}{\beta}Q^{-\frac{1}{2}}EQ^{-\frac{1}{2}}$, $\widetilde{C} = \frac{1}{\sqrt{\beta\tau}}R^{-\frac{1}{2}}CQ^{-\frac{1}{2}}$, $\widetilde{S} = \frac{1}{\tau}R^{-\frac{1}{2}}SR^{-\frac{1}{2}}$, $\widetilde{u} = \sqrt{\beta}Q^{\frac{1}{2}}u$ and $\widetilde{v} = \sqrt{\tau}R^{\frac{1}{2}}v$. Since, $\widetilde{E}$ and $\widetilde{S}$ are SPD matrices and $\widetilde{C}$ has full row rank, according to Proposition 2.12 in [25], we have the desired bounds of (3.5.6) and (3.5.7). ■

Remark 3.5.3. *The asymptotic behavior of the eigenvalue μ is analyzed according to Theorem 3.5.2 when the iteration parameters β and τ approach zero from the positive side. Let $\theta = \frac{\mu}{1-\boldsymbol{\omega}\mu}$, then this analysis is conducted in the following two cases:*

- *When $\Im(\theta) \neq 0$, we have $\Re(\theta) \to +\infty$ and $|\Im(\theta)| \to +\infty$ as $\beta, \tau \to 0_+$. Then,*

$$\mu = \left[\frac{1}{\boldsymbol{\omega}} - \frac{1 + \boldsymbol{\omega}\Re(\theta)}{(1+\boldsymbol{\omega}\Re(\theta))^2 + \boldsymbol{\omega}^2\Im(\theta)^2}\right] + \boldsymbol{i}\frac{\Im(\theta)}{(1+\boldsymbol{\omega}\Re(\theta))^2 + \boldsymbol{\omega}^2\Im(\theta)^2} \to \frac{1}{\boldsymbol{\omega}} \text{ as } \beta, \tau \to 0_+.$$

- *When $\Im(\theta) = 0$, we have $\theta \to +\infty$ as $\beta, \tau \to 0_+$. Then,*

$$\mu = \frac{\theta}{1+\boldsymbol{\omega}\theta} \to \frac{1}{\boldsymbol{\omega}}, \text{ as } \beta, \tau \to 0_+. \quad (3.5.13)$$

The following theorem establishes the spectral properties for the RGSS-II preconditioned matrix $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$.

Theorem 3.5.4. *Assume that A and E are SPD matrices, B and C have full row rank, and let R be an SPD matrix. Then the preconditioned matrix $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ has the eigenvalue $\dfrac{1}{\boldsymbol{\omega}}$ with multiplicity $n+p$. The remaining eigenvalues satisfy the generalized eigenvalue problem $(BA^{-1}B^T+C^TE^{-1}C+\tau R)\mathbf{x}=\lambda\mathcal{R}_2\mathbf{x}$, where $\mathcal{R}_2=\tau R+\boldsymbol{\omega}BA^{-1}B^T+\boldsymbol{\omega}C^TE^{-1}C$.*

Proof. From (3.5.5), we have

$$\begin{aligned}\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B} &= \begin{bmatrix}\boldsymbol{\omega}^{-1}A & \mathbf{0} & -A^{-1}B^T\mathcal{R}_2^{-1}\\ \mathbf{0} & \boldsymbol{\omega}^{-1}E & -E^{-1}C\mathcal{R}_2^{-1}\\ \mathbf{0} & \mathbf{0} & \mathcal{R}_2\end{bmatrix}\begin{bmatrix}A & \mathbf{0} & B^T\\ \mathbf{0} & E & C\\ \mathbf{0} & \mathbf{0} & \mathbf{X}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{\omega}^{-1}I & \mathbf{0} & \boldsymbol{\omega}^{-1}AB^T-A^{-1}B^T\mathcal{R}_2^{-1}\mathbf{X}\\ \mathbf{0} & \boldsymbol{\omega}^{-1}I & \boldsymbol{\omega}^{-1}EC-E^{-1}C\mathcal{R}_2^{-1}\mathbf{X}\\ \mathbf{0} & \mathbf{0} & \mathcal{R}_2^{-1}\mathbf{X}\end{bmatrix},\end{aligned} \tag{3.5.14}$$

where $\mathbf{X}=BA^{-1}B^T+C^TE^{-1}C+\tau R$. Thus, $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ has the eigenvalue $\lambda=\dfrac{1}{\boldsymbol{\omega}}$ with multiplicity at least $n+p$, and the rest of eigenvalues satisfy the generalized eigenvalue problem

$$(BA^{-1}B^T+C^TE^{-1}C+\tau R)\mathbf{x}=\lambda\mathcal{R}_2\mathbf{x}.$$

Hence, the proof is completed. ■

Corollary 3.5.1. *Suppose that the assumptions on Theorem 3.5.4 hold. Then, the eigenvalues of $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ satisfy*

$$\lambda\in[\Lambda_{\min},\Lambda_{\max}], \tag{3.5.15}$$

where

$$\Lambda_{\min}=\frac{\eta_{\min}+\xi_{\min}+\tau}{\boldsymbol{\omega}\eta_{\max}+\boldsymbol{\omega}\xi_{\max}+\tau},\quad \Lambda_{\max}=\frac{\eta_{\max}+\xi_{\max}+\tau}{\boldsymbol{\omega}\eta_{\min}+\boldsymbol{\omega}\xi_{\min}+\tau},$$

$\eta_{\min}=\lambda_{\min}(R^{-1}BA^{-1}B^T)$, $\eta_{\max}=\lambda_{\max}(R^{-1}BA^{-1}B^T)$, $\xi_{\min}=\lambda_{\min}(R^{-1}C^TE^{-1}C)$ *and* $\xi_{\max}=\lambda_{\max}(R^{-1}C^TE^{-1}C)$.

Proof. Premultiplying by $\mathbf{x}^T$ of the generalized eigenvalue problem $(BA^{-1}B^T+C^TE^{-1}C+\tau R)\mathbf{x}=\lambda\mathcal{R}_2\mathbf{x}$, we obtain

$$\begin{aligned}\lambda &= \frac{\mathbf{x}^T(BA^{-1}B^T+C^TE^{-1}C+\tau R)\mathbf{x}}{\mathbf{x}^T(\boldsymbol{\omega}BA^{-1}B^T+\boldsymbol{\omega}C^TE^{-1}C+\tau R)\mathbf{x}}\\ &= \frac{(R^{1/2}\mathbf{x})^TR^{-1/2}(BA^{-1}B^T+C^TE^{-1}C+\tau)R^{-1/2}(R^{1/2}\mathbf{x})}{(R^{1/2}\mathbf{x})^TR^{-1/2}(\boldsymbol{\omega}BA^{-1}B^T+\boldsymbol{\omega}C^TE^{-1}C+\tau)R^{-1/2}(R^{1/2}\mathbf{x})}.\end{aligned} \tag{3.5.16}$$

Since $R^{-1/2}BA^{-1}B^TR^{-1/2}$ is similar to $R^{-1}BA^{-1}B^T$ and $R^{-1/2}C^TA^{-1}CR^{-1/2}$ is similar to $R^{-1}C^TA^{-1}C$, and $\lambda_{\min}(X) \leq \theta \leq \lambda_{\max}(X)$ for any $\theta \in \sigma(X)$, where X is SPD, the proof follows from (3.5.16). ■

Next, we will examine the properties of the minimal polynomial of the preconditioned matrix $\mathscr{P}^{-1}_{\text{RGSS-II}}\mathfrak{B}$, which determines the dimension of the Krylov subspace.

Theorem 3.5.5. *Assume that $A \in \mathbb{R}^{n\times n}$ and $E \in \mathbb{R}^{p\times p}$ with A_H and E_H are SPD matrices, $B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times m}$ are full row rank matrices. Then, the degree of the minimal polynomial of the preconditioned matrix $\mathscr{P}^{-1}_{\text{RGSS}-\text{II}}\mathfrak{B}$ is at most $m+1$. Therefore, the dimension of the Krylov subspace $\mathcal{K}(\mathscr{P}^{-1}_{\text{RGSS}-\text{II}}\mathfrak{B}, \widetilde{\mathbf{d}})$ is at most $m+1$.*

Proof. From (3.5.14), we obtain

$$\mathscr{P}^{-1}_{\text{RGSS-II}}\mathfrak{B} = \begin{bmatrix} \boldsymbol{\omega}^{-1}I & \mathbf{0} & \Sigma_1 \\ \mathbf{0} & \boldsymbol{\omega}^{-1}I & \Sigma_2 \\ \mathbf{0} & \mathbf{0} & \Sigma_3 \end{bmatrix}, \tag{3.5.17}$$

where $\Sigma_1 = \boldsymbol{\omega}^{-1}AB^T - A^{-1}B^T\mathcal{R}_2^{-1}\mathbf{X}$, $\Sigma_2 = \boldsymbol{\omega}^{-1}EC - E^{-1}C\mathcal{R}_2^{-1}\mathbf{X}$, $\Sigma_3 = \mathcal{R}_2^{-1}\mathbf{X}$, and $\mathbf{X} = BA^{-1}B^T + C^TE^{-1}C + R$. Let μ_i, $i = 1, 2, \ldots, m$, be the eigenvalues of of Σ_3. Then they are also eigenvalues of the preconditioned matrix $\mathscr{P}^{-1}_{\text{RGSS-II}}\mathfrak{B}$. Then the characteristic polynomial of $\mathscr{P}^{-1}_{\text{RGSS-II}}\mathfrak{B}$ is given by

$$f(\lambda) = \left(\lambda - \frac{1}{\boldsymbol{\omega}}\right)^{n+p}\prod_{i=1}^{m}(\lambda - \mu_i).$$

Consider the polynomial $g(\lambda) = (\lambda - \frac{1}{\boldsymbol{\omega}})\prod_{i=1}^{m}(\lambda - \mu_i)$. Then

$$\begin{aligned} g(\mathscr{P}^{-1}_{\text{GSS-II}}\mathfrak{B}) &= (\mathscr{P}^{-1}_{\text{GSS-II}}\mathfrak{B} - \frac{1}{\boldsymbol{\omega}}I)\prod_{i=1}^{m}(\mathscr{P}^{-1}_{\text{GSS-II}}\mathfrak{B} - \mu_i I) \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} & \Sigma_1 \\ \mathbf{0} & \mathbf{0} & \Sigma_2 \\ \mathbf{0} & \mathbf{0} & \Sigma_3 - \boldsymbol{\omega}^{-1}I \end{bmatrix} \begin{bmatrix} \prod_{i=1}^{n}(\boldsymbol{\omega}^{-1}I - \mu_i I) & \mathbf{0} & \Sigma_1 \\ \mathbf{0} & \prod_{i=1}^{p}(\boldsymbol{\omega}^{-1}I - \mu_i I) & \Sigma_2 \\ \mathbf{0} & \mathbf{0} & \prod_{i=1}^{m}(\Sigma_3 - \mu_i I) \end{bmatrix}. \end{aligned} \tag{3.5.18}$$

Given that Σ_3 has the eigenvalues μ_i, $i = 1, 2, \ldots, m$, we obtain $\prod_{i=1}^{m}(\Sigma_3 - \mu_i I) = \mathbf{0}$. Therefore, from (3.5.18), we obtain $g(\mathscr{P}^{-1}_{\text{RGSS-II}}\mathfrak{B}) = \mathbf{0}$. Hence, by Cayley-Hamilton theorem, the degree of the minimal polynomial of $\mathscr{P}^{-1}_{\text{RGSS-II}}\mathfrak{B}$ is at most $m+1$.

As mentioned in [122], the degree of the minimal polynomial of a matrix and the dimension of the associated Krylov subspace are equal. Therefore, the Krylov subspace $\mathcal{K}(\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}, \widetilde{\mathbf{d}})$ has dimension at most $m+1$. Hence, the proof follows. ■

Based on the property outlined in Theorem 3.5.5, when we use RPGSS-II preconditioner, Krylov subspace methods like PGMRES require maximum $m+1$ iterations to solve the system (3.1.1).

3.6. Discussion on the Selection of Parameters

It is worth noting that proposed GSS, RGSS-I, and RGSS-II preconditioners involve the parameters α, β, τ, and $\boldsymbol{\omega}$. As a general criterion for a preconditioner to perform efficiently, it should be as close as possible to the coefficient matrix of the system [26], we find the optimal choices for the proposed GSS preconditioner by minimizing $\|\mathcal{N}_{\text{GSS}}\|_F = \|\mathscr{P}_{\text{GSS}} - \mathfrak{B}\|_F$. First, we define the function φ by

$$\varphi(\alpha, \beta, \tau, \boldsymbol{\omega}) = \|\mathcal{N}_{\text{GSS}}\|_F^2 = \mathtt{tr}(\mathcal{N}_{\text{GSS}}^T \mathcal{N}_{\text{GSS}}) > 0.$$

After some easy calculations, we obtain

$$\begin{aligned}\varphi(\alpha, \beta, \tau, \boldsymbol{\omega}) =& \alpha^2\|P\|_F^2 + \beta^2\|Q\|_F^2 + \tau^2\|R\|_F^2 + (1-\boldsymbol{\omega})^2\|A\|_F^2 + (1-\boldsymbol{\omega})^2\|E\|_F^2 \\ &+ 2\alpha(\boldsymbol{\omega}-1)\mathtt{tr}(PA) + 2\beta(\boldsymbol{\omega}-1)\mathtt{tr}(QE) + 2(1-\boldsymbol{\omega})^2\|B\|_F^2 \\ &+ 2(1-\boldsymbol{\omega})^2\|C\|_F^2.\end{aligned}$$

Now we need to select the parameters α, β, τ and $\boldsymbol{\omega}$ such that $\varphi(\alpha, \beta, \tau, \boldsymbol{\omega})$ is very small. Since

$$\lim_{\alpha,\beta,\tau,\boldsymbol{\omega}\to 0_+} \varphi(\alpha, \beta, \tau, \boldsymbol{\omega}) = (1-\boldsymbol{\omega})^2\|A\|_F^2 + (1-\boldsymbol{\omega})^2\|E\|_F^2 + 2(1-\boldsymbol{\omega})^2\|B\|_F^2 + 2(1-\boldsymbol{\omega})^2\|C\|_F^2,$$

we can select $\boldsymbol{\omega} = 1$ and α, β, τ, $\boldsymbol{\omega} \to 0_+$ such that $\varphi(\alpha, \beta, \tau, \boldsymbol{\omega}) \to 0_+$, and consequently, $\mathcal{N}_{\text{GSS}} \to \mathbf{0}$.

In the sequel, the GSS preconditioner is equivalently rewritten as: $\mathscr{P}_{\text{GSS}} = \boldsymbol{\omega}\widetilde{\mathscr{P}}_{\text{GSS}}$, where

$$\widetilde{\mathscr{P}}_{\text{GSS}} = \begin{bmatrix} \frac{\alpha}{\boldsymbol{\omega}}P + A & \mathbf{0} & B^T \\ \mathbf{0} & \frac{\beta}{\boldsymbol{\omega}}Q + E & C \\ -B & -C^T & \frac{\tau}{\boldsymbol{\omega}}R \end{bmatrix}$$

can be regarded as a scaled preconditioner. A preconditioner is considered efficient if it closely approximates the coefficient matrix, and for given $\alpha, \beta, \tau > 0$, $\widetilde{\mathscr{P}}_{\text{GSS}} - \mathfrak{B} = \frac{1}{\boldsymbol{\omega}}\Theta \to \mathbf{0}$ as $\boldsymbol{\omega} \to \infty$. Similar studies apply to the RGSS-I and RGSS-II preconditioners as well.

The performance of the proposed preconditioners by varying the parameters is shown in the numerical experiment section.

3.7. Numerical Experiments

To demonstrate the effectiveness and robustness of the proposed preconditioners GSS, RGSS-I and RGSS-II over the existing ones within the Krylov subspace method to solve the DSPP, in this section, we perform a few numerical experiments. We compare our proposed GSS, RGSS-I and RGSS-II PGMRES methods (abbreviated as "GSS", "RGSS-I" and "'RGSS-II", respectively) with the GMRES method and PGMRES methods with block diagonal [33], block preconditioner [83], dimension splitting [23], relaxed dimension factorization [27], and shift-splitting [61] preconditioners (abbreviated as "BD", "BP" "DS", "RDF", and "SS", respectively). Moreover, we have also compared proposed methods with BD preconditioner in conjunction with minimum residual method (MIRES) (abbreviated as "BD-MINRES"). The numerical results are presented in terms of iteration counts (abbreviated as "IT") and elapsed CPU time in seconds (abbreviated as "CPU").

The initial guess vector is set to $\widehat{\boldsymbol{w}}_0 = \mathbf{0} \in \mathbb{R}^{n+p+m}$ for all iterative methods, and the method terminates if

$$\texttt{RES} := \frac{\|\mathfrak{B}\widehat{\boldsymbol{w}}_{k+1} - \widetilde{\mathbf{d}}\|_2}{\|\widetilde{\mathbf{d}}\|_2} < 10^{-6}$$

or if the maximum number of iterations exceeds 5000. All the linear subsystems involved in Algorithms 3.3.1, 3.5.1 and 3.5.2 are solved using the LU or Cholesky Factorization. Numerical experiments are conducted in MATLAB R2024a on a Windows 11 system, using an Intel(R) Core(TM) i7-10700 CPU at 2.90 GHz with 16 GB of memory.

Example 3.7.1. [33, 115] **The Poisson control problem:** We consider DSSP arising from the distributed control problem (1.1.1). The MATLAB code downloaded from [113], generates the linear system of the form (1.1.2) by using the following setup: parameters in "*set_def_setup.m*" are selected as:
def_setup.bc ='dirichlet', *def_setup.beta* $=1e-2$, *def_setup.ob* $=1$, *def_setup.type* ='dist2d' and *def_setup.pow* $= 5$, 6 and 7. For these selection of *def_setup.pow*, size of the coefficient matrix $\mathfrak{B}$ is 2883, 11907 and 48378, respectively.
Parameter selection: For the DS preconditioner, we choose the parameter α (denoted by α_{DS}) as follow [23]:

$$\alpha_{\text{DS}} = \frac{\sqrt{\texttt{tr}(A^TA) + 2\texttt{tr}(BB^T)} + \sqrt{\texttt{tr}(E^TE) + 2\texttt{tr}(C^TC)}}{2(n+m+p)}.$$

Table 3.7.1: Experimental results of GMRES, BD, BD-MINRES, BP, DS, RDF, SS, GSS, RGSS-I and RGSS-II PGMRES methods for Example 3.7.1 when $\boldsymbol{\nu} = 0.1$

Process		*def_setup.pow* = 5	*def_setup.pow* = 6	*def_setup.pow* = 7
	size($\mathfrak{B}$)	2883	11907	48378
GMRES	IT	579	2149	−−
	CPU	2.9618	367.9274	−−
BD	IT	10	10	10
	CPU	2.1418	40.6138	1512.0707
BD-MINRES	IT	9	8	8
	CPU	1.5042	28.7551	1122.2980
BP	IT	3	3	3
	CPU	0.7222	17.1404	555.3527
DS	IT	31	40	51
	CPU	2.8514	54.3105	1706.9543
RDF	IT	6	6	8
	CPU	0.8733	8.1585	302.9091
SS	IT	16	22	46
	CPU	2.12460	45.9988	3985.6141
GSS	IT	2	2	2
$\boldsymbol{\omega}_{exp} = 30$	CPU	0.6909	7.3233	226.8625
RGSS-I	IT	2	2	2
$\boldsymbol{\omega}_{exp} = 25$	CPU	0.7796	7.72871	249.7657
RGSS-II	IT	2	2	2
$\boldsymbol{\omega}_{exp} = 30$	CPU	0.7946	6.8447	238.5576

−− signifies that the method does not converge within 5000 IT.

For the RDF preconditioner, the parameter α is selected from the interval $(0, 1)$ with a step size of 0.01. The optimal performance, in terms of minimal CPU times, is achieved with smaller values of α as found in [27]. For the SS preconditioner, we take $\alpha = 0.01$. For the GSS, RGSS-I, and RGSS-II preconditioners parameters are selected as follows: $\alpha = \beta = 0.01$, $\tau = 0.001$, $P = A$, $Q = CC^T$, and $R = I$. The optimal parameter $\boldsymbol{\omega}$ (denoted by $\boldsymbol{\omega}_{exp}$) is determined experimentally within the interval $[2, 30]$ with step size one, which yields minimal CPU times.

Numerical results: The numerical results for GMRES and various PGMRES methods with $\boldsymbol{\beta} = 0.1$ and 0.001 are presented in Tables 3.7.1 and 3.7.2. We observe that the

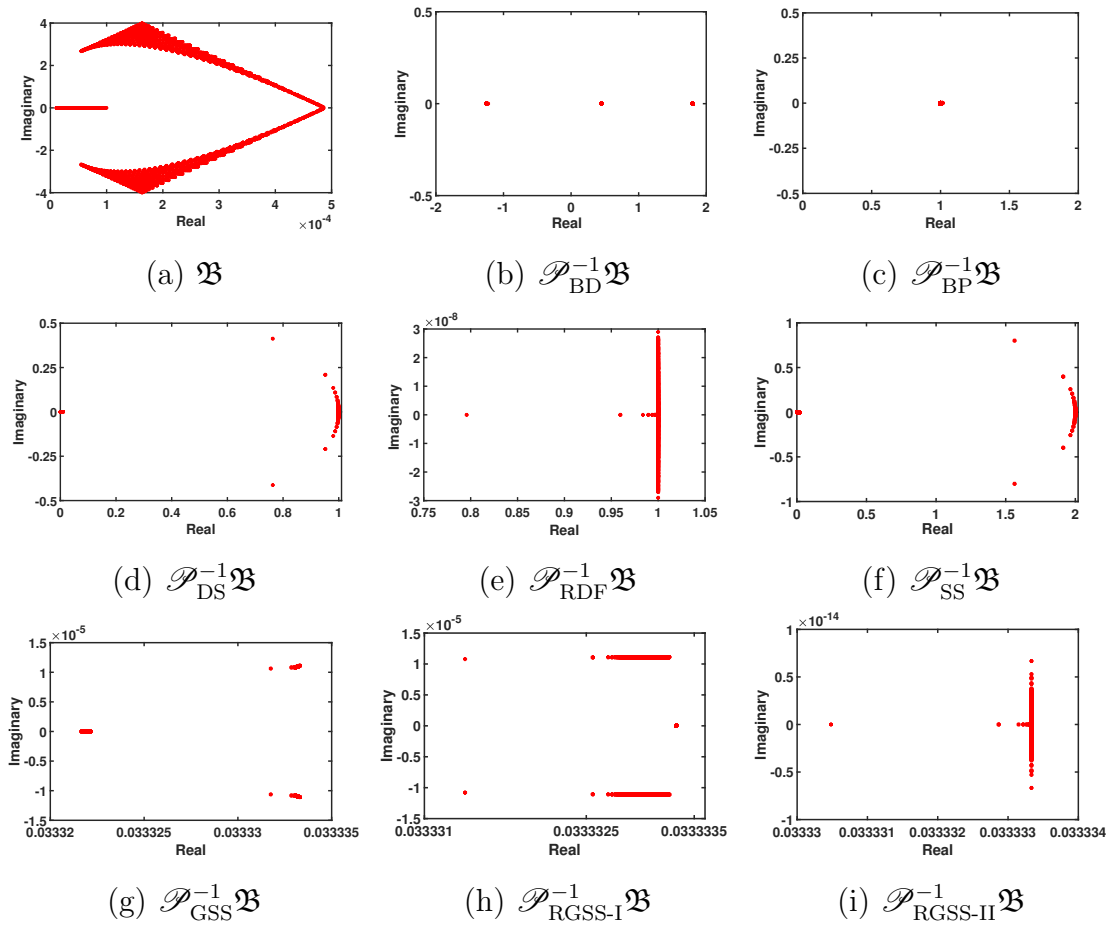
Table 3.7.2: Experimental results of GMRES, BD, BD-MINRES, BP, DS, RDF, SS, GSS, RGSS-I and RGSS-II PGMRES methods for Example 3.7.1 when $\boldsymbol{\nu} = 0.001$

Process		*def_setup.pow* = 5	*def_setup.pow* = 6	*def_setup.pow* = 7
	size($\mathfrak{B}$)	2883	11907	48378
GMRES	IT	919	2254	−−
	CPU	5.8702	409.5427	−−
BD	IT	19	19	19
	CPU	4.3498	84.9364	3034.4227
BD-MINRES	IT	9	8	8
	CPU	1.5042	28.7551	1122.2980
BP	IT	3	3	3
	CPU	0.7222	17.1404	555.3527
DS	IT	29	39	46
	CPU	2.6072	47.8576	1603.9128
RDF	IT	14	11	8
	CPU	1.4320	14.3846	324.8007
SS	IT	27	38	68
	CPU	3.3960	79.7055	5481.9674
GSS	IT	2	2	2
$\boldsymbol{\omega}_{exp} = 30$	CPU	0.7159	7.4041	242.8123
RGSS-I	IT	2	2	2
$\boldsymbol{\omega}_{exp} = 30$	CPU	0.78675	7.6515	221.0918
RGSS-II	IT	2	2	2
$\boldsymbol{\omega}_{exp} = 26$	CPU	0.6828	6.9229	247.4777

−− signifies that the method does not converge within 5000 IT.

GMRES method exhibits a significantly slower convergence rate compared to all other PGMRES methods, even does not converge within 5000 iterations when *def_setup.pow* = 7. On the other hand, we observe that our proposed preconditioners outperform all the compared preconditioners in terms of both IT and CPU times. For the DS and SS preconditioners, the IT increases as the size of $\mathfrak{B}$ grows. Whereas the proposed GSS, RGSS-I and RGSS-II preconditioners maintain a consistent IT regardless of matrix size.

Eigenvalue distributions: In order to better illustrate the superiority of the proposed GSS, RGSS-I and RGSS-II preconditioners, the eigenvalue distribution of $\mathfrak{B}$ and preconditioned matrices $\mathscr{P}_{\text{BD}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{DS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RDF}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{SS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ are



(a) $\mathfrak{B}$ (b) $\mathscr{P}_{\text{BD}}^{-1}\mathfrak{B}$ (c) $\mathscr{P}_{\text{BP}}^{-1}\mathfrak{B}$

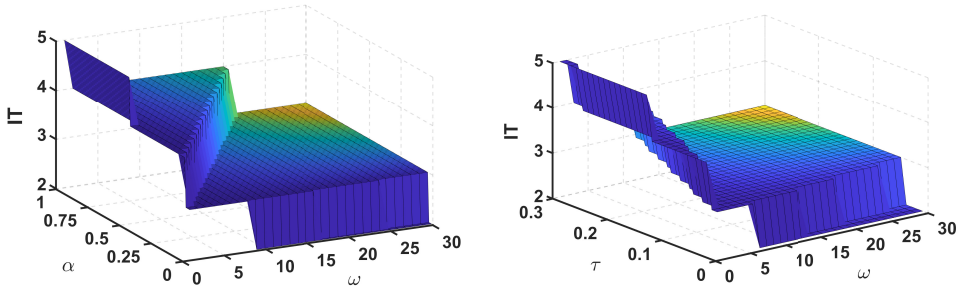
(d) $\mathscr{P}_{\text{DS}}^{-1}\mathfrak{B}$ (e) $\mathscr{P}_{\text{RDF}}^{-1}\mathfrak{B}$ (f) $\mathscr{P}_{\text{SS}}^{-1}\mathfrak{B}$

(g) $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$ (h) $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ (i) $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$

Figure 3.7.1: Eigenvalue distributions of $\mathfrak{B}$, $\mathscr{P}_{\text{BD}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{DS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{SS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ for def_set.pow $= 5$ with $\boldsymbol{\beta} = 0.1$ for Example 3.7.1.

displayed in Figure 3.7.1. From Figure 3.7.1, we observe that the eigenvalues of the preconditioned matrices $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$, and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ have clustered better than the coefficient matrix $\mathfrak{B}$, and preconditioned matrices $\mathscr{P}_{\text{BD}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{DS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RDF}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{SS}}^{-1}\mathfrak{B}$. This indicates enhanced computational efficiency, highlighting the superiority of the GSS, RGSS-I, and RGSS-II preconditioners over existing methods.

Influence of the parameters $\alpha, \beta, \tau, \boldsymbol{\omega}$: To demonstrate the influence of the parameter on the performance of the proposed preconditioners, we present graphs of IT counts versus parameters for the GSS and RGSS-II preconditioners in Figure 3.7.2. For the GSS preconditioner, we vary $\alpha = \beta$ from 0.01 to 1 with a step size of 0.01 and $\boldsymbol{\omega}$ from 1 to 30 with a step size of one. For the RGSS-II preconditioner, we vary τ from 0.01 to 0.3 with



(a) By varying $\alpha = \beta$ within the interval $[0.01, 1]$ and ω within the range $[1, 30]$ for the GSS preconditioner

(b) By varying τ within the interval $[0.01, 0.3]$ and $\boldsymbol{\omega}$ within the interval $[1, 30]$ for the RGSS-II preconditioner

Figure 3.7.2: Convergence curves of the GSS and RGSS-II RGMRES methods varying the parameters α, β, τ, $\boldsymbol{\omega}$ for Example 3.7.1 with $\boldsymbol{\beta} = 0.1$.

a step size of 0.01 and $\boldsymbol{\omega}$ from 1 to 30 with a step size of one. We can draw the following observation from Figure 3.7.2:

- For both preconditioners, IT exhibits minimal sensitivity to variations in the parameters.
- Although, for small values of $\boldsymbol{\omega}$, IT increases when α increase for GSS preconditioner and τ increases for RGSS-II preconditioner. Nonetheless, as $\boldsymbol{\omega}$ increases, IT decreases, even as the magnitude of α and τ continue to grow.

Therefore, the proposed preconditioners achieve high efficiency when α, β and τ are kept small and $\boldsymbol{\omega}$ is large.

Condition number (CN) analysis: To evaluate the robustness of the proposed GSS,

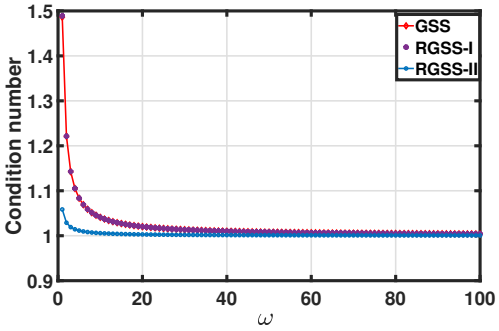


Figure 3.7.3: Relationship between CNs of the preconditioned matrices $\mathscr{P}_{\text{GSS}}^{-1}\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ varying the parameter $\boldsymbol{\omega}$ in $[1, 100]$ with $\boldsymbol{\beta} = 0.1$ for Example 3.7.1.

RGSS-I and RGSS-II preconditioners, we assess the CNs of the preconditioned matrices

$\mathscr{P}_{\text{GSS}}^{-1}\,\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$. For any nonsingular matrix A, its CN is defined as $\kappa(A) := \|A^{-1}\|_2\|A\|_2$. The CN of $\mathfrak{B}$ is $3.6396e+05$, which is comparatively large, making the system (3.1.1) ill-conditioned to solve. In Figure 3.7.3, we depict the effect of the parameter $\boldsymbol{\omega}$ ranges from 1 to 100 with a step size one on the preconditioned matrices $\mathscr{P}_{\text{GSS}}^{-1}\,\mathfrak{B}$, $\mathscr{P}_{\text{RGSS-I}}^{-1}\mathfrak{B}$ and $\mathscr{P}_{\text{RGSS-II}}^{-1}\mathfrak{B}$ for the case *def_setup.pow* $= 5$. We observe that for all values of $\boldsymbol{\omega}$, the CNs of the preconditioned matrices remain within the range $[1, 1.5]$. This indicates that the preconditioned systems are well-conditioned, demonstrating that the GSS, RGSS-I, and RGSS-II preconditioners are robust and effective.

3.8. Summary

This chapter proposed three preconditioners, termed GSS, RGSS-I, and RGSS-II for solving DSPPs arising from various applications. We provide a convergence analysis for the GSS iterative method, demonstrating that the method converges for any initial guess vector when the parameter $\boldsymbol{\omega} \geq 1/2$. Moreover, spectral bounds for the preconditioned matrices are derived. Additionally, we have shown that the RGSS-II preconditioner requires at most $m+1$ iterations to solve the DSPP. Numerical experiments for the DSPP arising from the PDE-constrained optimization problem are performed, which demonstrate that the proposed preconditioners are efficient and outperform the existing state-of-the-art preconditioners.

CHAPTER 4

Sparsity Preserving Structured Backward Errors for Saddle Point Problems * †‡

In this chapter, we investigate the structured backward errors (BEs) of GSPPs and DSPPs when the perturbation on the block matrices exploits the sparsity pattern as well as symmetric, Hermitian, circulant, Toeplitz, and symmetric-Toeplitz structures. Furthermore, we construct minimal perturbation matrices that preserve the sparsity pattern and the aforementioned structures. The developed frameworks are applied to compute BEs for the weighted regularized least squares (WRLS) problem. Finally, numerical experiments are performed to validate our findings, showcasing the utility of the obtained structured BEs in assessing the strong backward stability of numerical algorithms.

4.1. Structured Backward Errors for Generalized Saddle Point Problems

In this section, we consider the GSPP of the following form:

$$\mathcal{M}\boldsymbol{v} \triangleq \begin{bmatrix} A & B^T \\ B & D \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \triangleq \mathbf{b}, \tag{4.1.1}$$

where $A \in \mathbb{C}^{n\times n}, B \in \mathbb{C}^{m\times n}, D \in \mathbb{C}^{m\times m}, f \in \mathbb{C}^n$, and $g \in \mathbb{C}^m$. In general, the block matrices A, B and D are sparse [60]. Further, the block matrices in (4.1.1) can have symmetric, Toeplitz, symmetric-Toeplitz, or circulant structures. For instance, in the WRLS problem arising from image reconstruction [62] and image restoration with colored noise [82]. Also, GSPP involving circulant or Toeplitz block matrices often arise during the discretization of elasticity problems using finite difference scheme [32].

---

* S. S. Ahmad and **P. Khatun**, " Structured backward errors for special classes of saddle point problems with applications." *Linear Algebra and its Applications,* 13: 90-112, 2025.

† S. S. Ahmad and **P. Khatun**, "Structured backward error analysis for double saddle point problems." *Under Review.*

‡ S. S. Ahmad and **P. Khatun**, "Structured backward errors of sparse generalized saddle point problems with Hermitian block matrices." *Revision submitted in Electronic Transaction on Numerical Analysis.*

In recent times, a number of numerical algorithms have been developed to find the efficient solution of the GSPP (4.1.1) with circulant, Toeplitz, or symmetric-Toeplitz block matrices; see [17, 32, 157, 163]. However, the computed solution may still contain some errors and can potentially lead to insignificant results. Therefore, it is crucial to assess how closely the computed solution approximates the solution of the original problem. This prompts a natural inquiry: can an approximate solution obtained using a numerical algorithm serve as the exact solution to a nearly perturbed problem? The concept of BE is used to determine the minimal distance between the perturbed problem and the original problem. However, the perturbed coefficient matrix does not necessarily retain the special block structure of (4.1.1). This raises an interesting question: Does preserving the special structure of the block matrices in the perturbation matrix lead to a smaller BE? That is, are the numerical algorithms for solving the GSPP strongly backward stable or not? So far, significant research has been done in this direction, and structured BE has been extensively considered in [44, 97, 102, 129, 146, 162], where the perturbation matrix $\Delta\mathcal{M}$ preserve the only block structure of $\mathcal{M}$ and perturbation on block matrices A or D preserve the symmetric structures. However, it is noteworthy to mention a few drawbacks of the aforementioned studies:

- The block matrices of $\mathcal{M}$ in (4.1.1) are often sparse in many applications, making it essential to maintain their sparsity pattern in the perturbation matrices. The existing studies do not consider and preserve the sparsity pattern of the coefficient matrix $\mathcal{M}$. By preserving the sparsity pattern of the original matrices, structured BEs have been studied in the literature; see, for example, [2, 3, 159].
- Existing techniques are not applicable when the block matrices A, B and D in GSPP (4.1.1) have circulant, Toeplitz, or symmetric-Toeplitz structures.
- Moreover, the research available in the literature for structured BE analysis for (4.1.1) does not provide the explicit formulae for the minimal perturbation matrices for which structured BE is attained and preserves the inherent matrix structure.

This section addresses the aforementioned challenges by investigating structured BEs for the GSPP (4.1.1) by preserving both the inherent block structure and sparsity in the perturbation matrices under three scenarios. First, we consider the case where $n = m$ and the block matrices A, B, and D are circulant. Second, we consider A, B, and D are Toeplitz matrices. Third, we analyze the case when $n = m$, B is symmetric-Toeplitz, and A, $D \in \mathbb{C}^{n\times n}$. The following are the main contributions of this section:

- We investigate the structured BEs for the GSPP (4.1.1) when the block matrices A, B and D possesses circulant, Toeplitz, and symmetric-Toeplitz structure with or without preserving the sparsity pattern.
- We develop frameworks that give the minimal perturbation matrices that retain the circulant, Toeplitz, or symmetric-Toeplitz structures, as well as the sparsity pattern of the original matrices.
- We provide an application of our obtained results in finding the structured BEs for the WRLS problem when the coefficient matrix exhibits Toeplitz or symmetric-Toeplitz structure.
- Lastly, numerical experiments are performed to test the backward stability and strong backward stability of numerical algorithms to solve GSPP (4.1.1).

This section is organized as follows. In Subsection 4.1.1, we discuss preliminary definitions and results. In Subsections 4.1.2-4.1.4, structured BEs for circulant, Toeplitz, and symmetric-Toeplitz matrices are derived, respectively. Moreover, in Subsection 4.1.5, we discuss the unstructured BE for the GSPP (4.1.1) by only preserving the sparsity pattern. In Subsection 4.1.6, we provide an application of our developed theories in the WRLS problem.

4.1.1. Preliminaries

Let $w := [w_1, w_2, w_3, w_4, w_5]^T$, where w_i are nonnegative real numbers for $i = 1, 2, \ldots, 5$, with the convention that $w_i^{-1} = 0$, whenever $w_i = 0$. For any w, we define

$$\left|\!\left|\!\left| \begin{bmatrix} \mathcal{M} & \mathbf{b} \end{bmatrix} \right|\!\right|\!\right|_{w,F} = \left\| \begin{bmatrix} w_1\|A\|_F, & w_2\|B\|_F, & w_3\|D\|_F, & w_4\|f\|_2, & w_5\|g\|_2 \end{bmatrix} \right\|_2.$$

Note that $w_i = 0$ implies that the corresponding block matrix has no perturbation. Next, we recall the definitions of circulant, Toeplitz, and symmetric-Toeplitz matrices.

Definition 4.1.1. *A matrix $C \in \mathbb{C}^{n\times n}$ is called a circulant matrix if for any vector $c = [c_1, c_2, \ldots, c_n]^T \in \mathbb{C}^n$, it has the following form:*

$$\mathtt{Cr}(c) := C = \begin{bmatrix} c_1 & c_n & c_{n-1} & \cdots & c_2 \\ c_2 & c_1 & c_n & \cdots & c_3 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_{n-1} & c_{n-2} & \ddots & \ddots & c_n \\ c_n & c_{n-1} & \cdots & c_2 & c_1 \end{bmatrix}. \tag{4.1.2}$$

We denote the generator vector for the circulant matrix C as

$$\text{vec}_{\mathcal{C}}(C) := [c_1, c_2, \ldots, c_n]^T \in \mathbb{C}^n.$$

Definition 4.1.2. *A matrix $T = [t_{ij}] \in \mathbb{C}^{m\times n}$ is called a Toeplitz matrix if for any vector*

$$\text{vec}_{\mathcal{T}}(T) := [t_{-m+1}, t_{-m+2}, \ldots, t_{-1}, t_0, t_1, \ldots, t_{n-1}]^T \in \mathbb{C}^{n+m-1},$$

we have $t_{ij} = t_{j-i}$, for all $1 \le i \le m$ and $1 \le j \le n$.

We denote $\text{vec}_{\mathcal{T}}(T)$ as the generator vector for the Toeplitz matrix T. Also, for any vector $t \in \mathbb{C}^{m+n-1}$ corresponding generated Toeplitz matrix is denoted by $\mathcal{T}(t)$.

Remark 4.1.1. *The Toeplitz matrix T is known as a symmetric-Toeplitz matrix when $n = m$ and $t_{-m+1} = t_{n-1}, \ldots, t_{-1} = t_1$. In this context, we employ the notation*

$$\text{vec}_{\mathcal{ST}}(T) := [t_0, \ldots, t_{n-1}]^T \in \mathbb{C}^n$$

to denote its generator vector. Also, for any vector $t \in \mathbb{C}^n$, the corresponding symmetric-Toeplitz matrix is symbolized as $\mathcal{ST}(t)$.

Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the GSPP (4.1.1). Using the formula (1.3.4), the unstructured BE for the GSPP (4.3.1), denoted by $\boldsymbol{\eta}(\widetilde{\boldsymbol{v}})$, is expressed as:

$$\boldsymbol{\eta}(\widetilde{\boldsymbol{v}}) = \frac{\|\mathbf{b} - \mathcal{M}\widetilde{\boldsymbol{v}}\|_2}{\sqrt{\|\mathcal{M}\|_F^2\|\widetilde{\boldsymbol{v}}\|_2^2 + \|\mathbf{b}\|_2^2}}. \tag{4.1.3}$$

Throughout the section, we assume that the coefficient matrix $\mathcal{M}$ in (4.1.1) is non-singular. If the block matrices A, B, and D have circulant (or Toeplitz) structure, we identify (4.1.1) as circulant (or Toeplitz) structured GSPP. Moreover, we call (4.1.1) as symmetric-Toeplitz structured GSPP when $B \in \mathcal{ST}_n$ and $A, D \in \mathbb{C}^{n\times n}$.

We denote

$$\Delta\mathcal{M} := \begin{bmatrix} \Delta A & \Delta B^T \\ \Delta B & \Delta D \end{bmatrix} \text{ and } \Delta\mathbf{b} := \begin{bmatrix} \Delta f \\ \Delta g \end{bmatrix}.$$

Next, we define normwise structured BE for the GSPP (4.1.1).

Definition 4.1.3. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the GSPP (4.1.1). Then, the normwise structured BEs are defined as follows:*

$$\boldsymbol{\eta}^{\mathcal{S}_i}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \min_{\begin{pmatrix} \Delta A, \Delta B, \\ \Delta D, \Delta f, \Delta g \end{pmatrix} \in \mathcal{S}_i} \left|\!\left|\!\left| \begin{bmatrix} \Delta\mathcal{M} & \Delta\mathbf{b} \end{bmatrix} \right|\!\right|\!\right|_{w,F}, \quad \text{for } i = 1, 2, 3,$$

where

$$\mathcal{S}_1 = \left\{ \begin{pmatrix} \Delta A, \Delta B, \\ \Delta D, \Delta f, \Delta g \end{pmatrix} \Big| \begin{bmatrix} A+\Delta A & (B+\Delta B)^T \\ B+\Delta B & D+\Delta D \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{u}} \\ \widetilde{\boldsymbol{p}} \end{bmatrix} = \begin{bmatrix} f+\Delta f \\ g+\Delta g \end{bmatrix}, \right.$$
$$\left. \Delta A, \Delta B, \Delta C \in \mathcal{C}_n,\ \Delta f, \Delta g \in \mathbb{C}^n \right\}, \tag{4.1.4}$$

$$\mathcal{S}_2 = \left\{ \begin{pmatrix} \Delta A, \Delta B, \\ \Delta D, \Delta f, \Delta g \end{pmatrix} \Big| \begin{bmatrix} A+\Delta A & (B+\Delta B)^T \\ B+\Delta B & D+\Delta D \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{u}} \\ \widetilde{\boldsymbol{p}} \end{bmatrix} = \begin{bmatrix} f+\Delta f \\ g+\Delta g \end{bmatrix}, \right.$$
$$\left. \Delta A \in \mathcal{T}_{n\times n},\ \Delta B \in \mathcal{T}_{m\times n},\ \Delta D \in \mathcal{T}_{m\times m},\ \Delta f \in \mathbb{C}^n, \Delta g \in \mathbb{C}^m \right\}, \tag{4.1.5}$$

and

$$\mathcal{S}_3 = \left\{ \begin{pmatrix} \Delta A, \Delta B, \\ \Delta D, \Delta f, \Delta g \end{pmatrix} \Big| \begin{bmatrix} A+\Delta A & (B+\Delta B)^T \\ B+\Delta B & D+\Delta D \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{u}} \\ \widetilde{\boldsymbol{p}} \end{bmatrix} = \begin{bmatrix} f+\Delta f \\ g+\Delta g \end{bmatrix}, \right.$$
$$\left. \Delta B \in \mathcal{ST}_n,\ \Delta A, \Delta D \in \mathbb{C}^{n\times n},\ \Delta f, \Delta g \in \mathbb{C}^n \right\}. \tag{4.1.6}$$

In the following, we state the problem of finding structure-preserving minimal perturbation matrices for which the structured BE is attained.

Problem 4.1.2. Find out the minimal perturbation matrices $\widehat{\Delta A}$, $\widehat{\Delta B}, \widehat{\Delta D}$, $\widehat{\Delta f}$ and $\widehat{\Delta g}$ in $\mathcal{S}_i$ such that

$$\boldsymbol{\eta}^{\mathcal{S}_i}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left|\!\left|\!\left| \begin{bmatrix} \widehat{\Delta\mathcal{M}} & \widehat{\Delta d} \end{bmatrix} \right|\!\right|\!\right|_{w,F}, \text{ for } i = 1, 2, 3,$$

where $\widehat{\Delta\mathcal{M}} := \begin{bmatrix} \widehat{\Delta A} & \widehat{\Delta B}^T \\ \widehat{\Delta B} & \widehat{\Delta D} \end{bmatrix}$ and $\widehat{\Delta \mathbf{b}} := \begin{bmatrix} \widehat{\Delta f} \\ \widehat{\Delta g} \end{bmatrix}$.

Remark 4.1.3. *Our main focus is studying perturbations with the same sparsity pattern as the original matrices. To achieve this, we replace the perturbation matrices $\Delta A, \Delta B$ and ΔD by $\Delta A \odot \Theta_A, \Delta B \odot \Theta_B$ and $\Delta D \odot \Theta_D$, respectively, where the sparsity pattern of a matrix $A \in \mathbb{C}^{m\times n}$ is defined as $\Theta_A := \mathtt{sgn}(A) = [\mathtt{sgn}(a_{ij})]$. In this context, we denote the structured BEs by $\boldsymbol{\eta}^{\mathcal{S}_i}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$, $i = 1, 2, 3$. Further, the minimal perturbation matrices are denoted by $\widehat{\Delta A}_{\mathtt{sps}}$, $\widehat{\Delta B}_{\mathtt{sps}}$, $\widehat{\Delta D}_{\mathtt{sps}}$, $\widehat{\Delta f}_{\mathtt{sps}}$, and $\widehat{\Delta g}_{\mathtt{sps}}$. Note that, if $M \in \mathcal{C}_n$ (or $\mathcal{T}_{m\times n}$, or $\mathcal{ST}_n$), then $\Theta_M \in \mathcal{C}_n$ (or $\mathcal{T}_{m\times n}$, or $\mathcal{ST}_n$).*

4.1.2. Structured BEs for Circulant Structured GSPPs

In this subsection, we consider $n = m$ and derive explicit formulae for the structured BEs $\boldsymbol{\eta}^{\mathcal{S}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ and $\boldsymbol{\eta}^{\mathcal{S}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$, by preserving the circulant structure to the perturbation

matrices. Moreover, we provide minimal perturbation matrices to the Problem (4.1.2). In order to obtain structured BEs formulae, we derive the following lemma.

Lemma 4.1.4. *Let $A, B, M \in \mathcal{C}_n$ with generator vectors $\mathrm{vec}_{\mathcal{C}}(A) = [a_1, \ldots, a_n]^T \in \mathbb{C}^n$, $\mathrm{vec}_{\mathcal{C}}(B) = [b_1, \ldots, b_n]^T \in \mathbb{C}^n$, and $\mathrm{vec}_{\mathcal{C}}(M) = [m_1, \ldots, m_n]^T \in \mathbb{C}^n$, respectively. Suppose $x = [x_1, \ldots, x_n]^T \in \mathbb{C}^n$ and $y = [y_1, \ldots, y_n]^T \in \mathbb{C}^n$. Then*

$$(A \odot \Theta_M)x = \mathtt{Cr}\,(x)\,\mathfrak{D}_{c(M)}\mathrm{vec}_{\mathcal{C}}(A \odot \Theta_M) \text{ and}$$
$$(B \odot \Theta_M)^T y = \mathcal{H}_y \mathfrak{D}_{c(M)}\mathrm{vec}_{\mathcal{C}}(B \odot \Theta_M),$$

where $c(M) := \mathrm{vec}_{\mathcal{C}}(\Theta_M)$ and $\mathcal{H}_y \in \mathbb{C}^{n\times n}$ has the following form:

$$\mathcal{H}_y = \begin{bmatrix} y_1 & y_2 & \cdots & y_{n-1} & y_n \\ y_2 & y_3 & \cdots & y_n & y_1 \\ \vdots & \iddots & \iddots & \iddots & \vdots \\ y_{n-1} & \iddots & \iddots & y_{n-3} & y_{n-2} \\ y_n & y_1 & \cdots & y_{n-2} & y_{n-1} \end{bmatrix}. \tag{4.1.7}$$

Proof. Since ij^{th} entry of $A \odot \Theta_M$ is $(A \odot \Theta_M)_{ij} = a_{ij}\,\mathtt{sgn}(m_{ij})$, we get $A \odot \Theta_M \in \mathcal{C}_n$, and

$$\mathrm{vec}_{\mathcal{C}}(A \odot \Theta_M) = \begin{bmatrix} a_1\mathtt{sgn}(m_1) \\ \vdots \\ a_n\mathtt{sgn}(m_n) \end{bmatrix}.$$

Now, expanding $(A \odot \Theta_M)x$, we get the following:

$$(A \odot \Theta_M)x = \begin{bmatrix} a_1\,\mathtt{sgn}(m_1)x_1 + a_n\,\mathtt{sgn}(m_n)x_2 + \cdots + a_2\,\mathtt{sgn}(m_2)x_n \\ a_2\,\mathtt{sgn}(m_2)x_1 + a_1\,\mathtt{sgn}(m_1)x_2 + \cdots + a_3\,\mathtt{sgn}(m_3)x_n \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ a_n\,\mathtt{sgn}(m_n)x_1 + a_{n-1}\,\mathtt{sgn}(m_{n-1})x_2 + \cdots + a_1\,\mathtt{sgn}(m_1)x_n \end{bmatrix}.$$

Since $(\mathtt{sgn}(m_i))^2 = \mathtt{sgn}(m_i)$, rearrangement of the above gives

$$(A \odot \Theta_M)x = \begin{bmatrix} x_1\,\mathtt{sgn}(m_1) & x_n\,\mathtt{sgn}(m_2) & x_{n-1}\,\mathtt{sgn}(m_3) & \cdots & x_2\,\mathtt{sgn}(m_n) \\ x_2\,\mathtt{sgn}(m_1) & x_1\,\mathtt{sgn}(m_2) & x_n\,\mathtt{sgn}(m_3) & \cdots & x_3\,\mathtt{sgn}(m_n) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ x_{n-1}\,\mathtt{sgn}(m_1) & x_{n-2}\,\mathtt{sgn}(m_2) & \cdots & \ddots & x_n\,\mathtt{sgn}(m_n) \\ x_n\,\mathtt{sgn}(m_1) & x_{n-1}\,\mathtt{sgn}(m_2) & \cdots & x_2\,\mathtt{sgn}(m_{n-1}) & x_1\,\mathtt{sgn}(m_n) \end{bmatrix} \begin{bmatrix} a_1\mathtt{sgn}(m_1) \\ a_2\mathtt{sgn}(m_2) \\ \vdots \\ \vdots \\ a_n\mathtt{sgn}(m_n) \end{bmatrix}.$$

Hence, the above can be expressed as

$$(A \odot \Theta_M)x = \mathtt{Cr}(x)\mathfrak{D}_{c(M)}\mathrm{vec}_{\mathcal{C}}(A \odot \Theta_M).$$

Similarly, expanding $(B \odot \Theta_M)^T y$, we can obtain

$$(B \odot \Theta_M)^T y = \mathcal{H}_y \mathfrak{D}_{c(M)} \mathrm{vec}_{\mathcal{C}}(B \odot \Theta_M),$$

where $\mathcal{H}_y$ is given by (4.1.7). ■

For a better understanding of Lemma 4.1.4, we consider the following example.

Example 4.1.1. Consider

$$A = \begin{bmatrix} 5 & 7 & 3 \\ 3 & 5 & 7 \\ 7 & 3 & 5 \end{bmatrix} \in \mathcal{C}_3 \text{ and } M = \begin{bmatrix} 3 & 0 & 9 \\ 9 & 3 & 0 \\ 0 & 9 & 3 \end{bmatrix} \in \mathcal{C}_3.$$

Then, $\Theta_M = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$, $\mathrm{vec}_{\mathcal{C}}(\Theta_M) = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$, and we get

$$(A \odot \Theta_M)x = \left(\begin{bmatrix} 5 & 7 & 3 \\ 3 & 5 & 7 \\ 7 & 3 & 5 \end{bmatrix} \odot \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{4.1.8}$$

Then (4.1.8) can be rearranged in the following form:

$$(A \odot \Theta_M)x = \begin{bmatrix} x_1 & x_3 & x_2 \\ x_2 & x_1 & x_3 \\ x_3 & x_2 & x_1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix}. \tag{4.1.9}$$

The above equation can be written as: $(A \odot \Theta_M)x = \mathtt{Cr}(x)\mathfrak{D}_{c(M)}\mathrm{vec}_{\mathcal{C}}(A \odot \Theta_M)$.

Next, we present the main result of this section concerning the structured BE for circulant structured GSPP while preserving the sparsity pattern. Before that, we introduce the following notation:

$$\mathfrak{D}_{c(A)} = \mathrm{diag}(\mathrm{vec}_{\mathcal{C}}(\Theta_A)), \quad \mathfrak{D}_{c(B)} = \mathrm{diag}(\mathrm{vec}_{\mathcal{C}}(\Theta_B)), \tag{4.1.10}$$

$$\mathfrak{D}_{c(D)} = \mathrm{diag}(\mathrm{vec}_{\mathcal{C}}(\Theta_D)), \quad \text{and} \quad \mathfrak{D}_{\boldsymbol{a}} = \mathrm{diag}(\boldsymbol{a}), \tag{4.1.11}$$

where $\boldsymbol{a} = [\sqrt{n}, \ldots, \sqrt{n}]^T \in \mathbb{R}^n$.

Theorem 4.1.5. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be the approximate solution of the circulant structured GSPP (4.1.1), i.e., $A, B, D \in \mathcal{C}_n$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}^{\mathcal{S}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \mathcal{X}^H_{\mathtt{Cr}} (\mathcal{X}_{\mathtt{Cr}} \mathcal{X}^H_{\mathtt{Cr}})^{-1} r_{\mathtt{b}} \right\|_2, \tag{4.1.12}$$

where $\mathcal{X}_{\mathtt{Cr}} \in \mathbb{C}^{2n\times 5n}$ *is given by*

$$\mathcal{X}_{\mathtt{Cr}} = \begin{bmatrix} \frac{1}{w_1}\mathtt{Cr}(\widetilde{\boldsymbol{u}})\mathfrak{D}_{c(A)}\mathfrak{D}_{\boldsymbol{a}}^{-1} & \frac{1}{w_2}\mathcal{H}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{c(B)}\mathfrak{D}_{\boldsymbol{a}}^{-1} & \mathbf{0} & -\frac{1}{w_4}I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2}\mathtt{Cr}(\widetilde{\boldsymbol{u}})\mathfrak{D}_{c(B)}\mathfrak{D}_{\boldsymbol{a}}^{-1} & \frac{1}{w_3}\mathtt{Cr}(\widetilde{\boldsymbol{p}})\mathfrak{D}_{c(D)}\mathfrak{D}_{\boldsymbol{a}}^{-1} & \mathbf{0} & -\frac{1}{w_5}I_n \end{bmatrix},$$

$r_{\mathbf{b}} = [r_f^T,\ r_g^T]^T$, $r_f = f - A\widetilde{\boldsymbol{u}} - B^T\widetilde{\boldsymbol{p}}$, *and* $r_g = g - B\widetilde{\boldsymbol{u}} - D\widetilde{\boldsymbol{p}}$.

Furthermore, the minimum norm perturbations to the Problem 4.1.2 are given by

$$\widehat{\Delta A}_{\mathtt{sps}} = \mathtt{Cr}\left(\frac{1}{w_1}\mathfrak{D}_{\boldsymbol{a}}^{-1}\begin{bmatrix} I_n & \mathbf{0}_{n\times 4n}\end{bmatrix}\mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1}r_{\mathbf{b}}\right), \tag{4.1.13}$$

$$\widehat{\Delta B}_{\mathtt{sps}} = \mathtt{Cr}\left(\frac{1}{w_2}\mathfrak{D}_{\boldsymbol{a}}^{-1}\begin{bmatrix} \mathbf{0}_{n\times n} & I_n & \mathbf{0}_{n\times 3n}\end{bmatrix}\mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1}r_{\mathbf{b}}\right), \tag{4.1.14}$$

$$\widehat{\Delta D}_{\mathtt{sps}} = \mathtt{Cr}\left(\frac{1}{w_3}\mathfrak{D}_{\boldsymbol{a}}^{-1}\begin{bmatrix} \mathbf{0}_{n\times 2n} & I_n & \mathbf{0}_{n\times 2n}\end{bmatrix}\mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1}r_{\mathbf{b}}\right), \tag{4.1.15}$$

$$\widehat{\Delta f}_{\mathtt{sps}} = \frac{1}{w_4}\begin{bmatrix} \mathbf{0}_{n\times 3n} & I_n & \mathbf{0}_{n\times n}\end{bmatrix}\mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1}r_{\mathbf{b}}, \quad \textit{and} \tag{4.1.16}$$

$$\widehat{\Delta g}_{\mathtt{sps}} = \frac{1}{w_5}\begin{bmatrix} \mathbf{0}_{n\times 4n} & I_n\end{bmatrix}\mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1}r_{\mathbf{b}}. \tag{4.1.17}$$

Proof. Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the circulant structured GSPP of the form (4.1.1). We need to construct sparsity preserving perturbations ΔA, ΔB, $\Delta D \in \mathcal{C}_n$, and perturbations $\Delta f \in \mathbb{C}^n$ and $\Delta g \in \mathbb{C}^n$. By Definition 4.1.3, ΔA, ΔB, ΔD, Δf, and Δg satisfy

$$\Delta A\widetilde{\boldsymbol{u}} + \Delta B^T\widetilde{\boldsymbol{p}} - \Delta f = r_f \tag{4.1.18}$$

$$\text{and} \quad \Delta B\widetilde{\boldsymbol{u}} + \Delta D\widetilde{\boldsymbol{p}} - \Delta g = r_g. \tag{4.1.19}$$

To maintain the sparsity pattern of A, B and D on the perturbation matrices, we replace ΔA, ΔB and ΔD by $\Delta A \odot \Theta_A$, $\Delta B \odot \Theta_B$ and $\Delta D \odot \Theta_D$, respectively. Consequently, from (4.1.18) and (4.1.19), we get

$$w_1^{-1}w_1(\Delta A \odot \Theta_A)\widetilde{\boldsymbol{u}} + w_2^{-1}w_2(\Delta B \odot \Theta_B)^T\widetilde{\boldsymbol{p}} - w_4^{-1}w_4\Delta f = r_f, \tag{4.1.20}$$

$$\text{and} \quad w_2^{-1}w_2(\Delta B \odot \Theta_B)\widetilde{\boldsymbol{u}} + w_3^{-1}w_3(\Delta D \odot \Theta_D)\widetilde{\boldsymbol{p}} - w_5^{-1}w_5\Delta g = r_g. \tag{4.1.21}$$

Applying Lemma 4.1.4 in (4.1.20), we obtain

$$\begin{aligned} &w_1^{-1}\mathtt{Cr}\left(\widetilde{\boldsymbol{u}}\right)\mathfrak{D}_{c(A)}w_1\mathrm{vec}_{\mathcal{C}}(\Delta A \odot \Theta_A) \\ &\quad + w_2^{-1}\mathcal{H}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{c(B)}w_2\mathrm{vec}_{\mathcal{C}}(\Delta B \odot \Theta_B) - w_4^{-1}w_4\Delta f = r_f. \end{aligned} \tag{4.1.22}$$

Multiplying and dividing by $\mathfrak{D}_{\boldsymbol{a}}$ in (4.1.22), we get

$$w_1^{-1}\mathtt{Cr}(\widetilde{\boldsymbol{u}})\mathfrak{D}_{c(A)}\mathfrak{D}_{\boldsymbol{a}}^{-1}\mathfrak{D}_{\boldsymbol{a}}w_1\mathrm{vec}_{\mathcal{C}}(\Delta A \odot \Theta_A) + w_2^{-1}\mathcal{H}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{c(B)}\mathfrak{D}_{\boldsymbol{a}}^{-1}\mathfrak{D}_{\boldsymbol{a}}w_2\mathrm{vec}_{\mathcal{C}}(\Delta B \odot \Theta_B) - w_4^{-1}w_4\Delta f = r_f. \tag{4.1.23}$$

We can reformulate (4.1.23) as follows:

$$\mathcal{X}_1\Delta\mathcal{E} = r_f, \tag{4.1.24}$$

where

$$\mathcal{X}_1 = \begin{bmatrix} w_1^{-1}\mathtt{Cr}(\widetilde{\boldsymbol{u}})\mathfrak{D}_{c(A)}\mathfrak{D}_{\boldsymbol{a}}^{-1} & w_2^{-1}\mathcal{H}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{c(B)}\mathfrak{D}_{\boldsymbol{a}}^{-1} & \mathbf{0} & -w_4^{-1}I_n & \mathbf{0} \end{bmatrix} \in \mathbb{C}^{n\times 5n}$$

and

$$\Delta\mathcal{E} = \begin{bmatrix} w_1\mathfrak{D}_{\boldsymbol{a}}\mathrm{vec}_{\mathcal{C}}(\Delta A \odot \Theta_A) \\ w_2\mathfrak{D}_{\boldsymbol{a}}\mathrm{vec}_{\mathcal{C}}(\Delta B \odot \Theta_B) \\ w_3\mathfrak{D}_{\boldsymbol{a}}\mathrm{vec}_{\mathcal{C}}(\Delta D \odot \Theta_D) \\ w_4\Delta f \\ w_5\Delta g \end{bmatrix} \in \mathbb{C}^{5n}. \tag{4.1.25}$$

Here, the matrix $\mathfrak{D}_{\mathbf{a}}$ satisfy $\|\mathfrak{D}_{\mathbf{a}}\mathrm{vec}_{\mathcal{C}}(A)\|_2 = \|A\|_F$, for any $A \in \mathcal{C}_n$.
Similarly, applying Lemma 4.1.4 to (4.1.21), we obtain

$$w_2^{-1}\mathtt{Cr}(\widetilde{\boldsymbol{u}})\mathfrak{D}_{c(B)}w_2\mathrm{vec}_{\mathcal{C}}(\Delta B \odot \Theta_B) + w_3^{-1}\mathtt{Cr}\left(\widetilde{\boldsymbol{p}}\right)\mathfrak{D}_{c(D)}w_3\mathrm{vec}_{\mathcal{C}}(\Delta D \odot \Theta_D) - w_5^{-1}w_5\Delta g = r_g. \tag{4.1.26}$$

Thus, we can reformulate (4.1.26) as follows:

$$\mathcal{X}_2\Delta\mathcal{E} = r_g, \tag{4.1.27}$$

where $\mathcal{X}_2 = \begin{bmatrix} \mathbf{0} & w_2^{-1}\mathtt{Cr}(\widetilde{\boldsymbol{u}})\mathfrak{D}_{c(B)}\mathfrak{D}_{\mathbf{a}}^{-1} & w_3^{-1}\mathtt{Cr}\left(\widetilde{\boldsymbol{p}}\right)\mathfrak{D}_{c(D)}\mathfrak{D}_{\mathbf{a}}^{-1} & \mathbf{0} & -w_5^{-1}I_n \end{bmatrix} \in \mathbb{C}^{n\times 5n}$.
By combining (4.1.24) and (4.1.27), we get the following equivalent linear system of (4.1.18)–(4.1.19):

$$\mathcal{X}_{\mathtt{Cr}}\Delta\mathcal{E} \triangleq \begin{bmatrix} \mathcal{X}_1 \\ \mathcal{X}_2 \end{bmatrix}\Delta\mathcal{E} = r_{\mathbf{b}}, \tag{4.1.28}$$

where $\mathcal{X}_{\mathtt{Cr}} = \begin{bmatrix} \mathcal{X}_1 \\ \mathcal{X}_2 \end{bmatrix} \in \mathbb{C}^{2n\times 5n}$. Clearly, for $w_4, w_5 \neq 0$, the matrix $\mathcal{X}_{\mathtt{Cr}}$ has full row rank. Consequently, the linear system (4.1.28) is consistent, and $\mathcal{X}_{\mathtt{Cr}}^{\dagger} = \mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1}$, and by Lemma 1.3.1, the minimum norm solution of the system is given by

$$\Delta\mathcal{E}_{\min} := \mathcal{X}_{\mathtt{Cr}}^{\dagger}r_{\mathbf{b}} = \mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1}r_{\mathbf{b}}. \tag{4.1.29}$$

Next, the minimization problem in Definition 4.1.3 is equivalently written as

$$[\boldsymbol{\eta}^{\mathcal{S}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{p}})]^2 = \min\Big\{ w_1^2\|\Delta A\odot\Theta_A\|_F^2 + w_2^2\|\Delta B\odot\Theta_B\|_F^2 + w_3^3\|\Delta D\odot\Theta_D\|_F^2 + w_4^2\|\Delta f\|_2^2 + w_5^2\|\Delta g\|_2^2 \,\Big|\, \begin{pmatrix} \Delta A\odot\Theta_A, \Delta B\odot\Theta_B, \\ \Delta D\odot\Theta_D, \Delta f, \Delta g \end{pmatrix} \in \mathcal{S}_1 \Big\}$$
$$= \min\left\{ \|\Delta\mathcal{E}\|_2^2 \,\middle|\, \mathcal{X}_{\mathtt{Cr}}\Delta\mathcal{E} = r_{\mathbf{b}} \right\} = \|\Delta\mathcal{E}_{\min}\|_2^2. \tag{4.1.30}$$

Hence, using (4.1.29) and (4.1.30), the structured BE is given by

$$\boldsymbol{\eta}^{\mathcal{S}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{p}}) = \left\| \mathcal{X}_{\mathtt{Cr}}^H(\mathcal{X}_{\mathtt{Cr}}\mathcal{X}_{\mathtt{Cr}}^H)^{-1} r_{\mathbf{b}} \right\|_2.$$

From (4.1.25), we get $w_1\mathfrak{D}_{\boldsymbol{a}}\mathrm{vec}_{\mathcal{C}}(\Delta A\odot\Theta_A) = \begin{bmatrix} I_n & \mathbf{0}_{n\times 4n} \end{bmatrix}\Delta\mathcal{E}$. Thus, the minimal perturbation $\widehat{\Delta A}_{\mathtt{sps}}$ is given by

$$\widehat{\Delta A}_{\mathtt{sps}} = \mathtt{Cr}\left(\frac{1}{w_1}\mathfrak{D}_{\boldsymbol{a}}^{-1} \begin{bmatrix} I_n & \mathbf{0}_{n\times 4n} \end{bmatrix} \Delta\mathcal{E}_{\min} \right).$$

Similarly, we can obtain other minimal perturbations. Hence, the proof is completed. ■

In the following corollary, we present an explicit formula for the structured BE $\boldsymbol{\eta}^{\mathcal{S}_1}(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{p}})$ for the circulant structured GSPP (4.1.1) without maintaining the sparsity pattern in the perturbation matrices.

Corollary 4.1.1. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be the approximate solution of the circulant structured GSPP, i.e., $A, B, D \in \mathcal{C}_n$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}^{\mathcal{S}_1}(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{p}}) = \left\| \widehat{\mathcal{X}}_{\mathtt{Cr}}^H(\widehat{\mathcal{X}}_{\mathtt{Cr}}\widehat{\mathcal{X}}_{\mathtt{Cr}}^H)^{-1} r_{\mathbf{b}} \right\|_2, \tag{4.1.31}$$

where

$$\widehat{\mathcal{X}}_{\mathtt{Cr}} = \begin{bmatrix} \frac{1}{w_1}\mathtt{Cr}\left(\widetilde{\boldsymbol{u}}\right)\mathfrak{D}_{\boldsymbol{a}}^{-1} & \frac{1}{w_2}\mathcal{H}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{\boldsymbol{a}}^{-1} & \mathbf{0} & -\frac{1}{w_4}I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2}\mathtt{Cr}\left(\widetilde{\boldsymbol{u}}\right)\mathfrak{D}_{\boldsymbol{a}}^{-1} & \frac{1}{w_3}\mathtt{Cr}\left(\widetilde{\boldsymbol{p}}\right)\mathfrak{D}_{\boldsymbol{a}}^{-1} & \mathbf{0} & -\frac{1}{w_5}I_n \end{bmatrix}. \tag{4.1.32}$$

Proof. Since we are not maintaining the sparsity pattern to the perturbation matrices, we consider $\Theta_A = \Theta_B = \Theta_D = \mathbf{1}_{n\times n}$. Then $\Delta A\odot\Theta_A = \Delta A$, $\Delta B\odot\Theta_B = \Delta B$ and $\Delta C\odot\Theta_C = \Delta C$. Also, $\mathrm{vec}_{\mathcal{C}}(\Theta_A) = \mathrm{vec}_{\mathcal{C}}(\Theta_B) = \mathrm{vec}_{\mathcal{C}}(\Theta_C) = \mathbf{1}_n$. Consequently, the proof is completed using the formula stated in Theorem 4.1.5. ■

The minimal perturbations $\widehat{\Delta A}$, $\widehat{\Delta B}$, $\widehat{\Delta D}$, $\widehat{\Delta f}$, and $\widehat{\Delta g}$ to the Problem 4.1.2 are given by formulae (4.1.13)-(4.1.17) with $\mathcal{X}_{\mathtt{Cr}} = \widehat{\mathcal{X}}_{\mathtt{Cr}}$.

4.1.3. Structured BEs for Toeplitz Structured GSPPs

This subsection focuses on the derivation of compact formulae for the structured BEs $\boldsymbol{\eta}^{\mathcal{S}_2}_{\mathsf{sps}}(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{p}})$ and $\boldsymbol{\eta}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{p}})$ for Toeplitz structured GSPPs with and without preserving sparsity pattern, respectively. In addition, the minimal perturbations are provided for the Problem 4.1.2 for which the structured BEs are obtained. To accomplish this, we first derive the following lemma.

Lemma 4.1.6. *Let $A, B, M \in \mathcal{T}_{m\times n}$ with generator vectors*

$$\mathrm{vec}_{\mathcal{T}}(A) = [a_{-m+1},\ldots,a_{-1},a_0,a_1\ldots,a_{n-1}]^T \in \mathbb{C}^{n+m-1},$$

$$\mathrm{vec}_{\mathcal{T}}(B) = [b_{-m+1},\ldots,b_{-1},b_0,b_1\ldots,b_{n-1}]^T \in \mathbb{C}^{n+m-1} \text{ and}$$

$$\mathrm{vec}_{\mathcal{T}}(M) = [m_{-m+1},\ldots,m_{-1},m_0,m_1\ldots,m_{n-1}]^T \in \mathbb{C}^{n+m-1},$$

respectively. Suppose $x = [x_1,\ldots,x_n]^T \in \mathbb{C}^n$ and $y = [y_1,\ldots,y_m]^T \in \mathbb{C}^m$. Then

$$(A \odot \Theta_M)x = \mathcal{K}_x \mathfrak{D}_{t(M)} \mathrm{vec}_{\mathcal{T}}(A \odot \Theta_M) \quad \text{and}$$

$$(B \odot \Theta_M)^T y = \mathcal{G}_y \mathfrak{D}_{t(M)} \mathrm{vec}_{\mathcal{T}}(B \odot \Theta_M),$$

where $t(M) = \mathrm{vec}_{\mathcal{T}}(\Theta_M)$,

$$\mathcal{K}_x = \begin{bmatrix} 0 & \cdots & \cdots & 0 & \overset{m^{th}term}{\overset{\uparrow}{x_1}} & \cdots & \cdots & \cdots & x_{n-1} & x_n \\ \vdots & \cdots & 0 & x_1 & x_2 & \cdots & \cdots & x_{n-1} & x_n & 0 \\ \vdots & \iddots & \iddots & \iddots & & & \iddots & \iddots & \iddots & \vdots \\ 0 & \iddots & \iddots & & & \iddots & \iddots & \iddots & & \vdots \\ x_1 & x_2 & \cdots & & x_{n-1} & x_n & 0 & \cdots & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{m\times(m+n-1)},$$

$$\mathcal{G}_y = \begin{bmatrix} y_m & y_{m-1} & \cdots & y_1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & y_m & y_{m-1} & \cdots & y_1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \cdots & & 0 & y_m & y_{m-1} & \cdots & y_1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & \underset{\downarrow}{\underset{n^{th}term}{y_m}} & y_{m-1} & \cdots & y_1 \end{bmatrix} \in \mathbb{C}^{n\times(m+n-1)}.$$

Proof. The proof proceeds in a similar manner to the proof of Lemma 4.1.4. ■

In the following theorem, we derive the explicit formula for the structured BE $\boldsymbol{\eta}_{\mathtt{sps}}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$. Prior to that, we introduce the following notation:

$$\mathfrak{D}_{t(A)} = \mathrm{diag}(\mathrm{vec}_{\mathcal{T}}(\Theta_A)), \quad \mathfrak{D}_{t(B)} = \mathrm{diag}(\mathrm{vec}_{\mathcal{T}}(\Theta_B)), \tag{4.1.33}$$

$$\mathfrak{D}_{t(D)} = \mathrm{diag}(\mathrm{vec}_{\mathcal{T}}(\Theta_D)) \quad \text{and} \quad \mathfrak{D}_{\mathbf{t}_{mn}} = \mathrm{diag}(\mathbf{t}_{mn}), \tag{4.1.34}$$

where $\mathbf{t}_{mn} = [1, \sqrt{2}, \ldots, \sqrt{m-1}, \sqrt{\min\{m,n\}}, \sqrt{n-1}, \ldots, \sqrt{2}, 1]^T \in \mathbb{R}^{m+n-1}$. When $n = m$, we write $\mathbf{t}_{mn} = \mathbf{t}_n$ (or $\mathbf{t}_m$).

Theorem 4.1.7. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be the approximate solution of the Toeplitz structured GSPP (4.1.1), i.e., $A \in \mathcal{T}_{n\times n}, B \in \mathcal{T}_{m\times n}, D \in \mathcal{T}_{m\times m}$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}_{\mathtt{sps}}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \mathcal{X}_{\mathcal{T}}^H (\mathcal{X}_{\mathcal{T}} \mathcal{X}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}} \right\|_2, \tag{4.1.35}$$

where $\mathcal{X}_{\mathcal{T}} \in \mathbb{C}^{(n+m)\times(4n+4m-3)}$ is given by

$$\mathcal{X}_{\mathcal{T}} = \begin{bmatrix} \frac{1}{w_1}\mathcal{K}_{\widetilde{\boldsymbol{u}}}\mathfrak{D}_{t(A)}\mathfrak{D}_{\mathbf{t}_n}^{-1} & \frac{1}{w_2}\mathcal{G}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{t(B)}\mathfrak{D}_{\mathbf{t}_{mn}}^{-1} & \mathbf{0} & -\frac{1}{w_4}I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2}\mathcal{K}_{\widetilde{\boldsymbol{u}}}\mathfrak{D}_{t(B)}\mathfrak{D}_{\mathbf{t}_{mn}}^{-1} & \frac{1}{w_3}\mathcal{K}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{t(D)}\mathfrak{D}_{\mathbf{t}_m}^{-1} & \mathbf{0} & -\frac{1}{w_5}I_m \end{bmatrix},$$

$r_{\mathbf{b}} = \begin{bmatrix} r_f^T, & r_g^T \end{bmatrix}^T$, $r_f = f - A\widetilde{\boldsymbol{u}} - B^T\widetilde{\boldsymbol{p}}$, *and* $r_g = g - B\widetilde{\boldsymbol{u}} - D\widetilde{\boldsymbol{p}}$.

Furthermore, the minimal perturbations to the Problem 4.1.2 are given by

$$\widehat{\Delta A}_{\mathtt{sps}} = \mathcal{T}\left(\frac{1}{w_1}\mathfrak{D}_{\mathbf{t}_n}^{-1}\begin{bmatrix} I_{2n-1} & \mathbf{0}_{(2n-1)\times(2n+4m-2)} \end{bmatrix} \mathcal{X}_{\mathcal{T}}^H (\mathcal{X}_{\mathcal{T}} \mathcal{X}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}}\right), \tag{4.1.36}$$

$$\widehat{\Delta B}_{\mathtt{sps}} = \mathcal{T}\left(\frac{1}{w_2}\mathfrak{D}_{\mathbf{t}_{mn}}^{-1}\begin{bmatrix} \mathbf{0}_{(m+n-1)\times(2n-1)} & I_{m+n-1} & \mathbf{0}_{(m+n-1)\times(3m+n-1)} \end{bmatrix} \mathcal{X}_{\mathcal{T}}^H (\mathcal{X}_{\mathcal{T}} \mathcal{X}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}}\right), \tag{4.1.37}$$

$$\widehat{\Delta D}_{\mathtt{sps}} = \mathcal{T}\left(\frac{1}{w_3}\mathfrak{D}_{\mathbf{t}_m}^{-1}\begin{bmatrix} \mathbf{0}_{(2m-1)\times(3n+3m-2)} & I_{2m-1} & \mathbf{0}_{(2m-1)\times(n+m)} \end{bmatrix} \mathcal{X}_{\mathcal{T}}^H (\mathcal{X}_{\mathcal{T}} \mathcal{X}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}}\right), \tag{4.1.38}$$

$$\widehat{\Delta f}_{\mathtt{sps}} = \frac{1}{w_4}\begin{bmatrix} \mathbf{0}_{n\times(3n+3m-3)} & I_n & \mathbf{0}_{n\times m} \end{bmatrix} \mathcal{X}_{\mathcal{T}}^H (\mathcal{X}_{\mathcal{T}} \mathcal{X}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}}, \text{ and} \tag{4.1.39}$$

$$\widehat{\Delta g}_{\mathtt{sps}} = \frac{1}{w_5}\begin{bmatrix} \mathbf{0}_{m\times(4n+3m-3)} & I_m \end{bmatrix} \mathcal{X}_{\mathcal{T}}^H (\mathcal{X}_{\mathcal{T}} \mathcal{X}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}}. \tag{4.1.40}$$

Proof. We need to construct perturbation matrices $\Delta A \in \mathcal{T}_{n\times n}, \Delta B \in \mathcal{T}_{m\times n}, \Delta D \in \mathcal{T}_{m\times m}$ (which preserves the sparsity pattern of the original matrices), $\Delta f \in \mathbb{C}^n$ and $\Delta g \in \mathbb{C}^m$ for the approximate solution $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$. By Definition 4.1.3, $\Delta A, \Delta B, \Delta D, \Delta f$, and Δg satisfy

$$\Delta A\widetilde{\boldsymbol{u}} + \Delta B^T\widetilde{\boldsymbol{p}} - \Delta f = r_f,$$

$$\Delta B\widetilde{\boldsymbol{u}} + \Delta D\widetilde{\boldsymbol{p}} - \Delta g = r_g.$$

Following the proof method of Theorem 4.1.5 and using Lemma 4.1.6, we obtain $r_{\mathbf{b}} = \mathcal{X}_{\mathcal{T}}\Delta\mathcal{E}$, where

$$\Delta\mathcal{E} = \begin{bmatrix} w_1\mathfrak{D}_{\mathbf{t}_n}\text{vec}_{\mathcal{T}}(\Delta A \odot \Theta_A) \\ w_2\mathfrak{D}_{\mathbf{t}_{mn}}\text{vec}_{\mathcal{T}}(\Delta B \odot \Theta_B) \\ w_3\mathfrak{D}_{\mathbf{t}_m}\text{vec}_{\mathcal{T}}(\Delta D \odot \Theta_D) \\ w_4\Delta f \\ w_5\Delta g \end{bmatrix} \in \mathbb{C}^{4n+4m-3}. \tag{4.1.41}$$

Hence, an analogous way to proof of Theorem 4.1.5, we get

$$\boldsymbol{\eta}^{\mathcal{S}_2}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \mathcal{X}_{\mathcal{T}}^H(\mathcal{X}_{\mathcal{T}}\mathcal{X}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}} \right\|_2.$$

From (4.1.41), we get $w_1\mathfrak{D}_{\mathbf{t}_n}\text{vec}_{\mathcal{T}}(\Delta A \odot \Theta_A) = \begin{bmatrix} I_{2n-1} & \mathbf{0}_{(2n-1)\times(2n+4m-2)} \end{bmatrix} \Delta\mathcal{E}$. Therefore, the minimal perturbation matrix $\widehat{\Delta A}_{\mathtt{sps}}$, which preserve the sparsity pattern of A is given by

$$\widehat{\Delta A}_{\mathtt{sps}} = \frac{1}{w_1}\mathfrak{D}_{\mathbf{t}_n}^{-1} \begin{bmatrix} I_{2n-1} & \mathbf{0}_{(2n-1)\times(2n+4m-2)} \end{bmatrix} \Delta\mathcal{E}_{\min}.$$

Similarly, we can obtain the minimal perturbations $\widehat{\Delta B}_{\mathtt{sps}}, \widehat{\Delta D}_{\mathtt{sps}}, \widehat{\Delta f}_{\mathtt{sps}}$ and $\widehat{\Delta g}_{\mathtt{sps}}$. Hence, the proof is completed. ■

The next corollary presents a formula for $\boldsymbol{\eta}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ without considering the sparsity pattern in the input matrices.

Corollary 4.1.2. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be the approximate solution of the Toeplitz structured GSPP (4.1.1), i.e., $A \in \mathcal{T}_{n\times n}$, $B \in \mathcal{T}_{m\times n}$, $D \in \mathcal{T}_n$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \widehat{\mathcal{X}}_{\mathcal{T}}^H(\widehat{\mathcal{X}}_{\mathcal{T}}\widehat{\mathcal{X}}_{\mathcal{T}}^H)^{-1} r_{\mathbf{b}} \right\|_2, \tag{4.1.42}$$

where $\widehat{\mathcal{X}}_{\mathcal{T}} \in \mathbb{C}^{(n+m)\times(4n+4m-3)}$ is given by

$$\widehat{\mathcal{X}}_{\mathcal{T}} = \begin{bmatrix} \frac{1}{w_1}\mathcal{K}_{\widetilde{\boldsymbol{u}}}\mathfrak{D}_{\mathbf{t}_n}^{-1} & \frac{1}{w_2}\mathcal{G}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{\mathbf{t}_{mn}}^{-1} & \mathbf{0} & -\frac{1}{w_4}I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2}\mathcal{K}_{\widetilde{\boldsymbol{u}}}\mathfrak{D}_{\mathbf{t}_{mn}}^{-1} & \frac{1}{w_3}\mathcal{K}_{\widetilde{\boldsymbol{p}}}\mathfrak{D}_{\mathbf{t}_m}^{-1} & \mathbf{0} & -\frac{1}{w_5}I_m \end{bmatrix}. \tag{4.1.43}$$

Proof. Since the sparsity pattern of the perturbation matrices is not taken care of, we consider $\Theta_A = \mathbf{1}_{n\times n}$, $\Theta_B = \mathbf{1}_{m\times n}$ and $\Theta_D = \mathbf{1}_{m\times m}$. Then $\Delta A \odot \Theta_A = \Delta A$, $\Delta B \odot \Theta_B = \Delta B$, and $\Delta D \odot \Theta_D = \Delta D$. Also, $\text{vec}_{\mathcal{T}}(\Theta_A) = \mathbf{1}_{2n-1}$, $\text{vec}_{\mathcal{T}}(\Theta_B) = \mathbf{1}_{m+n-1}$ and $\text{vec}_{\mathcal{T}}(\Theta_D) = \mathbf{1}_{2m-1}$. As a result, the proof follows using the formula stated in Theorem 4.1.7. ■

Note that the minimal perturbations $\widehat{\Delta A}, \widehat{\Delta B}, \widehat{\Delta D}, \widehat{\Delta f}$, and $\widehat{\Delta g}$ to the Problem 4.1.2 are given by formulae (4.1.36)-(4.1.40) with $\mathcal{X}_{\mathcal{T}} = \widehat{\mathcal{X}}_{\mathcal{T}}$.

4.1.4. Structured BEs Symmetric-Toeplitz Structured GSPPs

In this subsection, we derive concise formulae for the structured BE $\boldsymbol{\eta}_{\mathbf{sps}}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ and $\boldsymbol{\eta}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ for symmetric-Toeplitz structured GSPPs with and without preserving the sparsity pattern, respectively. Since the $(1,2)$ block matrix B is symmetric, thus the case $n = m$ follows. In many applications, such as the WRLS problem, the block matrices A and D do not follow any particular structure, in this subsection, we focus on the structured BE when the perturbation matrix ΔB follows symmetric-Toeplitz structure of B. To find the structured BE, we present the following lemmas that are crucial in establishing our main results.

Lemma 4.1.8. *Let $A, M \in \mathcal{ST}_n$ with generator vectors $\text{vec}_{\mathcal{ST}}(A) = [a_0, a_1 \ldots, a_{n-1}]^T \in \mathbb{C}^n$ and $\text{vec}_{\mathcal{ST}}(M) = [m_0, m_1 \ldots, m_{n-1}]^T \in \mathbb{C}^n$, respectively. Suppose $x = [x_1, \ldots, x_n]^T \in \mathbb{C}^n$, then*

$$(A \odot \Theta_M)x = \mathcal{I}_x \mathfrak{D}_{s(M)} \text{vec}_{\mathcal{ST}}(A \odot \Theta_M),$$

where $s(M) = \text{vec}_{\mathcal{ST}}(\Theta_M)$ and $\mathcal{I}_x \in \mathbb{C}^{n \times n}$ is given by

$$\mathcal{I}_x = \begin{bmatrix} x_1 & \cdots & \cdots & x_{n-1} & x_n \\ x_2 & \cdots & \cdots & x_n & 0 \\ \vdots & & \iddots & \iddots & \vdots \\ \vdots & x_n & 0 & \cdots & \vdots \\ x_n & 0 & \cdots & \cdots & 0 \end{bmatrix} + \begin{bmatrix} 0 & \cdots & \cdots & & 0 \\ 0 & x_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & x_{n-2} & \cdots & x_1 & 0 \\ 0 & x_{n-1} & \cdots & \cdots & x_1 \end{bmatrix}. \tag{4.1.44}$$

Proof. The proof proceeds in a similar manner to the proof of Lemma 4.1.4. ■

Lemma 4.1.9. *Let $A, B, M \in \mathbb{C}^{m \times n}$ be three matrices. Suppose that $x = [x_1, \ldots, x_n]^T \in \mathbb{C}^n$ and $y = [y_1, \ldots, y_m]^T \in \mathbb{C}^m$. Then, $(A \odot \Theta_M)x = M_x^m \mathfrak{D}_{\text{vec}(\Theta_M)} \text{vec}(A \odot \Theta_M)$ and $(B \odot \Theta_M)^T y = N_y^n \mathfrak{D}_{\text{vec}(\Theta_M)} \text{vec}(B \odot \Theta_M)$, where $M_x^m = x^T \otimes I_m \in \mathbb{C}^{m \times mn}$ and $N_y^n = I_n \otimes y^T \in \mathbb{C}^{n \times mn}$.*

Proof. The proof proceeds in a similar method to the proof of Lemma 4.1.4. ■

Next, we derive concrete formulae for $\boldsymbol{\eta}_{\mathbf{sps}}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ and $\boldsymbol{\eta}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$, which are the main result of this subsection. Before proceeding, we introduce the following notations:

$$\mathfrak{D}_{s(A)} = \text{diag}(\text{vec}_{\mathcal{ST}}(\Theta_A)), \quad \mathfrak{D}_{s(B)} = \text{diag}(\text{vec}_{\mathcal{ST}}(\Theta_b)), \tag{4.1.45}$$

$$\mathfrak{D}_{s(D)} = \text{diag}(\text{vec}_{\mathcal{ST}}(\Theta_D)) \quad \text{and} \quad \mathfrak{D}_{\mathbf{s}} = \text{diag}(\mathbf{s}), \tag{4.1.46}$$

where $\mathbf{s} = [\sqrt{n}, \sqrt{2(n-1)}, \sqrt{2(n-2)}, \ldots, \sqrt{2}]^T \in \mathbb{R}^n$.

Theorem 4.1.10. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the symmetric-Toeplitz structured GSPP (4.1.1), i.e., $B \in \mathcal{ST}_n$, $A, D \in \mathbb{C}^{n\times n}$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}_{\mathtt{sps}}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \mathcal{Z}_{\mathcal{ST}}^H (\mathcal{Z}_{\mathcal{ST}} \mathcal{Z}_{\mathcal{ST}}^H)^{-1} r_{\mathbf{b}} \right\|_2, \tag{4.1.47}$$

where $\mathcal{Z}_{\mathcal{ST}} \in \mathbb{C}^{2n\times l}$ is given by

$$\mathcal{Z}_{\mathcal{ST}} = \begin{bmatrix} \frac{1}{w_1} M_{\widetilde{\boldsymbol{u}}}^n \mathfrak{D}_{\mathrm{vec}(\Theta_A)} & \frac{1}{w_2} \mathcal{I}_{\widetilde{\boldsymbol{p}}} \mathfrak{D}_{s(B)} \mathfrak{D}_{\mathbf{s}}^{-1} & \mathbf{0} & -\frac{1}{w_4} I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2} \mathcal{I}_{\widetilde{\boldsymbol{u}}} \mathfrak{D}_{s(B)} \mathfrak{D}_{\mathbf{s}}^{-1} & \frac{1}{w_3} M_{\widetilde{\boldsymbol{p}}}^n \mathfrak{D}_{\mathrm{vec}(\Theta_D)} & \mathbf{0} & -\frac{1}{w_5} I_n \end{bmatrix}, \tag{4.1.48}$$

$r_{\mathbf{b}} = [r_f^T, \ r_g^T]^T$, $r_f = f - A\widetilde{\boldsymbol{u}} - B\widetilde{\boldsymbol{p}}$, $r_g = g - B\widetilde{\boldsymbol{u}} - D\widetilde{\boldsymbol{p}}$, *and* $l = 2n^2 + 3n$.

Furthermore, the minimal perturbations to the Problem 4.1.2 are given by

$$\mathrm{vec}(\widehat{\Delta A}_{\mathtt{sps}}) = \frac{1}{w_1} \begin{bmatrix} I_{n^2} & \mathbf{0}_{n^2\times(n^2+3n)} \end{bmatrix} \mathcal{Z}_{\mathcal{ST}}^H (\mathcal{Z}_{\mathcal{ST}} \mathcal{Z}_{\mathcal{ST}}^H)^{-1} r_{\mathbf{b}}, \tag{4.1.49}$$

$$\widehat{\Delta B}_{\mathtt{sps}} = \mathcal{ST}\left(\frac{1}{w_2} \mathfrak{D}_{\mathbf{s}}^{-1} \begin{bmatrix} \mathbf{0}_{n\times n^2} & I_n & \mathbf{0}_{n\times(2n^2+2n)} \end{bmatrix} \mathcal{Z}_{\mathcal{ST}}^H (\mathcal{Z}_{\mathcal{ST}} \mathcal{Z}_{\mathcal{ST}}^H)^{-1} r_{\mathbf{b}} \right), \tag{4.1.50}$$

$$\mathrm{vec}(\widehat{\Delta D}_{\mathtt{sps}}) = \frac{1}{w_3} \begin{bmatrix} \mathbf{0}_{n^2\times(n^2+n)} & I_{n^2} & \mathbf{0}_{n^2\times(n^2+2n)} \end{bmatrix} \mathcal{Z}_{\mathcal{ST}}^H (\mathcal{Z}_{\mathcal{ST}} \mathcal{Z}_{\mathcal{ST}}^H)^{-1} r_{\mathbf{b}}, \tag{4.1.51}$$

$$\widehat{\Delta f}_{\mathtt{sps}} = \frac{1}{w_4} \begin{bmatrix} \mathbf{0}_{n\times(2n^2+n)} & I_n & \mathbf{0}_{n\times(2n^2+n)} \end{bmatrix} \mathcal{Z}_{\mathcal{ST}}^H (\mathcal{Z}_{\mathcal{ST}} \mathcal{Z}_{\mathcal{ST}}^H)^{-1} r_{\mathbf{b}}, \quad \textit{and} \tag{4.1.52}$$

$$\widehat{\Delta g}_{\mathtt{sps}} = \frac{1}{w_5} \begin{bmatrix} \mathbf{0}_{n\times(2n^2+2n)} & I_n \end{bmatrix} \mathcal{Z}_{\mathcal{ST}}^H (\mathcal{Z}_{\mathcal{ST}} \mathcal{Z}_{\mathcal{ST}}^H)^{-1} r_{\mathbf{b}}. \tag{4.1.53}$$

Proof. For an approximate solution $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$, we require to construct sparsity preserving perturbation matrices $\Delta B \in \mathcal{ST}_n, \Delta A, \Delta D \in \mathbb{C}^{n\times n}$, and perturbations $\Delta f \in \mathbb{C}^n$ and $\Delta g \in \mathbb{C}^n$. By Definition 4.1.3, we have

$$\Delta A\widetilde{\boldsymbol{u}} + \Delta B^T \widetilde{\boldsymbol{p}} - \Delta f = r_f, \tag{4.1.54}$$

$$\Delta B\widetilde{\boldsymbol{u}} + \Delta D\widetilde{\boldsymbol{p}} - \Delta g = r_g, \tag{4.1.55}$$

where $\Delta B \in \mathcal{ST}_n$, and $\Delta A, \Delta D \in \mathbb{C}^{n\times n}$.

Following the proof method of Theorem 4.1.5 and using Lemmas 4.1.8 and 4.1.9, we obtain $r_{\mathbf{b}} = \mathcal{Z}_{\mathcal{ST}} \Delta\mathcal{E}$, where

$$\Delta\mathcal{E} = \begin{bmatrix} w_1 \mathfrak{D}_{\mathbf{s}} \mathrm{vec}(\Delta A \odot \Theta_A) \\ w_2 \mathfrak{D}_{\mathbf{s}} \mathrm{vec}_{\mathcal{ST}}(\Delta B \odot \Theta_B) \\ w_3 \mathfrak{D}_{\mathbf{s}} \mathrm{vec}(\Delta D \odot \Theta_D) \\ w_4 \Delta f \\ w_5 \Delta g \end{bmatrix} \in \mathbb{C}^l. \tag{4.1.56}$$

Hence, following an analogous way to proof of Theorem 4.1.5, we get the desired structured BE and perturbation matrices. ■

Next, we present the formula for $\boldsymbol{\eta}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ without preserving the sparsity pattern.

Corollary 4.1.3. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the symmetric-Toeplitz structured GSPP (4.1.1), i.e., $B \in \mathcal{ST}_n$, $A, D \in \mathbb{C}^{n\times n}$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \widehat{\mathcal{Z}}_{\mathcal{ST}}^H (\widehat{\mathcal{Z}}_{\mathcal{ST}} \widehat{\mathcal{Z}}_{\mathcal{ST}}^H)^{-1} r_{\mathbf{b}} \right\|_2, \tag{4.1.57}$$

where

$$\widehat{\mathcal{Z}}_{\mathcal{ST}} = \begin{bmatrix} \frac{1}{w_1} M_{\widetilde{\boldsymbol{u}}}^n & \frac{1}{w_2} \mathcal{I}_{\widetilde{\boldsymbol{p}}} \mathfrak{D}_{\mathbf{s}}^{-1} & \mathbf{0} & -\frac{1}{w_4} I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2} \mathcal{I}_{\widetilde{\boldsymbol{u}}} \mathfrak{D}_{\mathbf{s}}^{-1} & \frac{1}{w_3} M_{\widetilde{\boldsymbol{p}}}^n & \mathbf{0} & -\frac{1}{w_5} I_n \end{bmatrix}. \tag{4.1.58}$$

Proof. Because we are not considering the sparsity pattern in the perturbation matrices, taking $\Theta_A = \Theta_B = \Theta_D = \mathbf{1}_{n\times n}$ in the Theorem 4.1.10 yields the desired result. ■

The minimal perturbation matrices $\widehat{\Delta A}, \widehat{\Delta B}, \widehat{\Delta D}, \widehat{\Delta f}$, and $\widehat{\Delta g}$ to the Problem 4.1.2 are given by formulae (4.1.49)-(4.1.53) with $\mathcal{Z}_{\mathcal{ST}} = \widehat{\mathcal{Z}}_{\mathcal{ST}}$.

Remark 4.1.11. *Applying our framework developed in this subsection and Subsections 4.1.2 and 4.1.3, we can obtain the structured BEs for the GSPP (4.1.1) when the block matrices possess only symmetric structure or Hankel structure (which is symmetric as well).*

4.1.5. Unstructured BEs with Preserving Sparsity Pattern

In this section, we address the scenario where A, B, and D in (4.1.1) are unstructured. Although the previous studies such as [146, 44, 162] have explored the BEs for the GSPP (4.1.1), their investigations do not take into account the sparsity pattern of the block matrices. Consequently, in this scenario, first, we define the unstructured BE as follows:

$$\eta(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) := \min_{\begin{pmatrix} \Delta A, \Delta B, \Delta D, \\ \Delta f, \Delta g \end{pmatrix} \in \mathcal{S}^0} \left\| \left[\Delta \mathcal{M} \quad \Delta \mathbf{b} \right] \right\|_{w,F} \tag{4.1.59}$$

where

$$\mathcal{S}^0 = \left\{ \begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta f, \Delta g \end{pmatrix} \Bigg| \begin{bmatrix} A + \Delta A & (B + \Delta B)^T \\ B + \Delta B & D + \Delta D \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{u}} \\ \widetilde{\boldsymbol{p}} \end{bmatrix} = \begin{bmatrix} f + \Delta f \\ g + \Delta g \end{bmatrix} \right\}.$$

The following result gives the formula for the unstructured BE of the GSPPs (4.1.1) when the sparsity pattern in the perturbation matrices is preserved, and in this case, we denote it by $\boldsymbol{\eta}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$.

Theorem 4.1.12. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the GSPP (4.1.1) with $A \in \mathbb{C}^{n\times n}, B \in \mathbb{C}^{m\times n}$, $D \in \mathbb{C}^{m\times m}$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\|\mathcal{N}^H(\mathcal{N}\mathcal{N}^H)^{-1} r_{\mathbf{b}}\right\|_2,$$

where $\mathcal{N} \in \mathbb{C}^{(m+n)\times k}$ is given by

$$\mathcal{N} = \begin{bmatrix} \frac{1}{w_1} M^n_{\widetilde{\boldsymbol{u}}} \mathfrak{D}_{\mathrm{vec}(\Theta_A)} & \frac{1}{w_2} N^n_{\widetilde{\boldsymbol{p}}} \mathfrak{D}_{\mathrm{vec}(\Theta_B)} & \mathbf{0} & -\frac{1}{w_4} I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2} M^m_{\widetilde{\boldsymbol{u}}} \mathfrak{D}_{\mathrm{vec}(\Theta_B)} & \frac{1}{w_3} N^m_{\widetilde{\boldsymbol{p}}} \mathfrak{D}_{\mathrm{vec}(\Theta_D)} & \mathbf{0} & -\frac{1}{w_5} I_m \end{bmatrix} \tag{4.1.60}$$

and $k = n^2 + m^2 + mn + n + m$.

Proof. The proof follows similarly to the proof method of Theorem 4.1.10. ■

The next corollary presents the BE formula when the sparsity is not considered.

Corollary 4.1.4. *Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the GSPP (4.1.1) with $A \in \mathbb{C}^{n\times n}, B \in \mathbb{C}^{m\times n}$ and $D \in \mathbb{C}^{m\times m}$, and $w_4, w_5 \neq 0$. Then, we have*

$$\boldsymbol{\eta}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\|\widehat{\mathcal{N}}^H(\widehat{\mathcal{N}}\widehat{\mathcal{N}}^H)^{-1} r_{\mathbf{b}}\right\|_2,$$

where $\widehat{\mathcal{N}} \in \mathbb{C}^{(m+n)\times l}$ is given by

$$\widehat{\mathcal{N}} = \begin{bmatrix} \frac{1}{w_1} M^n_{\widetilde{\boldsymbol{u}}} & \frac{1}{w_2} N^n_{\widetilde{\boldsymbol{p}}} & \mathbf{0} & -\frac{1}{w_4} I_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{w_2} M^m_{\widetilde{\boldsymbol{u}}} & \frac{1}{w_3} N^m_{\widetilde{\boldsymbol{p}}} & \mathbf{0} & -\frac{1}{w_5} I_m \end{bmatrix}. \tag{4.1.61}$$

Proof. The proof is followed by taking $\Theta_A = \mathbf{1}_{m\times m}$, $\Theta_B = \mathbf{1}_{m\times n}$ and $\Theta_D = \mathbf{1}_{n\times n}$ in the expression of $\boldsymbol{\eta}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$, presented in Theorem 4.1.12. ■

Note that when w_4 or w_5 are zero, the desired BE is achieved if $\mathcal{N}$ and $\widehat{\mathcal{N}}$ have full row rank. Nevertheless, in [90, 146], formulas for BEs with no special structure on block matrices are discussed, the following example illustrates that our BE can be smaller than theirs.

Example 4.1.2. Consider the GSPP (4.1.1), where $A = I_4$, $B = \begin{bmatrix} 2 & 1 & 3 & 1 \\ -1 & 2 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 4}$, $D = \mathbf{0}$, $f = [-1, 0, 2, 3]^T$, and $g = \mathbf{0}$. We take the approximate solution $[\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T = [-1.495, 1.505, 1.505, 1.505, 1.005, -0.495]^T$. Then, employing the formula provided in [90] with $\theta = 1$, the computed BE is 0.0410. Since, the (1,1) block in [90] has no perturbation, by considering $w_1 = 0$ in Corollary 4.1.4 (with $w_2 = w_4 = 1$, $w_3 = w_5 = 0$), the BE is 0.0288. This comparison highlights that our computed BE using Corollary 4.1.4 and [90] are of the same order, illustrating the reliability of our obtained BE.

4.1.6. Application to Derive the Structured BEs for the WRLS Problems

In this subsection, we present an application of our developed theory in deriving the structured BE for the WRLS problem (1.1.3). The minimization problem (1.1.3) can be reformulated as the following GSPP:

$$\widehat{\mathcal{M}}\begin{bmatrix}\mathbf{r}\\ z\end{bmatrix} \triangleq \begin{bmatrix}W^{-1} & K^T\\ K & -\lambda I_m\end{bmatrix}\begin{bmatrix}\mathbf{r}\\ z\end{bmatrix} = \begin{bmatrix}f\\ \mathbf{0}\end{bmatrix}, \tag{4.1.62}$$

where K is a Toeplitz or symmetric-Toeplitz matrix.

Since the weighting matrix W and the regularization matrix $-\lambda I_m$ are not allowed to be perturbed, we consider $(1,1)$ block and $(2,2)$ block has no perturbation. Let $[\widetilde{\boldsymbol{r}}^T,\ \widetilde{z}^T]^T$ be the approximate solution of (4.1.62), i.e., $\widetilde{z}$ be an approximate solution of the WRLS problem. Then, we define structured BE for the WRLS problem as follows:

$$\boldsymbol{\zeta}(\widetilde{z}) := \min_{(\Delta K,\Delta B)\in\mathcal{S}^{ls}} \left\| \left[[w_2\|\Delta K\|_F,\quad w_4\|\Delta f\|_2 \right] \right\|_2, \tag{4.1.63}$$

where

$$\mathcal{S}^{ls} := \left\{ (\Delta K, \Delta f) \ : \ \begin{bmatrix}W^{-1} & (K+\Delta K)^T\\ K+\Delta K & -\lambda I_m\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{r}}\\ \widetilde{z}\end{bmatrix} = \begin{bmatrix}f+\Delta f\\ \mathbf{0}\end{bmatrix},\ \Delta K \in \{\mathcal{T}_{m\times n}, \mathcal{ST}_n\} \right\}.$$

Before proceeding, we define $\mathcal{X}_{ls} \in \mathbb{R}^{(n+m)\times(2n+m-1)}$ and $\mathcal{Z}_{ls} \in \mathbb{R}^{2n\times 2n}$ as follows:

$$\mathcal{X}_{ls} = \begin{bmatrix}\frac{1}{w_2}\mathcal{G}_{\widetilde{z}}\mathfrak{D}_{t(K)}\mathfrak{D}_{\mathbf{t}_{mn}}^{-1} & -\frac{1}{w_4}I_n\\ \frac{1}{w_2}\mathcal{K}_{\widetilde{\mathbf{r}}}\mathfrak{D}_{t(K)}\mathfrak{D}_{\mathbf{t}_{mn}}^{-1} & \mathbf{0}\end{bmatrix} \quad \text{and} \quad \mathcal{Z}_{ls} = \begin{bmatrix}\frac{1}{w_2}\mathcal{I}_{\widetilde{z}}\mathfrak{D}_{t(K)}\mathfrak{D}_{\mathbf{s}}^{-1} & -\frac{1}{w_4}I_n\\ \frac{1}{w_2}\mathcal{I}_{\widetilde{\mathbf{r}}}\mathfrak{D}_{t(K)}\mathfrak{D}_{\mathbf{s}}^{-1} & \mathbf{0}\end{bmatrix},$$

where $t(K) = \mathrm{vec}_{\mathcal{T}}(\Theta_K)$, $\widetilde{r}_f = f - W^{-1}\widetilde{\mathbf{r}} - K^T\widetilde{z}$, and $\widetilde{r}_g = \lambda\widetilde{z} - K\widetilde{\mathbf{r}}$.

Theorem 4.1.13. *Let $\widetilde{z}$ be an approximate solution of the WRLS problem (1.1.3) with $K \in \{\mathcal{T}_{m\times n}, \mathcal{ST}_n\}$. Let $\widetilde{r}_d = [\widetilde{r}_f^T,\ \ \widetilde{r}_g^T]^T$, then*

1. *when $K \in \mathcal{T}_{m\times n}$ and $\mathrm{rank}(\mathcal{X}_{ls}) = \mathrm{rank}\left([\mathcal{X}_{ls}\ \widetilde{r}_d]\right)$, we have*

$$\boldsymbol{\zeta}(\widetilde{z}) = \left\| \mathcal{X}_{ls}^{\dagger}\widetilde{r}_d \right\|_2, \tag{4.1.64}$$

2. *when $K \in \mathcal{ST}_n$ and $\mathrm{rank}(\mathcal{Z}_{ls}) = \mathrm{rank}\left([\mathcal{Z}_{ls}\ \widetilde{r}_d]\right)$, we have*

$$\boldsymbol{\zeta}(\widetilde{z}) = \left\| \mathcal{Z}_{ls}^{\dagger}\widetilde{r}_d \right\|_2. \tag{4.1.65}$$

Proof. First, we consider $K \in \mathcal{T}_{m\times n}$. Since, W and $-\lambda I_m$ are not required to be perturbed, we take $w_1 = 0$ and $w_3 = 0$. Following the proof method of Theorem 4.1.7, we obtain that $(\Delta K, \Delta B) \in \mathcal{S}^{ls}$, $\Delta K \in \mathcal{T}_{m\times n}$ if and only if

$$\mathcal{X}_{ls}\Delta\mathcal{E}_{ls} = \widetilde{r}_d, \tag{4.1.66}$$

where $\Delta\mathcal{E}_{ls} = \begin{bmatrix} w_2\mathfrak{D}_{\mathbf{t}_{mn}}\mathrm{vec}_{\mathcal{T}}(\Delta K \odot \Theta_K) \\ w_4\Delta B \end{bmatrix}$. Thus, when $\mathrm{rank}(\mathcal{X}_{ls}) = \mathrm{rank}\left([\mathcal{X}_{ls}\ \widetilde{r}_d]\right)$, the minimum norm solution of (4.1.66) is $\Delta\mathcal{E}^{ls}_{\min} = \mathcal{X}^{\dagger}_{ls}\widetilde{r}_d$. Hence, the structured BE $\boldsymbol{\zeta}(\widetilde{z})$ in (4.1.64) is attained. Similarly, for $K \in \mathcal{ST}_n\,(n = m)$, we can derive the structured BE given in (4.1.65). ■

Note that we can similarly obtain the structured BE for the WRLS when K is circulant.

4.2. Structured Backward Errors of Generalized Saddle Point Problems with Hermitian Block Matrices

In this section, we derive the structured BE for a class of GSPP by preserving the Hermitian and sparsity structures of the block matrices. Additionally, we construct the minimal backward perturbation matrices for which the structured BE is achieved. Our analysis further explores the structured BE when the sparsity structure is not preserved.

We consider the GSPP of the following form:

$$\mathcal{M}_0\boldsymbol{v} \triangleq \begin{bmatrix} A & B^H \\ B & D \end{bmatrix}\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} q \\ r \end{bmatrix} \triangleq \boldsymbol{f}, \tag{4.2.1}$$

where $A \in \mathbb{C}^{n\times n}, B, C \in \mathbb{C}^{m\times n}, D \in \mathbb{C}^{m\times m}$, $q \in \mathbb{C}^n$ and $r \in \mathbb{C}^m$. We investigate the structured BE for the GSPP (4.2.1) when $A \in \mathbb{C}^{n\times n}$ is Hermitian, $B \in \mathbb{C}^{m\times n}$, and $D \in \mathbb{C}^{m\times m}$ by preserving the the sparsity of $\mathcal{M}_0$.

4.2.1. Basic Definitions and Lemmas

We use $\mathcal{S}_n$, $\mathbb{SKR}^{n\times n}$ and $\mathbb{HC}^{n\times n}$ represent the set of all $n\times n$ real symmetric matrices, real skew-symmetric matrices and Hermitian matrices, respectively. For $X \in \mathbb{C}^{n\times m}$, $\mathfrak{R}(X)$ and $\mathfrak{I}(X)$ represent the real part and imaginary part of X, respectively. Given a positive weight vector $\boldsymbol{\sigma} = [\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2]^T$. Then, the corresponding weighted Frobenius norms are defined as follows:

$$\left|\left|\left|\begin{bmatrix}\Delta\mathcal{M}_0 & \Delta\boldsymbol{f}\end{bmatrix}\right|\right|\right|_{\boldsymbol{\sigma},F} = \left\|\begin{bmatrix}\alpha_1\|\Delta A\|_F, & \alpha_2\|\Delta B\|_F, & \alpha_3\|\Delta D\|_F, & \beta_1\|\Delta q\|_2, & \beta_2\|\Delta r\|_2\end{bmatrix}\right\|_2.$$

Following the next definition, we introduce the concept of structured BE for the GSPP (4.2.1). Throughout the section, we assume that the coefficient matrix $\mathcal{M}_0$ in (4.2.1) is nonsingular.

Definition 4.2.1. *Assume that $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ is a computed solution of the GSPP (4.2.1). Then, the normwise structured BEs $\boldsymbol{\eta}^{\mathcal{G}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$, is defined as follows:*

$$\boldsymbol{\eta}^{\mathcal{G}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \min_{\left(\begin{array}{c}\Delta A, \Delta B,\\ \Delta D, \Delta q, \Delta r\end{array}\right) \in \mathcal{G}} \left\| \left[\Delta \mathcal{M}_0 \quad \Delta \boldsymbol{f}\right] \right\|_{\boldsymbol{\sigma},F},$$

where

$$\mathcal{G} = \left\{ \left(\begin{array}{c}\Delta A, \Delta B,\\ \Delta D, \Delta q, \Delta r\end{array}\right) \Big| \begin{bmatrix} A+\Delta A & (B+\Delta B)^H \\ B+\Delta B & D+\Delta D \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{u}} \\ \widetilde{\boldsymbol{p}} \end{bmatrix} = \begin{bmatrix} q+\Delta q \\ r+\Delta r \end{bmatrix}, \right.$$
$$\left. \Delta A \in \mathbb{HC}^{n\times n}, \Delta B \in \mathbb{C}^{m\times n}, \Delta D \in \mathbb{C}^{m\times m}, \Delta q \in \mathbb{C}^n, \Delta r \in \mathbb{C}^m \right\}. \tag{4.2.2}$$

By choosing $\alpha_1 = \frac{1}{\|A\|_F}$, $\alpha_2 = \frac{1}{\|B\|_F}$, $\alpha_3 = \frac{1}{\|D\|_F}$, $\beta_1 = \frac{1}{\|q\|_2}$, and $\beta_2 = \frac{1}{\|r\|_2}$, we obtain relative structured BEs for the GSPP (4.2.1).

Remark 4.2.1. *The minimal backward perturbations for the structured BEs are denoted by $\widehat{\Delta A}_{\mathtt{sps}}$, $\widehat{\Delta B}_{\mathtt{sps}}$, $\widehat{\Delta D}_{\mathtt{sps}}$, $\widehat{\Delta q}_{\mathtt{sps}}$, $\widehat{\Delta r}_{\mathtt{sps}}$. Therefore, the following holds:*

$$\boldsymbol{\eta}^{\mathcal{G}_i}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \boldsymbol{\eta}^{\sigma_1}(\widehat{\Delta A}_{\mathtt{sps}}, \widehat{\Delta B}_{\mathtt{sps}}, \widehat{\Delta D}_{\mathtt{sps}}, \widehat{\Delta q}_{\mathtt{sps}}, \widehat{\Delta r}_{\mathtt{sps}}).$$

The structured BE by preserving sparsity pattern is denoted as $\boldsymbol{\eta}^{\mathcal{G}}_{\mathtt{sps}}(\boldsymbol{u}, \boldsymbol{p})$.

Next, we define $\Theta_{\mathcal{M}_0} := \begin{bmatrix} \Theta_A & \Theta_B^T \\ \Theta_B & \Theta_D \end{bmatrix}$ and discuss some important definitions and lemmas.

Definition 4.2.2. *Let $Z \in \mathcal{S}_m$, then we define its generator vector by*

$$\mathrm{vec}_S(Z) := [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \ldots, \boldsymbol{z}_m^T]^T \in \mathbb{R}^{\frac{m(m+1)}{2}},$$

where $\boldsymbol{z}_1 = [z_{11}, z_{21}, \ldots, z_{m1}]^T \in \mathbb{R}^m$, $\boldsymbol{z}_2 = [z_{22}, z_{32}, \ldots, z_{m2}]^T \in \mathbb{R}^{m-1}, \ldots, \boldsymbol{z}_{m-1} = [z_{(m-1)(m-1)}, z_{m(m-1)}]^T \in \mathbb{R}^2$, and $\boldsymbol{z}_m = [z_{mm}] \in \mathbb{R}$.

Definition 4.2.3. *Let $Z \in \mathbb{SKR}^{m\times m}$, then we define its generator vector by*

$$\mathrm{vec}_{SK}(Z) := [\boldsymbol{z}_1^T, \boldsymbol{z}_2^T, \ldots, \boldsymbol{z}_{m-1}^T]^T \in \mathbb{R}^{\frac{m(m-1)}{2}},$$

where $\boldsymbol{z}_1 = [z_{21}, \ldots, z_{m1}]^T \in \mathbb{R}^{m-1}$, $\boldsymbol{z}_2 = [z_{32}, \ldots, z_{m2}]^T \in \mathbb{R}^{m-2}$, $\ldots$, and $\boldsymbol{z}_{m-1} = [z_{(m-1)(m-1)}]^T \in \mathbb{R}$.

Lemma 4.2.2. *Let $M \in \mathcal{S}_m$. Then* $\mathrm{vec}(M) = \mathcal{J}_S^m \mathrm{vec}_S(M)$, *where*

$$\mathcal{J}_S^m = \begin{bmatrix} \mathcal{J}_S^{(1)} & \mathcal{J}_S^{(2)} & \cdots & \mathcal{J}_S^{(m)} \end{bmatrix} \in \mathbb{R}^{m^2 \times \frac{m(m+1)}{2}}$$

and $\mathcal{J}_S^{(i)} \in \mathbb{R}^{m \times (m-i+1)}$ are defined by

$$\mathcal{J}_S^{(1)} = \begin{bmatrix} e_1^m & e_2^m & e_3^m & \cdots & e_{m-1}^m & e_m^m \\ \mathbf{0} & e_1^m & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & e_1^m & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & e_1^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & e_1^m \end{bmatrix}, \mathcal{J}_S^{(2)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ e_2^m & e_3^m & \cdots & \cdots & e_m^m \\ \mathbf{0} & e_2^m & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & e_2^m & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & e_2^m \end{bmatrix}, \ldots, \mathcal{J}_S^{(m)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ e_m^m \end{bmatrix}.$$

Lemma 4.2.3. *Let $M \in \mathbb{SKR}^{m \times m}$. Then* $\mathrm{vec}(M) = \mathcal{J}_{SK}^m \mathrm{vec}_{SK}(M)$, *where*

$$\mathcal{J}_{SK}^m = \begin{bmatrix} \mathcal{J}_{SK}^{(1)} & \mathcal{J}_{SK}^{(2)} & \cdots \mathcal{J}_{SK}^{(m-1)} \end{bmatrix} \in \mathbb{R}^{m^2 \times \frac{m(m-1)}{2}}$$

and $\mathcal{J}_{SK}^{(i)} \in \mathbb{R}^{m \times (m-i)}$ are defined by

$$\mathcal{J}_{SK}^{(1)} = \begin{bmatrix} e_2^m & e_3^m & \cdots & e_{m-1}^m & e_m^m \\ -e_1^m & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & -e_1^m & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & -e_1^m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & -e_1^m \end{bmatrix}, \mathcal{J}_{SK}^{(2)} = \begin{bmatrix} \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ e_3^m & \cdots & \cdots & e_m^m \\ -e_2^m & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & -e_2^m & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \cdots & \cdots & \mathbf{0} & -e_2^m \end{bmatrix}, \ldots, \mathcal{J}_{SK}^{(m-1)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ e_m^m \\ -e_{m-1}^m \end{bmatrix}.$$

Next, we introduce a key lemma that is essential for computing structured BE while preserving sparsity.

Lemma 4.2.4. *Assume $M \in \mathbb{R}^{m \times m}$ and $X \in \mathbb{HC}^{n \times n}$. Then, the following holds:*

1. *When $M \in \mathcal{S}_m$. Then, we have*

$$\mathrm{vec}(M \odot \Theta_X) = \mathcal{J}_S^m \Phi_X \mathrm{vec}_S(M \odot \Theta_X), \tag{4.2.3}$$

   where $\Phi_X = \mathrm{diag}(\mathrm{vec}_S(\Theta_X))$.

2. *When $M \in \mathbb{SKR}^{m \times m}$. Then, we have*

$$\mathrm{vec}(M \odot \Theta_X) = \mathcal{J}_{SK}^m \Psi_X \mathrm{vec}_{SK}(M \odot \Theta_X), \tag{4.2.4}$$

   where $\Psi_X = \mathrm{diag}([\Theta_X(1, 2:n), \Theta_X(2, 3:n), \ldots, \Theta_X(n-1:n)]^T)$.

Proof. Let $M \in \mathcal{S}_m$ and $X \in \mathbb{HC}^{n\times n}$. By definition of the matrix Θ_X, we have $\Theta_X \in \mathcal{S}_m$ and consequently, $M \odot \Theta_X \in \mathcal{S}_m$. Then

$$\begin{aligned}\mathrm{vec}(M \odot \Theta_X) &= \mathcal{J}_S^m \mathrm{vec}_S(M \odot \Theta_X)\\ &= \mathcal{J}_S^m \Phi_X \mathrm{vec}_S(M \odot \Theta_X).\end{aligned}$$

Hence, (4.2.3) follows. Similarly, we can prove (4.2.4) when $M \in \mathbb{SKR}^{m\times m}$. ■

Remark 4.2.5. *When $M, X \in \mathbb{C}^{m\times n}$, we have* $\mathrm{vec}(M \odot \Theta_X) = \Sigma_X \mathrm{vec}(M \odot \Theta_X)$, *where* $\Sigma_X = \mathrm{diag}(\mathrm{vec}(\Theta_X))$.

4.2.2. Computation of Structured BEs

In this subsection, we compute the closed-form expressions for the structured BEs $\boldsymbol{\eta}_{\mathtt{sps}}^{\mathcal{G}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ by preserving the sparsity of the coefficient matrix and the Hermitian structure of the matrix A. The following lemma plays a crucial role in the computation of the structured BE.

Lemma 4.2.6. *Consider the following GSPP:*

$$\begin{bmatrix} A+\Delta A & (B+\Delta B)^H \\ B+\Delta B & D+\Delta D \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} q+\Delta q \\ r+\Delta r \end{bmatrix}. \tag{4.2.5}$$

Then (4.2.5) can be reformulated as the following system of linear equations:

$$\begin{cases} \Re(\Delta A)\Re(\boldsymbol{u}) - \Im(\Delta A)\Im(\boldsymbol{u}) + \Re(\Delta B)^T\Re(\boldsymbol{p}) + \Im(\Delta B)^T\Im(\boldsymbol{p}) - \Re(\Delta q) = \Re(Q), \\ \Re(\Delta A)\Im(\boldsymbol{u}) + \Im(\Delta A)\Re(\boldsymbol{u}) + \Re(\Delta B)^T\Im(\boldsymbol{p}) - \Im(\Delta B)^T\Re(\boldsymbol{p}) - \Im(\Delta q) = \Im(Q), \end{cases} \tag{4.2.6}$$

and

$$\begin{cases} \Re(\Delta B)\Re(\boldsymbol{u}) - \Im(\Delta B)\Im(\boldsymbol{u}) + \Re(\Delta D)\Re(\boldsymbol{p}) - \Im(\Delta D)\Im(\boldsymbol{p}) - \Re(\Delta r) = \Re(R), \\ \Re(\Delta B)\Im(\boldsymbol{u}) + \Im(\Delta B)\Re(\boldsymbol{u}) + \Re(\Delta D)\Im(\boldsymbol{p}) + \Im(\Delta D)\Re(\boldsymbol{p}) - \Im(\Delta r) = \Im(R), \end{cases} \tag{4.2.7}$$

where

$$Q = q - A\boldsymbol{u} - B^H\boldsymbol{p} \quad \textit{and} \quad R = r - B\boldsymbol{u} - D\boldsymbol{p}. \tag{4.2.8}$$

Proof. The GSPP (4.2.5) can be equivalently rewritten as follows:

$$\Delta A\boldsymbol{u} + \Delta B^H\boldsymbol{p} - \Delta q = q - A\boldsymbol{u} - B^H\boldsymbol{p}, \tag{4.2.9}$$

$$\Delta B\boldsymbol{u} + \Delta D\boldsymbol{p} - \Delta r = r - B\boldsymbol{u} - D\boldsymbol{p}. \tag{4.2.10}$$

Now separating the real and imaginary parts from (4.2.9) and (4.2.10), the proof follows. ■

Prior to stating the main theorem of this section, we construct the following matrices: Let $\mathfrak{D}_{\mathcal{S}_n} \in \mathbb{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ and $\mathfrak{D}_{SK_n} \in \mathbb{R}^{\frac{n(n-1)}{2}\times\frac{n(n-1)}{2}}$ are the diagonal matrices with

$$\mathfrak{D}_{\mathcal{S}_n}(j,j) := \begin{cases} 1, & \text{for } \ j = \frac{(2n-(i-2))(i-1)}{2} + 1,\ i = 1,2,\ldots,n, \\ \sqrt{2}, & \text{otherwise,} \end{cases}$$

and

$$\mathfrak{D}_{SK_n}(j,j) := \sqrt{2}, \ \text{ for } j = 1,2,\ldots,\frac{n(n-1)}{2}.$$

Further, set

$$N_1 := \mathcal{J}_S^n \Phi_A \mathfrak{D}_{\mathcal{S}_n}^{-1} \in \mathbb{R}^{n^2\times\frac{n(n+1)}{2}} \ \text{ and } \ N_2 := \mathcal{J}_{SK}^n \Psi_A \mathfrak{D}_{SK_n}^{-1} \in \mathbb{R}^{n^2\times\frac{n(n-1)}{2}}. \tag{4.2.11}$$

Let $\boldsymbol{s} = n^2 + 2mn + 2m^2$, $\mathbf{X}_1 \in \mathbb{R}^{2n\times\boldsymbol{s}}$ and $\mathbf{X}_2 \in \mathbb{R}^{2m\times\boldsymbol{s}}$ be defined as follows:

$$\mathbf{X}_1 := [\widehat{\mathbf{X}}_1 \ \ \mathbf{0}_{2n\times 2m^2}] \ \text{ and } \ \mathbf{X}_2 := [\mathbf{0}_{2m\times n^2} \ \ \widehat{\mathbf{X}}_2],$$

where

$$\widehat{\mathbf{X}}_1 = \begin{bmatrix} \alpha_1^{-1}(\mathfrak{R}(\widetilde{\boldsymbol{u}})^T \otimes I_n)N_1 & -\alpha_1^{-1}(\mathfrak{I}(\widetilde{\boldsymbol{u}})^T \otimes I_n)N_2 & \alpha_2^{-1}(I_n \otimes \mathfrak{R}(\widetilde{\boldsymbol{p}})^T)\Sigma_B & \alpha_2^{-1}(I_n \otimes \mathfrak{I}(\widetilde{\boldsymbol{p}})^T)\Sigma_B \\ \alpha_1^{-1}(\mathfrak{I}(\widetilde{\boldsymbol{u}})^T \otimes I_n)N_1 & \alpha_1^{-1}(\mathfrak{R}(\widetilde{\boldsymbol{u}})^T \otimes I_n)N_2 & \alpha_2^{-1}(I_n \otimes \mathfrak{I}(\widetilde{\boldsymbol{p}})^T)\Sigma_B & -\alpha_2^{-1}(I_n \otimes \mathfrak{R}(\widetilde{\boldsymbol{p}})^T)\Sigma_B \end{bmatrix}$$

and

$$\widehat{\mathbf{X}}_2 = \begin{bmatrix} \alpha_2^{-1}(\mathfrak{R}(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B & -\alpha_2^{-1}(\mathfrak{I}(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B & \alpha_3^{-1}(\mathfrak{R}(\widetilde{\boldsymbol{p}})^T \otimes I_m)\Sigma_D & -\alpha_3^{-1}(\mathfrak{I}(\widetilde{\boldsymbol{p}})^T \otimes I_m)\Sigma_D \\ \alpha_2^{-1}(\mathfrak{I}(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B & \alpha_2^{-1}(\mathfrak{R}(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B & \alpha_3^{-1}(\mathfrak{I}(\widetilde{\boldsymbol{p}})^T \otimes I_m)\Sigma_D & \alpha_3^{-1}(\mathfrak{R}(\widetilde{\boldsymbol{p}})^T \otimes I_m)\Sigma_D \end{bmatrix}.$$

Set

$$\Delta\mathcal{Y} := \begin{bmatrix} \alpha_1\mathfrak{D}_{S,n}\mathrm{vec}_S(\mathfrak{R}(\Delta A \odot \Theta_A)) \\ \alpha_1\mathfrak{D}_{SK_n}\mathrm{vec}_{SK}(\mathfrak{I}(\Delta A \odot \Theta_A)) \\ \alpha_2\mathrm{vec}(\mathfrak{R}(\Delta B \odot \Theta_B)) \\ \alpha_2\mathrm{vec}(\mathfrak{I}(\Delta B \odot \Theta_B)) \\ \alpha_3\mathrm{vec}(\mathfrak{R}(\Delta D \odot \Theta_D)) \\ \alpha_3\mathrm{vec}(\mathfrak{I}(\Delta D \odot \Theta_D)) \end{bmatrix} \in \mathbb{R}^{\boldsymbol{s}} \ \text{ and } \ \Delta\mathcal{Z} := \begin{bmatrix} \beta_1\mathfrak{R}(\Delta q) \\ \beta_1\mathfrak{I}(\Delta q) \\ \beta_2\mathfrak{R}(\Delta r) \\ \beta_2\mathfrak{I}(\Delta r) \end{bmatrix} \in \mathbb{R}^{2(n+m)}. \tag{4.2.12}$$

Note that

$$\begin{aligned} \alpha_1^2\|A\|_F^2 &= \alpha_1^2\|\mathfrak{R}(A)\|_F^2 + \alpha_1^2\|\mathfrak{I}(A)\|_F^2 \\ &= \|\alpha_1\mathfrak{D}_{\mathcal{S}_n}\mathrm{vec}_S(\mathfrak{R}(\Delta A))\|_2^2 + \|\alpha_1\mathfrak{D}_{SK_n}\mathrm{vec}_{SK}(\mathfrak{I}(\Delta A))\|_2^2 \\ &= \left\| \begin{bmatrix} \alpha_1\mathfrak{D}_{\mathcal{S}_n}\mathrm{vec}_S(\mathfrak{R}(\Delta A)) \\ \alpha_1\mathfrak{D}_{SK_n}\mathrm{vec}_{SK}(\mathfrak{I}(\Delta A)) \end{bmatrix} \right\|_2^2. \end{aligned} \tag{4.2.13}$$

In the next theorem, we present a compact expression for the structured BE $\boldsymbol{\eta}_{\mathtt{sps}}^{\mathcal{G}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$.

Theorem 4.2.7. *Assume that* $A \in \mathbb{HC}^{n\times n}$, $B \in \mathbb{C}^{m\times n}$, $D \in \mathbb{C}^{m\times m}$ *and* $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ *is a computed solution of the GSPP* (4.2.1). *Then, the structured BE* $\boldsymbol{\eta}^{\mathcal{G}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ *with preserving sparsity is given by*

$$\boldsymbol{\eta}^{\mathcal{G}}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix}^T \left(\begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix}^T \right)^{-1} \begin{bmatrix} \Re(Q) \\ \Im(Q) \\ \Re(R) \\ \Im(R) \end{bmatrix} \right\|_2, \tag{4.2.14}$$

where $Q = q - A\boldsymbol{u} - B^H\boldsymbol{p}$, $R = r - B\boldsymbol{u} - D\boldsymbol{p}$, $\mathcal{I}_1 = \begin{bmatrix} -\beta_1^{-1} I_{2n} & 0_{2n\times 2m} \end{bmatrix} \in \mathbb{R}^{2n\times 2(n+m)}$ *and* $\mathcal{I}_2 = \begin{bmatrix} 0_{2m\times 2n} & -\beta_2^{-1} I_{2m} \end{bmatrix} \in \mathbb{R}^{2m\times 2(n+m)}$.

Proof. Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be a computed solution of the GSPP (4.2.1) with $A \in \mathbb{HC}^{n\times n}$. Then, we need to find the perturbations $\Delta q \in \mathbb{C}^n$, $\Delta r \in \mathbb{C}^m$ and sparsity preserving perturbation matrices $\Delta A \in \mathbb{HC}^{n\times n}$, $\Delta B \in \mathbb{C}^{m\times n}$ and $\Delta D \in \mathbb{C}^{m\times m}$ so that (4.2.2) holds. Therefore, we replace ΔA, ΔB and ΔD with $\Delta A \odot \Theta_A$, $\Delta B \odot \Theta_B$ and $\Delta D \odot \Theta_D$, respectively, such the following holds:

$$\begin{bmatrix} A + (\Delta A \odot \Theta_A) & (B + (\Delta B \odot \Theta_B))^H \\ B + (\Delta B \odot \Theta_B) & D + (\Delta D \odot \Theta_D) \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{u}} \\ \widetilde{\boldsymbol{p}} \end{bmatrix} = \begin{bmatrix} q + \Delta q \\ r + \Delta r \end{bmatrix}. \tag{4.2.15}$$

Then using Lemma 4.2.6, we have

$$\begin{cases} \Re(\Delta A \odot \Theta_A)\Re(\widetilde{\boldsymbol{u}}) - \Im(\Delta A \odot \Theta_A)\Im(\widetilde{\boldsymbol{u}}) + \Re(\Delta B \odot \Theta_B)^T\Re(\widetilde{\boldsymbol{p}}) \\ \qquad + \Im(\Delta B \odot \Theta_B)^T\Im(\widetilde{\boldsymbol{p}}) - \Re(\Delta q) = \Re(Q), \\ \Re(\Delta A \odot \Theta_A)\Im(\widetilde{\boldsymbol{u}}) + \Im(\Delta A \odot \Theta_A)\Re(\widetilde{\boldsymbol{u}}) + \Re(\Delta B \odot \Theta_B)^T\Im(\widetilde{\boldsymbol{p}}) \\ \qquad - \Im(\Delta B \odot \Theta_B)^T\Re(\widetilde{\boldsymbol{p}}) - \Im(\Delta q) = \Im(Q), \end{cases} \tag{4.2.16}$$

and

$$\begin{cases} \Re(\Delta B \odot \Theta_B)\Re(\widetilde{\boldsymbol{u}}) - \Im(\Delta B \odot \Theta_B)\Im(\widetilde{\boldsymbol{u}}) + \Re(\Delta D \odot \Theta_D)\Re(\widetilde{\boldsymbol{p}}) \\ \qquad - \Im(\Delta D \odot \Theta_D)\Im(\widetilde{\boldsymbol{p}}) - \Re(\Delta r) = \Re(R), \\ \Re(\Delta B \odot \Theta_B)\Im(\widetilde{\boldsymbol{u}}) + \Im(\Delta B \odot \Theta_B)\Re(\boldsymbol{u}) + \Re(\Delta D \odot \Theta_D)\Im(\widetilde{\boldsymbol{p}}) \\ \qquad + \Im(\Delta D \odot \Theta_D)\Re(\widetilde{\boldsymbol{p}}) - \Im(\Delta r) = \Im(R). \end{cases} \tag{4.2.17}$$

Now, using the properties of the vec operator and Kronecker product on (4.2.16), we obtain

$$\begin{cases} (\Re(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathrm{vec}(\Re(\Delta A \odot \Theta_A)) - (\Im(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathrm{vec}(\Im(\Delta A \odot \Theta_A)) \\ +(I_n \otimes \Re(\widetilde{\boldsymbol{p}})^T)\mathrm{vec}(\Re(\Delta B \odot \Theta_B)) - (I_n \otimes \Im(\widetilde{\boldsymbol{p}})^T)\mathrm{vec}(\Im(\Delta B \odot \Theta_B)) - \Re(\Delta q) = \Re(Q), \\ (\Im(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathrm{vec}(\Re(\Delta A \odot \Theta_A)) + (\Re(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathrm{vec}(\Im(\Delta A \odot \Theta_A) \\ +(I_n \otimes \Im(\widetilde{\boldsymbol{p}})^T)\mathrm{vec}(\Re(\Delta B \odot \Theta_B)) - (I_n \otimes \Re(\widetilde{\boldsymbol{p}})^T)\mathrm{vec}(\Im(\Delta B \odot \Theta_B)) - \Im(\Delta q) = \Im(Q). \end{cases} \tag{4.2.18}$$

As $\Delta A \in \mathbb{HC}^{n\times n}$, we have $\Re(\Delta A) \in \mathcal{S}_n$ and $\Im(\Delta A) \in \mathbb{SKR}^{n\times n}$. Further, we have $\Delta A \odot \Theta_A \in \mathbb{HC}^{n\times n}$, $\Re(\Delta A \odot \Theta_A) \in \mathcal{S}_n$ and $\Im(\Delta A \odot \Theta_A) \in \mathbb{SKR}^{n\times n}$. Hence, using Lemma 4.2.4 on (4.2.18), we get

$$\begin{cases} (\Re(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_S^n \Phi_A \mathrm{vec}_S(\Re(\Delta A \odot \Theta_A)) - (\Im(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_{SK}^n \Psi_A \mathrm{vec}_{SK}(\Im(\Delta A \odot \Theta_A)) \\ +(I_n \otimes \Re(\widetilde{\boldsymbol{p}})^T)\Sigma_B \mathrm{vec}(\Re(\Delta B \odot \Theta_B)) - (I_n \otimes \Im(\widetilde{\boldsymbol{p}})^T)\Sigma_B \mathrm{vec}(\Im(\Delta B \odot \Theta_B)) - \Re(\Delta q) = \Re(Q), \\ (\Im(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_S^n \Phi_A \mathrm{vec}_S(\Re(\Delta A \odot \Theta_A)) + (\Re(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_{SK}^n \Psi_A \mathrm{vec}_{SK}(\Im(\Delta A \odot \Theta_A) \\ +(I_n \otimes \Im(\widetilde{\boldsymbol{p}})^T)\Sigma_B \mathrm{vec}(\Re(\Delta B \odot \Theta_B)) - (I_n \otimes \Re(\widetilde{\boldsymbol{p}})^T)\Sigma_B \mathrm{vec}(\Im(\Delta B \odot \Theta_B)) - \Im(\Delta q) = \Im(Q). \end{cases} \tag{4.2.19}$$

Then, (4.2.19) can be reformulate as

$$\mathbf{X}_1 \Delta\mathcal{Y} + \begin{bmatrix} -\alpha_4^{-1} I_{2n} & 0_{2m} \end{bmatrix} \Delta\mathcal{Z} = \begin{bmatrix} \Re(Q) \\ \Im(Q) \end{bmatrix}. \tag{4.2.20}$$

In a similar manner, from (4.2.17), we can deduce the following:

$$\mathbf{X}_2 \Delta\mathcal{Y} + \begin{bmatrix} 0_{2n} & -\alpha_5^{-1} I_{2m} \end{bmatrix} \Delta\mathcal{Z} = \begin{bmatrix} \Re(R) \\ \Im(R) \end{bmatrix}. \tag{4.2.21}$$

Therefore, combining (4.2.20) and (4.2.21), we get

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \Delta\mathcal{Y} + \begin{bmatrix} \mathcal{I}_1 \\ \mathcal{I}_2 \end{bmatrix} \Delta\mathcal{Z} = \begin{bmatrix} \Re(Q) \\ \Im(Q) \\ \Re(R) \\ \Im(R) \end{bmatrix} \Longleftrightarrow \begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix} \begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix} = \begin{bmatrix} \Re(Q) \\ \Im(Q) \\ \Re(R) \\ \Im(R) \end{bmatrix}. \tag{4.2.22}$$

Since, the matrix $\begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix}$ has full row rank, then

$$\begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix}^T \left(\begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix}^T \right)^{-1}$$

and the consistency condition in Lemma 1.3.1 is fulfilled and the minimal norm solution of (4.2.22) is given by

$$\begin{bmatrix}\Delta\mathcal{Y}\\ \Delta\mathcal{Z}\end{bmatrix}_{\min} = \begin{bmatrix}\mathbf{X}_1 & \mathcal{I}_1\\ \mathbf{X}_2 & \mathcal{I}_2\end{bmatrix}^T \left(\begin{bmatrix}\mathbf{X}_1 & \mathcal{I}_1\\ \mathbf{X}_2 & \mathcal{I}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1 & \mathcal{I}_1\\ \mathbf{X}_2 & \mathcal{I}_2\end{bmatrix}^T\right)^{-1} \begin{bmatrix}\Re(Q)\\ \Im(Q)\\ \Re(R)\\ \Im(R)\end{bmatrix}.$$

According to the Definition 4.2.1, we have

$$\begin{aligned}\boldsymbol{\eta}^{\mathcal{G}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{p}}) &= \min_{\begin{pmatrix}\Delta A\odot A, \Delta B\odot B,\\ \Delta D\odot D, \Delta q, \Delta r\end{pmatrix}\in\mathcal{G}_1} \left|\!\left|\!\left|\begin{bmatrix}\Delta\mathcal{M}_0\odot\Theta_{\mathcal{M}_0} & \Delta d\end{bmatrix}\right|\!\right|\!\right|_{\boldsymbol{\sigma}_1,F}\\ &= \min_{\begin{pmatrix}\Delta A\odot A, \Delta B\odot B,\\ \Delta D\odot \Theta_D, \Delta q, \Delta r\end{pmatrix}\in\mathcal{G}_1} \left\|\begin{bmatrix}\alpha_1\mathrm{vec}(\Re(\Delta A\odot\Theta_A))\\ \alpha_1\mathrm{vec}(\Im(\Delta A\odot\Theta_A))\\ \alpha_2\mathrm{vec}(\Re(\Delta B\odot\Theta_B))\\ \alpha_2\mathrm{vec}(\Im(\Delta B\odot\Theta_B))\\ \alpha_3\mathrm{vec}(\Re(\Delta D\odot\Theta_D))\\ \alpha_3\mathrm{vec}(\Im(\Delta D\odot\Theta_D))\\ \beta_1\Re(\Delta q)\\ \beta_1\Im(\Delta q)\\ \beta_2\Re(\Delta r)\\ \beta_2\Im(\Delta r)\end{bmatrix}\right\|_2\\ &= \min_{\begin{pmatrix}\Delta A\odot A, \Delta B\odot B,\\ \Delta D\odot \Theta_D, \Delta q, \Delta r\end{pmatrix}\in\mathcal{G}_1} \left\|\begin{bmatrix}\alpha_1\mathfrak{D}_{\mathcal{S}_n}\mathrm{vec}_S(\Re(\Delta A\odot\Theta_A))\\ \alpha_1\mathfrak{D}_{SK_n}\mathrm{vec}_{SK}(\Im(\Delta A\odot\Theta_A))\\ \alpha_2\mathrm{vec}(\Re(\Delta B\odot\Theta_B))\\ \alpha_2\mathrm{vec}(\Im(\Delta B\odot\Theta_B))\\ \alpha_3\mathrm{vec}(\Re(\Delta D\odot\Theta_D))\\ \alpha_3\mathrm{vec}(\Im(\Delta D\odot\Theta_D))\\ \beta_1\Re(\Delta q)\\ \beta_1\Im(\Delta q)\\ \beta_2\Re(\Delta r)\\ \beta_2\Im(\Delta r)\end{bmatrix}\right\|_2\end{aligned}$$

$$= \min \left\{ \left\| \begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix} \right\|_2 \middle| \begin{bmatrix} \mathbf{X}_1 & \mathcal{I}_1 \\ \mathbf{X}_2 & \mathcal{I}_2 \end{bmatrix} \begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix} = \begin{bmatrix} \Re(Q) \\ \Im(Q) \\ \Re(R) \\ \Im(R) \end{bmatrix} \right\}$$

$$= \left\| \begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix}_{\min} \right\|_2 . \tag{4.2.23}$$

Hence, the proof is completed. ■

Remark 4.2.8. *Using* (4.2.12) *and* (4.2.23)*, the minimal perturbation matrix* $\widehat{\Delta A}_{\text{sps}}$ *is given by*

$$\widehat{\Delta A}_{\text{sps}} = \Re(\widehat{\Delta A}_{\text{sps}}) + i\Im(\widehat{\Delta A}_{\text{sps}}),$$

where

$$\text{vec}_S(\Re(\widehat{\Delta A}_{\text{sps}})) = \alpha_1^{-1}\mathfrak{D}_{\mathcal{S}_n}^{-1}\begin{bmatrix} I_{\boldsymbol{n}_1} & 0_{\boldsymbol{n}_1\times(\boldsymbol{l}-\boldsymbol{n}_1)} \end{bmatrix}\begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix}_{\text{sps}},$$

$$\text{vec}_{SK}(\Im(\widehat{\Delta A}_{\text{sps}})) = \alpha_1^{-1}\mathfrak{D}_{SK_n}^{-1}\begin{bmatrix} 0_{\boldsymbol{n}_2\times\boldsymbol{n}_1} & I_{\boldsymbol{n}_2} & 0_{\boldsymbol{n}_2\times(\boldsymbol{l}-n^2)} \end{bmatrix}\begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix}_{\text{sps}},$$

$\boldsymbol{n}_1 = \frac{n(n+1)}{2}$, $\boldsymbol{n}_2 = \frac{n(n-1)}{2}$, *and* $\boldsymbol{l} = \boldsymbol{s} + 2(m+n)$. *Similarly, minimal perturbation matrix* $\widehat{\Delta B}_{\text{sps}}$, $\widehat{\Delta D}_{\text{sps}}$, $\widehat{\Delta r}_{\text{sps}}$ *and* $\widehat{\Delta q}_{\text{sps}}$ *are given by*

$$\text{vec}(\widehat{\Delta B}_{\text{sps}}) = \alpha_2^{-1}\begin{bmatrix} 0_{nm\times n^2} & I_{nm} & \boldsymbol{i}I_{nm} & 0_{nm\times(\boldsymbol{l}-n^2-2nm)} \end{bmatrix}\begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix}_{\text{sps}},$$

$$\text{vec}(\widehat{\Delta D}_{\text{sps}}) = \alpha_3^{-1}\begin{bmatrix} 0_{m^2\times(n^2+2nm)} & I_{m^2} & \boldsymbol{i}I_{m^2} & 0_{m^2\times 2(m+n)} \end{bmatrix}\begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix}_{\text{sps}},$$

$$\widehat{\Delta q}_{\text{sps}} = \beta_1^{-1}\begin{bmatrix} 0_{n\times(\boldsymbol{l}-2(n+m))} & I_n & \boldsymbol{i}I_n & 0_{n\times 2m} \end{bmatrix}\begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix}_{\text{sps}},$$

and

$$\widehat{\Delta r}_{\text{sps}} = \beta_2^{-1}\begin{bmatrix} 0_{m\times(\boldsymbol{l}-2m)} & I_m & \boldsymbol{i}I_m \end{bmatrix}\begin{bmatrix} \Delta\mathcal{Y} \\ \Delta\mathcal{Z} \end{bmatrix}_{\text{sps}}.$$

Next, we derive the structured BE when the sparsity structure of the coefficient matrix $\mathcal{M}_0$ is not preserved.

Corollary 4.2.1. *Assume that $A \in \mathbb{HC}^{n\times n}$, $B \in \mathbb{C}^{m\times n}$ and $D \in \mathbb{C}^{m\times m}$ and $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ is a computed solution of the GSPP (4.2.1). Then, the structured BE $\boldsymbol{\eta}^{\mathcal{G}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ without preserving sparsity is given by*

$$\boldsymbol{\eta}^{\mathcal{G}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \begin{bmatrix} \mathbf{Y}_1 & \mathcal{I}_1 \\ \mathbf{Y}_2 & \mathcal{I}_2 \end{bmatrix}^T \left(\begin{bmatrix} \mathbf{Y}_1 & \mathcal{I}_1 \\ \mathbf{Y}_2 & \mathcal{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 & \mathcal{I}_1 \\ \mathbf{Y}_2 & \mathcal{I}_2 \end{bmatrix}^T \right)^{-1} \begin{bmatrix} \Re(Q) \\ \Im(Q) \\ \Re(R) \\ \Im(R) \end{bmatrix} \right\|_2,$$

where

$$\mathbf{Y}_1 = [\widehat{\mathbf{Y}}_1 \quad \mathbf{0}_{2n\times 2m^2}], \quad \mathbf{Y}_2 = [\mathbf{0}_{2m^2\times n^2} \quad \widehat{\mathbf{Y}}_2],$$

$$\widehat{\mathbf{Y}}_1 = \begin{bmatrix} \alpha_1^{-1}(\Re(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_S^n\mathfrak{D}_{\mathcal{S}_n}^{-1} & -\alpha_1^{-1}(\Im(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_{SK}^n\mathfrak{D}_{SK_n}^{-1} & \alpha_2^{-1}(I_n \otimes \Re(\widetilde{\boldsymbol{p}})^T) & \alpha_2^{-1}(I_n \otimes \Im(\widetilde{\boldsymbol{p}})^T) \\ \alpha_1^{-1}(\Im(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_S^n\mathfrak{D}_{\mathcal{S}_n}^{-1} & \alpha_1^{-1}(\Re(\widetilde{\boldsymbol{u}})^T \otimes I_n)\mathcal{J}_{SK}^n\mathfrak{D}_{SK_n}^{-1} & \alpha_2^{-1}(I_n \otimes \Im(\widetilde{\boldsymbol{p}})^T) & -\alpha_2^{-1}(I_n \otimes \Re(\widetilde{\boldsymbol{p}})^T) \end{bmatrix},$$

and

$$\widehat{\mathbf{Y}}_2 = \begin{bmatrix} \alpha_2^{-1}(\Re(\widetilde{\boldsymbol{u}})^T \otimes I_m) & -\alpha_2^{-1}(\Im(\widetilde{\boldsymbol{u}})^T \otimes I_m) & \alpha_3^{-1}(\Re(\widetilde{\boldsymbol{p}})^T \otimes I_m) & -\alpha_3^{-1}(\Im(\widetilde{\boldsymbol{p}})^T \otimes I_m) \\ \alpha_2^{-1}(\Im(\widetilde{\boldsymbol{u}})^T \otimes I_m) & \alpha_2^{-1}(\Re(\widetilde{\boldsymbol{u}})^T \otimes I_m) & \alpha_3^{-1}(\Im(\widetilde{\boldsymbol{p}})^T \otimes I_m) & \alpha_3^{-1}(\Re(\widetilde{\boldsymbol{p}})^T \otimes I_m) \end{bmatrix}.$$

Proof. As we are not preserving the sparsity structure of A, B and D, by setting $\Theta_A = \mathbf{1}_{n\times n}$, $\Theta_B = \mathbf{1}_{m\times n}$ and $\Theta_D = \mathbf{1}_{m\times m}$, the proof proceeds accordingly. ■

In the following, we derive the structured BE for the GSPP (4.2.1) when $A \in \mathbb{HC}^{n\times n}$ and $D = \mathbf{0}_{m\times m}$.

Corollary 4.2.2. *Assume that $A \in \mathbb{HC}^{n\times n}$, $B \in \mathbb{C}^{m\times n}$ and $D = \mathbf{0}_{m\times m}$ and $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ is a computed solution of the GSPP (4.2.1). Then, the structured BE $\boldsymbol{\eta}_{\mathtt{sps}}^{\mathcal{G}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ with preserving sparsity is given by*

$$\boldsymbol{\eta}_{\mathtt{sps}}^{\mathcal{G}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = \left\| \begin{bmatrix} \widehat{\mathbf{X}}_1 & \mathcal{I}_1 \\ \mathbf{Z} & \mathcal{I}_2 \end{bmatrix}^T \left(\begin{bmatrix} \widehat{\mathbf{X}}_1 & \mathcal{I}_1 \\ \mathbf{Z} & \mathcal{I}_2 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{X}}_1 & \mathcal{I}_1 \\ \mathbf{Z} & \mathcal{I}_2 \end{bmatrix}^T \right)^{-1} \begin{bmatrix} \Re(Q) \\ \Im(Q) \\ \Re(\widehat{R}) \\ \Im(\widehat{R}) \end{bmatrix} \right\|_2, \tag{4.2.24}$$

where$\widehat{R} = r - B\widetilde{\boldsymbol{u}}$, $\mathbf{Z} = [\mathbf{0}_{2m^2\times n^2} \quad \widehat{\mathbf{Z}}]$,

$$\widehat{\mathbf{Z}} = \begin{bmatrix} \alpha_2^{-1}(\Re(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B & -\alpha_2^{-1}(\Im(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B \\ \alpha_2^{-1}(\Im(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B & \alpha_2^{-1}(\Re(\widetilde{\boldsymbol{u}})^T \otimes I_m)\Sigma_B \end{bmatrix}.$$

Proof. The proof follows by taking $\alpha_3 \to 0$ and $D = \mathbf{0}_{m\times m}$ in Theorem 4.2.7. ■

4.3. Structured Backward Errors for Double Saddle Point Problems

In this section, we consider the following general form of the DSPP:

$$\mathfrak{B}\boldsymbol{w} := \begin{bmatrix} A & B^T & \mathbf{0} \\ F & -D & C^T \\ \mathbf{0} & G & E \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} = \begin{bmatrix} f \\ g \\ h \end{bmatrix} =: \mathbf{d}, \tag{4.3.1}$$

where $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$, $E \in \mathbb{R}^{p\times p}$, $B, F \in \mathbb{R}^{m\times n}$, $C, G \in \mathbb{R}^{p\times m}$, $\boldsymbol{x}, f \in \mathbb{R}^n$, $\boldsymbol{y}, g \in \mathbb{R}^m$ and $\boldsymbol{z}, h \in \mathbb{R}^p$.

This section answers the fundamental question raised in Chapter 1: Whether a backward stable algorithm for solving (4.3.1) exhibits strong backward stability or not to the DSPP? The notion of structured BE facilitates us in addressing the aforementioned question, where we study the BE by preserving the inherent structure of the coefficient matrix.

The DSPP of the form (4.3.1) can be converted into a GSPP [11]. Considerable research effort has been devoted to structured BEs for the GSSP in the past years; see [44, 146, 97, 162, 102]. Nevertheless, due to the special block structure of $\mathfrak{B}$, these investigations do not provide exact structured BEs for the DSPP (4.3.1). Recently, Lv and Zheng [96] have investigated the structured BE of (4.3.1), when $A = A^T$, $D = \mathbf{0}$ and $E = \mathbf{0}$. Further, Lv [95] studied the structured BEs of the equivalent form of the DSPP (4.3.1) given by

$$\widehat{\mathfrak{B}}\widehat{\boldsymbol{w}} := \begin{bmatrix} A & \mathbf{0} & B^T \\ \mathbf{0} & E & C \\ -B & -C^T & D \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{z} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} f \\ h \\ -g \end{bmatrix}, \tag{4.3.2}$$

with $B = F$, $C = G$, A and E are non-symmetric, and D is symmetric. When $D = \mathbf{0}$ and $E = \mathbf{0}$, computable expressions for the structured BEs are obtained in [98] in three the cases: first, $A^T = A, B \neq F$ and $C = G$; second, $A^T = A, B = F$ and $C \neq G$; and third, $A^T = A, B \neq F$ and $C \neq G$. However, these studies lack the following investigations: (a) the coefficient matrix $\mathfrak{B}$ in (4.3.1) is generally sparse, and the existing studies do not preserve the sparsity pattern to the perturbation matrices, (b) existing research does not provide explicit formulae for the minimal perturbation matrices that preserve the inherent structures of original matrices for which an approximate solution becomes the exact solution of a nearly perturbed DSPP.

To address the aforementioned drawbacks, in this section, we investigate the structured BEs by preserving the sparsity in three cases: (i) $A^T = A$, $B = F$, $D^T = D$, $C = G$ and $E^T = E$; (ii) $A^T = A$, $B \neq F$, $D^T = D$, $C = G$ and $E^T = E$; (iii) $A^T \neq A$, $B = F$, $D^T = D$, $C \neq G$ and $E^T = E$. The main contributions of this section are as follows:

- We investigate the structured BEs when the perturbation matrices preserve the structures mentioned in the cases (i), (ii) and (iii) as well as preserve the sparsity patterns of the block matrices of the coefficient matrix $\mathfrak{B}$.
- We derive explicit formulae for the minimal perturbation matrices for which the structured BE is attained. These perturbation matrices preserve the inherent structures of the original matrices as well as their sparsity pattern.
- Numerical experiments are performed to validate our theoretical findings, to test the backward stability and strong backward stability of numerical algorithms for solving the DSPPs.

The organization of this section is as follows. In Subsection 4.3.1, we present basic definitions and preliminary results. In Subsections 4.3.2, 4.3.3 and 4.3.4, we derive explicit formulae for the structured BEs corresponding to cases (i), (ii) and (iii), respectively.

4.3.1. Preliminaries

In this subsection, we introduce the definitions of structured BEs for the three cases (i)-(iii). Furthermore, we establish two pivotal lemmas essential for deriving structured BEs. Throughout the section, we assume that $\mathfrak{B}$ is nonsingular. Let $\widetilde{\boldsymbol{w}} = [\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ be an approximate solution of the DSPP (4.3.1). Using the formula (1.3.4), the unstructured BE for the DSPP (4.3.1), denoted by $\boldsymbol{\eta}(\widetilde{\boldsymbol{w}})$, is expressed as:

$$\boldsymbol{\eta}(\widetilde{\boldsymbol{w}}) = \frac{\|\mathbf{d} - \mathfrak{B}\widetilde{\boldsymbol{w}}\|_2}{\sqrt{\|\mathfrak{B}\|_F^2\|\widetilde{\boldsymbol{w}}\|_2^2 + \|\mathbf{d}\|_2^2}}. \tag{4.3.3}$$

Let $\boldsymbol{\alpha} := [\theta_1, \theta_2, \ldots, \theta_{10}]^T$, where θ_i are nonnegative real numbers for $i = 1, 2, \ldots, 10$, with the convention that $\theta_i^{-1} = 0$, whenever $\theta_i = 0$. We define

$$\left|\!\left|\!\left| \begin{bmatrix} \Delta\mathfrak{B} & \Delta\mathbf{d} \end{bmatrix} \right|\!\right|\!\right|_{\boldsymbol{\alpha},F} = \Big\| \Big[\theta_1\|\Delta A\|_F,\ \theta_2\|\Delta B\|_F,\ \theta_3\|\Delta F\|_F,\ \theta_4\|\Delta D\|_F,\ \theta_5\|\Delta C\|_F,$$
$$\theta_6\|\Delta G\|_F,\ \theta_7\|\Delta E\|_F,\ \theta_8\|\Delta f\|_2,\ \theta_9\|\Delta g\|_2,\ \theta_{10}\|\Delta h\|_2\Big] \Big\|_2$$

Next, we define structured BE for an approximate solution of the DSPP (4.3.1).

Definition 4.3.1. *Let $\widetilde{\boldsymbol{w}} = [\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ be an approximate solution of the DSPP (4.3.1). Then, we define the structured BEs, denoted by $\boldsymbol{\eta}^{\mathbb{S}_i}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$, $i = 1, 2, 3$, for the cases (i)-(iii) as follows:*

$$\boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \min_{\begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta f, \\ \Delta g, \Delta h \end{pmatrix} \in \mathbb{S}_1} \left|\!\left|\!\left| \begin{bmatrix} \Delta\mathfrak{B} & \Delta\mathbf{d} \end{bmatrix} \right|\!\right|\!\right|_{\boldsymbol{\alpha}_1, F},$$

$$\boldsymbol{\eta}^{\mathbb{S}_2}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \min_{\begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta F, \\ \Delta f, \Delta g, \Delta h \end{pmatrix} \in \mathbb{S}_2} \left|\!\left|\!\left| \begin{bmatrix} \Delta\mathfrak{B} & \Delta\mathbf{d} \end{bmatrix} \right|\!\right|\!\right|_{\boldsymbol{\alpha}_2, F},$$

and

$$\boldsymbol{\eta}^{\mathbb{S}_3}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \min_{\begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta G, \\ \Delta f, \Delta g, \Delta h \end{pmatrix} \in \mathbb{S}_3} \left|\!\left|\!\left| \begin{bmatrix} \Delta\mathfrak{B} & \Delta\mathbf{d} \end{bmatrix} \right|\!\right|\!\right|_{\boldsymbol{\alpha}_3, F},$$

respectively, where $\boldsymbol{\alpha}_1 := \boldsymbol{\alpha}$ with $\theta_3 = 0$ and $\theta_6 = 0$; $\boldsymbol{\alpha}_2 := \boldsymbol{\alpha}$ with $\theta_6 = 0$; $\boldsymbol{\alpha}_3 := \boldsymbol{\alpha}$ with $\theta_3 = 0$;

$$\mathbb{S}_1 = \left\{ \begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta f, \\ \Delta g, \Delta h \end{pmatrix} \middle| \begin{bmatrix} A + \Delta A & (B + \Delta B)^T & \mathbf{0} \\ B + \Delta B & -(D + \Delta D) & (C + \Delta C)^T \\ \mathbf{0} & C + \Delta C & (E + \Delta E)^T \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{x}} \\ \widetilde{\boldsymbol{y}} \\ \widetilde{\boldsymbol{z}} \end{bmatrix} = \begin{bmatrix} f + \Delta f \\ g + \Delta g \\ h + \Delta h \end{bmatrix}, \right.$$
$$\Delta A \in \mathcal{S}_n, \Delta D \in \mathcal{S}_m, \Delta E \in \mathcal{S}_p, \Delta B \in \mathbb{R}^{m\times n}, \Delta C \in \mathbb{R}^{p\times m},$$
$$\left. \Delta f \in \mathbb{R}^n, \Delta g \in \mathbb{R}^m, \Delta h \in \mathbb{R}^p \right\}. \tag{4.3.4}$$

$$\mathbb{S}_2 = \left\{ \begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta F, \\ \Delta f, \Delta g, \Delta h \end{pmatrix} \middle| \begin{bmatrix} A + \Delta A & (B + \Delta B)^T & \mathbf{0} \\ F + \Delta F & -(D + \Delta D) & (C + \Delta C)^T \\ \mathbf{0} & C + \Delta C & (E + \Delta E)^T \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{x}} \\ \widetilde{\boldsymbol{y}} \\ \widetilde{\boldsymbol{z}} \end{bmatrix} = \begin{bmatrix} f + \Delta f \\ g + \Delta g \\ h + \Delta h \end{bmatrix}, \right.$$
$$\Delta A \in \mathcal{S}_n, \Delta D \in \mathcal{S}_m, \Delta E \in \mathcal{S}_p, \Delta B, \Delta F \in \mathbb{R}^{m\times n}, \Delta C \in \mathbb{R}^{p\times m},$$
$$\left. \Delta f \in \mathbb{R}^n, \Delta g \in \mathbb{R}^m, \Delta h \in \mathbb{R}^p \right\}, \tag{4.3.5}$$

and

$$\mathbb{S}_3 = \left\{ \begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta G, \\ \Delta f, \Delta g, \Delta h \end{pmatrix} \middle| \begin{bmatrix} A+\Delta A & (B+\Delta B)^T & \mathbf{0} \\ B+\Delta B & -(D+\Delta D) & (C+\Delta C)^T \\ \mathbf{0} & G+\Delta G & (E+\Delta E)^T \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{x}} \\ \widetilde{\boldsymbol{y}} \\ \widetilde{\boldsymbol{z}} \end{bmatrix} = \begin{bmatrix} f+\Delta f \\ g+\Delta g \\ h+\Delta h \end{bmatrix}, \right.$$
$$\Delta A \in \mathbb{R}^{n\times n}, \Delta D \in \mathcal{S}_m, \Delta E \in \mathcal{S}_p, \Delta B \in \mathbb{R}^{m\times n}, \Delta C, \Delta G \in \mathbb{R}^{p\times m},$$
$$\left. \Delta f \in \mathbb{R}^n, \Delta g \in \mathbb{R}^m, \Delta h \in \mathbb{R}^p \right\}. \tag{4.3.6}$$

Next, we state the problem of finding structure-preserving minimal perturbation matrices for which the structured BE is attained.

Problem 4.3.1. Find out the minimal perturbation matrices $\widehat{\Delta A}$, $\widehat{\Delta B}$, $\widehat{\Delta C}$, $\widehat{\Delta D}$, $\widehat{\Delta E}$, $\widehat{\Delta F}$, $\widehat{\Delta G}$, $\widehat{\Delta f}$, $\widehat{\Delta g}$, and $\widehat{\Delta h}$ such that

$$\boldsymbol{\eta}^{\mathbb{S}_i}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \left|\!\left|\!\left| \begin{bmatrix} \widehat{\Delta\mathfrak{B}} & \widehat{\Delta\mathbf{d}} \end{bmatrix} \right|\!\right|\!\right|_{\boldsymbol{\alpha}_i, F}, \quad i = 1, 2, 3,$$

where

$$\left|\!\left|\!\left| \begin{bmatrix} \widehat{\Delta\mathfrak{B}} & \widehat{\Delta\mathbf{d}} \end{bmatrix} \right|\!\right|\!\right|_{\boldsymbol{\alpha}, F} = \left\| \begin{bmatrix} \theta_1 \|\widehat{\Delta A}\|_F, \ \theta_2 \|\widehat{\Delta B}\|_F, \ \theta_3 \|\widehat{\Delta F}\|_F, \ \theta_4 \|\widehat{\Delta D}\|_F, \ \theta_5 \|\widehat{\Delta C}\|_F, \\ \theta_6 \|\widehat{\Delta G}\|_F, \ \theta_7 \|\widehat{\Delta E}\|_F, \ \theta_8 \|\widehat{\Delta f}\|_2, \ \theta_9 \|\widehat{\Delta g}\|_2, \ \theta_{10} \|\widehat{\Delta h}\|_2 \end{bmatrix} \right\|_2.$$

Remark 4.3.2. *When $\theta_i = 0$ for any given i ($i = 1, 2, \ldots, 10$), it indicates that the corresponding block matrix has no perturbation.*

Remark 4.3.3. *To preserve sparsity pattern of the block matrices, we substitute the perturbation matrices ΔA, ΔB, ΔC, ΔD, ΔE, ΔF, and ΔG by $\Delta A \odot \Theta_A$, $\Delta B \odot \Theta_B$, $\Delta C \odot \Theta_C$, $\Delta D \odot \Theta_D$, $\Delta E \odot \Theta_E$, $\Delta F \odot \Theta_F$ and $\Delta G \odot \Theta_G$, respectively. Within this framework, we denote the structured BEs as $\boldsymbol{\eta}^{\mathbb{S}_i}_{\boldsymbol{sps}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$, $i = 1, 2, 3$. Moreover, the minimal perturbation matrices are denoted by $\widehat{\Delta A}_{\mathtt{sps}}$, $\widehat{\Delta B}_{\mathtt{sps}}$, $\widehat{\Delta C}_{\mathtt{sps}}$, $\widehat{\Delta D}_{\mathtt{sps}}$, $\widehat{\Delta E}_{\mathtt{sps}}$, $\widehat{\Delta F}_{\mathtt{sps}}$, $\widehat{\Delta G}_{\mathtt{sps}}$, $\widehat{\Delta f}_{\mathtt{sps}}$, $\widehat{\Delta g}_{\mathtt{sps}}$, and $\widehat{\Delta h}_{\mathtt{sps}}$.*

When the structured BEs $\boldsymbol{\eta}^{\mathbb{S}_i}(\widetilde{\boldsymbol{x}}, \widetilde{y}, \widetilde{\boldsymbol{z}})$ and $\boldsymbol{\eta}^{\mathbb{S}_i}_{\mathtt{sps}}(\widetilde{\boldsymbol{x}}, \widetilde{y}, \widetilde{\boldsymbol{z}})$ are around an order of unit round-off error, then the approximate solution $\widetilde{\boldsymbol{w}} = [\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ becomes an actual solution of nearly perturbed structure-preserving DSPP of the form (4.3.1). Thus, the corresponding algorithm is referred to as strongly backward stable. We set $\mu = \frac{n(n+1)}{2}$, $\sigma = \frac{m(m+1)}{2}$ and $\tau = \frac{p(p+1)}{2}$. To obtain the structured BEs formulae, the following lemmas play a pivotal role.

Lemma 4.3.4. *Let $A, H \in \mathcal{S}_n$ with generator vectors $\text{vec}_{\mathcal{S}}(A) = [\boldsymbol{a}_1^T, \boldsymbol{a}_2^T \ldots, \boldsymbol{a}_n^T]^T$ and $\text{vec}_{\mathcal{S}}(H) = [\boldsymbol{h}_1^T, \boldsymbol{h}_2^T \ldots, \boldsymbol{h}_n^T]^T$, respectively. Suppose $x = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, then*

$$(A \odot \Theta_H)x = \mathcal{K}_x \Phi_H \text{vec}_{\mathcal{S}}(A \odot \Theta_H),$$

where $\Phi_H = \text{diag}(\text{vec}_{\mathcal{S}}(\Theta_H))$, $\mathcal{K}_x = \begin{bmatrix} K_x^1 & K_x^2 & \cdots & K_x^n \end{bmatrix} \in \mathbb{R}^{n \times \mu}$ and $K_x^i \in \mathbb{R}^{n \times (n-i+1)}$ are given by

$$K_x^1 = \begin{bmatrix} x_1 & x_2 & \cdots & \cdots & x_n \\ 0 & x_1 & 0 & \cdots & 0 \\ 0 & 0 & x_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & x_1 \end{bmatrix}, \; K_x^2 = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ x_2 & x_3 & \cdots & \cdots & x_n \\ 0 & x_2 & 0 & \cdots & 0 \\ 0 & 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & x_2 \end{bmatrix}, \ldots, \; K_x^n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ x_n \end{bmatrix}.$$

Proof. The proof follows similary to the proof of Lemma 4.1.4. ■

4.3.2. Derivation of Structured BEs for Case (i)

In this subsection, we discuss the structured BEs for the DSPP (4.3.1) for the case (i), i.e., $A \in \mathcal{S}_n, D \in \mathcal{S}_m$, $E \in \mathcal{S}_p$, $B = F$ and $C = G$, and perturbation matrices belongs to set $\mathbb{S}_1$. Prior to that, we construct the diagonal matrix $\mathfrak{D}_{\mathcal{S}_n} \in \mathbb{R}^{\mu \times \mu}$, where

$$\begin{cases} \mathfrak{D}_{\mathcal{S}_n}(k,k) = 1, & \text{for} \quad k = \frac{(2n-(i-2))(i-1)}{2} + 1, \; i = 1, 2, \ldots, n, \\ \mathfrak{D}_{\mathcal{S}_n}(k,k) = \sqrt{2}, & \text{otherwise.} \end{cases}$$

The matrix $\mathfrak{D}_{\mathcal{S}_n}$ has the property, $\|A\|_F = \|\mathfrak{D}_{\mathcal{S}_n} \text{vec}_{\mathcal{S}}(A)\|_2$. Further, we introduce the following notation:

$$\Phi_A = \text{diag}(\text{vec}_{\mathcal{S}}(\Theta_A)), \; \Phi_B = \text{diag}(\text{vec}(\Theta_B)), \; \Phi_C = \text{diag}(\text{vec}(\Theta_C)), \tag{4.3.7}$$

$$\Phi_D = \text{diag}(\text{vec}_{\mathcal{S}}(\Theta_D)), \; \Phi_E = \text{diag}(\text{vec}_{\mathcal{S}}(\Theta_E)), \tag{4.3.8}$$

and

$$\mathcal{I} = \begin{bmatrix} -\frac{1}{\theta_8} I_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\frac{1}{\theta_9} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_{10}} I_p \end{bmatrix}.$$

Theorem 4.3.5. *Let $[\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ be an approximate solution of the DSPP (4.3.1) with $A \in \mathcal{S}_n$, $D \in \mathcal{S}_m$, $E \in \mathcal{S}_p$, and $\theta_8, \theta_9, \theta_{10} \neq 0$. Then, we have*

$$\boldsymbol{\eta}_{\mathbf{sps}}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \left\| \mathcal{J}_{\mathbb{S}_1}^T (\mathcal{J}_{\mathbb{S}_1} \mathcal{J}_{\mathbb{S}_1}^T)^{-1} R_{\mathbf{d}} \right\|_2, \tag{4.3.9}$$

where $\mathcal{J}_{\mathbb{S}_1} = [\widetilde{\mathcal{J}}_{\mathbb{S}_1} \;\; \mathcal{I}] \in \mathbb{R}^{(n+m+p)\times \boldsymbol{l}}$ *and* $\widetilde{\mathcal{J}}_{\mathbb{S}_1}$ *is given by*

$$\widetilde{\mathcal{J}}_{\mathbb{S}_1} = \begin{bmatrix} \frac{1}{\theta_1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\Phi_A\mathfrak{D}_{\mathcal{S}_n}^{-1} & \frac{1}{\theta_2}\mathcal{N}_{\widetilde{\boldsymbol{y}}}^{n}\Phi_B & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\theta_2}M_{\widetilde{\boldsymbol{x}}}^{m}\Phi_B & -\frac{1}{\theta_4}\mathcal{K}_{\widetilde{\boldsymbol{y}}}\Phi_D\mathfrak{D}_{\mathcal{S}_m}^{-1} & \frac{1}{\theta_5}N_{\widetilde{\boldsymbol{z}}}^{m}\Phi_C & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{\theta_5}\mathcal{M}_{\widetilde{\boldsymbol{y}}}^{p}\Phi_C & \frac{1}{\theta_7}\mathcal{K}_{\widetilde{\boldsymbol{z}}}\Phi_E\mathfrak{D}_{\mathcal{S}_p}^{-1} \end{bmatrix},$$

$R_f = f - A\widetilde{\boldsymbol{x}} - B^T\widetilde{\boldsymbol{y}}$, $R_g = g - B\widetilde{\boldsymbol{x}} + D\widetilde{\boldsymbol{y}} - C^T\widetilde{\boldsymbol{z}}$, $R_h = h - C\widetilde{\boldsymbol{y}} - E\widetilde{\boldsymbol{z}}$, $R_{\mathbf{d}} = [R_f^T,\, R_g^T,\, R_h^T]^T$, *and* $\boldsymbol{l} = \mu + \sigma + \tau + mn + mp + m + n + p$.

The minimal perturbation matrices for the Problem 4.3.1 are given by the following generating vectors:

$$\begin{aligned}
\mathrm{vec}_{\mathcal{S}}(\widehat{\Delta A}_{\mathtt{sps}}) &= \theta_1^{-1}\mathfrak{D}_{\mathcal{S}_n}^{-1}\begin{bmatrix} I_\mu & \mathbf{0}\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}},\\
\mathrm{vec}(\widehat{\Delta B}_{\mathtt{sps}}) &= \theta_2^{-1}\begin{bmatrix}\mathbf{0} & I_{mn} & \mathbf{0}\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}},\\
\mathrm{vec}(\widehat{\Delta C}_{\mathtt{sps}}) &= \theta_5^{-1}\begin{bmatrix}\mathbf{0} & I_{mp} & \mathbf{0}\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}},\\
\mathrm{vec}_{\mathcal{S}}(\widehat{\Delta D}_{\mathtt{sps}}) &= \theta_4^{-1}\mathfrak{D}_{\mathcal{S}_m}^{-1}\begin{bmatrix}\mathbf{0} & I_\sigma & \mathbf{0}\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}},\\
\mathrm{vec}_{\mathcal{S}}(\widehat{\Delta E}_{\mathtt{sps}}) &= \theta_7^{-1}\mathfrak{D}_{\mathcal{S}_p}^{-1}\begin{bmatrix}\mathbf{0} & I_\tau & \mathbf{0}\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}},\\
\widehat{\Delta f}_{\mathtt{sps}} &= \theta_8^{-1}\begin{bmatrix}\mathbf{0} & I_n & \mathbf{0}\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}},\\
\widehat{\Delta g}_{\mathtt{sps}} &= \theta_9^{-1}\begin{bmatrix}\mathbf{0} & I_m & \mathbf{0}\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}}, \quad \textit{and}\\
\widehat{\Delta h}_{\mathtt{sps}} &= \theta_{10}^{-1}\begin{bmatrix}\mathbf{0} & I_p\end{bmatrix}\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}}.
\end{aligned}$$

Proof. For the approximate solution $[\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$, we need to construct perturbation matrices $\Delta A \in \mathcal{S}_n$, $\Delta B \in \mathbb{R}^{m\times n}$, $\Delta D \in \mathcal{S}_m$, $\Delta C \in \mathbb{R}^{p\times m}$, $\Delta E \in \mathcal{S}_p$, which maintain the sparsity pattern of A, B, C, D, E, respectively, and the perturbations $\Delta f \in \mathbb{R}^n$, $\Delta g \in \mathbb{R}^m$, and $\Delta h \in \mathbb{R}^p$. By (4.3.4), $\begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta f, \\ \Delta g, \Delta h \end{pmatrix} \in \mathbb{S}_1$ if and only if $\Delta A, \Delta B, \Delta C, \Delta D$, $\Delta E, \Delta f$, Δg and Δh satisfy

$$\left.\begin{aligned}
\Delta A\widetilde{\boldsymbol{x}} + \Delta B^T\widetilde{\boldsymbol{y}} - \Delta f &= R_f,\\
\Delta B\widetilde{\boldsymbol{x}} - \Delta D\widetilde{\boldsymbol{y}} + \Delta C^T\widetilde{\boldsymbol{z}} - \Delta g &= R_g,\\
\Delta C\widetilde{\boldsymbol{y}} + \Delta E\widetilde{\boldsymbol{z}} - \Delta h &= R_h,
\end{aligned}\right\} \tag{4.3.10}$$

and $\Delta A \in \mathcal{S}_n$, $\Delta D \in \mathcal{S}_m$, $\Delta E \in \mathcal{S}_p$. To maintain the sparsity pattern of A, B, C, D and E to the perturbation matrices, we replace $\Delta A, \Delta B, \Delta C, \Delta D$, and ΔE by $\Delta A \odot \Theta_A$, $\Delta B \odot \Theta_B$,

$\Delta C \odot \Theta_C, \Delta D \odot \Theta_D$, and $\Delta E \odot \Theta_E$, respectively. Thus (4.3.10) can be equivalently reformulated as:

$$\theta_1^{-1}\theta_1(\Delta A \odot \Theta_A)\widetilde{\boldsymbol{x}} + \theta_2^{-1}\theta_2(\Delta B \odot \Theta_B)^T\widetilde{\boldsymbol{y}} - \theta_8^{-1}\theta_8\Delta f = R_f, \tag{4.3.11}$$

$$\theta_2^{-1}\theta_2(\Delta B \odot \Theta_B)\widetilde{\boldsymbol{x}} - \theta_4^{-1}\theta_4(\Delta D \odot \Theta_D)\widetilde{\boldsymbol{y}} + \theta_5^{-1}\theta_5(\Delta C \odot \Theta_C)^T\widetilde{\boldsymbol{z}} - \theta_9^{-1}\theta_9\Delta g = R_g, \tag{4.3.12}$$

$$\theta_5^{-1}\theta_5(\Delta C \odot \Theta_C)\widetilde{\boldsymbol{x}} + \theta_7^{-1}\theta_7(\Delta E \odot \Theta_E)^T\widetilde{\boldsymbol{z}} - \theta_{10}^{-1}\theta_{10}\Delta h = R_h. \tag{4.3.13}$$

Using Lemma 4.3.4 in (4.3.11), we get

$$\theta_1^{-1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\Phi_A\theta_1\mathrm{vec}_{\mathcal{S}}(A \odot \Theta_A) + \theta_2^{-1}N^n_{\widetilde{\boldsymbol{y}}}\Phi_B\theta_2\mathrm{vec}(B \odot \Theta_B) - \theta_8^{-1}\theta_8\Delta f = R_f. \tag{4.3.14}$$

Further, (4.3.14) can be expressed as

$$\theta_1^{-1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\Phi_A\mathfrak{D}^{-1}_{\mathcal{S}_n}\mathfrak{D}_{\mathcal{S}_n}\theta_1\mathrm{vec}_{\mathcal{S}}(A \odot \Theta_A) + \theta_2^{-1}N^n_{\widetilde{\boldsymbol{y}}}\Phi_B\theta_2\mathrm{vec}(B \odot \Theta_B) - \theta_8^{-1}\theta_8\Delta f = R_f. \tag{4.3.15}$$

Equivalently, (4.3.15) can be written as follows:

$$\mathcal{J}^1_{\mathbb{S}_1}\Delta X = R_f, \tag{4.3.16}$$

where $\mathcal{J}^1_{\mathbb{S}_1} = \begin{bmatrix}\theta_1^{-1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\Phi_A\mathfrak{D}^{-1}_{\mathcal{S}_n} & \theta_2^{-1}N^n_{\widetilde{\boldsymbol{y}}}\Phi_B & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\theta_8^{-1}I_n & \mathbf{0} & \mathbf{0}\end{bmatrix} \in \mathbb{R}^{n\times \boldsymbol{l}}$ and

$$\begin{aligned}\Delta X =&[\theta_1\mathfrak{D}_{\mathcal{S}_n}\mathrm{vec}_{\mathcal{S}}(\Delta A \odot \Theta_A)^T, \theta_2\mathrm{vec}(\Delta B \odot \Theta_B)^T, \theta_4\mathfrak{D}_{\mathcal{S}_m}\mathrm{vec}_{\mathcal{S}}(\Delta D \odot \Theta_D)^T,\\ &\theta_5\mathrm{vec}(\Delta C \odot \Theta_C)^T, \theta_7\mathfrak{D}_{\mathcal{S}_p}\mathrm{vec}_{\mathcal{S}}(\Delta E \odot \Theta_E)^T, \theta_8\Delta f^T, \theta_9\Delta g^T, \theta_{10}\Delta h^T]^T \in \mathbb{R}^{\boldsymbol{l}}.\end{aligned} \tag{4.3.17}$$

Similarly, using Lemma 4.3.4 to (4.3.12) and (4.3.13), we obtain

$$\mathcal{J}^2_{\mathbb{S}_1}\Delta X = R_g \quad \text{and} \quad \mathcal{J}^3_{\mathbb{S}_1}\Delta X = R_h, \tag{4.3.18}$$

where $\mathcal{J}^2_{\mathbb{S}_1} \in \mathbb{R}^{m\times \boldsymbol{l}}$ and $\mathcal{J}^3_{\mathbb{S}_1} \in \mathbb{R}^{p\times \boldsymbol{l}}$ are given by

$$\mathcal{J}^2_{\mathbb{S}_1} = \begin{bmatrix}\mathbf{0} & \theta_2^{-1}M^m_{\widetilde{\boldsymbol{x}}}\Phi_B & -\theta_4^{-1}\mathcal{K}_{\widetilde{\boldsymbol{y}}}\Phi_D\mathfrak{D}^{-1}_{\mathcal{S}_m} & \theta_5^{-1}N^m_{\widetilde{\boldsymbol{z}}}\Phi_C & \mathbf{0} & \mathbf{0} & -\theta_9^{-1}I_m & \mathbf{0}\end{bmatrix}$$

and

$$\mathcal{J}^3_{\mathbb{S}_1} = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0} & \theta_5^{-1}\mathcal{M}^p_{\widetilde{\boldsymbol{y}}}\Phi_C & \theta_7^{-1}\mathcal{K}_{\widetilde{\boldsymbol{z}}}\mathfrak{D}^{-1}_{\mathcal{S}_p}\Phi_E & \mathbf{0} & \mathbf{0} & -\theta_{10}^{-1}I_p\end{bmatrix},$$

respectively. Combining (4.3.16) and (4.3.18), we obtain the following equivalent system

$$\mathcal{J}_{\mathbb{S}_1}\Delta X = R_{\mathbf{d}}. \tag{4.3.19}$$

Clearly, for $\theta_8, \theta_9, \theta_{10} \neq 0$, $\mathcal{J}_{\mathbb{S}_1}$ has full row rank. Therefore, by Lemma 1.3.1, the minimum norm solution of (4.3.19) is given by

$$\Delta X_{\min} = \mathcal{J}^{\dagger}_{\mathbb{S}_1}R_{\mathbf{d}} = \mathcal{J}^T_{\mathbb{S}_1}(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}^T_{\mathbb{S}_1})^{-1}R_{\mathbf{d}}. \tag{4.3.20}$$

On the other hand, the minimization problem in Definition 4.3.1 can be reformulated as:

$$
\begin{aligned}
[\boldsymbol{\eta}^{\mathbb{S}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}},\widetilde{\boldsymbol{z}})]^2 = \min\Bigg\{ & \theta_1^2\|\Delta A\odot\Theta_A\|_F^2+\theta_2^2\|\Delta B\odot\Theta_B\|_F^2+\theta_4^2\|\Delta D\odot\Theta_D\|_F^2+\theta_5^2\|\Delta C\odot\Theta_C\|_F^2 \\
& +\theta_7^2\|\Delta E\odot\Theta_E\|_F^2+\theta_8^2\|\Delta f\|_2^2+\theta_9^2\|\Delta g\|_2^2+\theta_{10}^2\|\Delta h\|_2^2 \;\Big| \\
& \begin{pmatrix} \Delta A\odot\Theta_A,\Delta B\odot\Theta_B,\Delta C\odot\Theta_C, \\ \Delta D\odot\Theta_D,\Delta E\odot\Theta_E,\Delta f, \\ \Delta g,\Delta h \end{pmatrix}\in\mathbb{S}_1\Bigg\} \\
= \min\Bigg\{ & \theta_1^2\|\mathfrak{D}_{\mathcal{S}_n}\mathrm{vec}_{\mathcal{S}}(\Delta A\odot\Theta_A)\|_2^2+\theta_2^2\|\mathrm{vec}(\Delta B\odot\Theta_B)\|_2^2 \\
& +\theta_4^2\|\mathfrak{D}_{\mathcal{S}_m}\mathrm{vec}_{\mathcal{S}}(\Delta D\odot\Theta_D)\|_2^2+\theta_5^2\|\mathrm{vec}(\Delta C\odot\Theta_C)\|_2^2 \\
& +\theta_7^2\|\mathfrak{D}_{\mathcal{S}_p}\mathrm{vec}_{\mathcal{S}}(\Delta E\odot\Theta_E)\|_2^2+\theta_8^2\|\Delta f\|_2^2+\theta_9^2\|\Delta g\|_2^2+\theta_{10}^2\|\Delta h\|_2^2 \;\Big| \\
& \mathcal{J}_{\mathbb{S}_1}\Delta X = R_{\mathbf{d}}\Bigg\} \qquad (4.3.21) \\
= \min\Big\{ & \|\Delta X\|_2^2 \;\Big|\; \mathcal{J}_{\mathbb{S}_1}\Delta X = R_{\mathbf{d}}\Big\} = \|\Delta X_{\min}\|_2^2. \qquad (4.3.22)
\end{aligned}
$$

Consequently, substituting (4.3.20) into (4.3.21), we obtain

$$
\boldsymbol{\eta}^{\mathbb{S}_1}_{\mathtt{sps}}(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}},\widetilde{\boldsymbol{z}}) = \left\|\mathcal{J}_{\mathbb{S}_1}^T(\mathcal{J}_{\mathbb{S}_1}\mathcal{J}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}}\right\|_2.
$$

From (4.3.17), we have $\theta_1\mathfrak{D}_{\mathcal{S}_n}\mathrm{vec}_{\mathcal{S}}(\Delta A\odot\Theta_A) = \begin{bmatrix} I_\mu & \mathbf{0}\end{bmatrix}\Delta X$. Therefore, the generating vector for the minimal perturbation matrix $\widehat{\Delta A}_{\mathtt{sps}}$ which also preserves the sparsity pattern is given by

$$
\mathrm{vec}_{\mathcal{S}}(\widehat{\Delta A}_{\mathtt{sps}}) = \theta_1^{-1}\mathfrak{D}_{\mathcal{S}_n}^{-1}\begin{bmatrix} I_\mu & \mathbf{0}\end{bmatrix}\Delta X_{\min}.
$$

Similarly, the generating vectors for other minimal perturbation matrices can be obtained. Hence, the proof is completed. ■ The following result gives the structured BE without preserving sparsity.

Corollary 4.3.1. *Suppose the approximate solution of the DSPP (4.3.1) with $A\in\mathcal{S}_n$, $D\in\mathcal{S}_m$, $E\in\mathcal{S}_p$, and $\theta_8,\theta_9,\theta_{10}\neq 0$ is $[\widetilde{\boldsymbol{x}}^T,\widetilde{\boldsymbol{y}}^T,\widetilde{\boldsymbol{z}}^T]^T$. Then, we have*

$$
\boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{y}},\widetilde{\boldsymbol{z}}) = \left\|\widehat{\mathcal{J}}_{\mathbb{S}_1}^T(\widehat{\mathcal{J}}_{\mathbb{S}_1}\widehat{\mathcal{J}}_{\mathbb{S}_1}^T)^{-1}R_{\mathbf{d}}\right\|_2, \qquad (4.3.23)
$$

where $\widehat{\mathcal{J}}_{\mathbb{S}_1}\in\mathbb{R}^{(n+m+p)\times\boldsymbol{l}}$ is given by

$$
\widehat{\mathcal{J}}_{\mathbb{S}_1} = \begin{bmatrix}
\frac{1}{\theta_1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\mathfrak{D}_{\mathcal{S}_n}^{-1} & \frac{1}{\theta_2}\mathcal{N}_{\widetilde{\boldsymbol{y}}}^n & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_8}I_n & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \frac{1}{\theta_2}M_{\widetilde{\boldsymbol{x}}}^m & -\frac{1}{\theta_4}\mathcal{K}_{\widetilde{\boldsymbol{y}}}\mathfrak{D}_{\mathcal{S}_m}^{-1} & \frac{1}{\theta_5}N_{\widetilde{\boldsymbol{z}}}^m & \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_9}I_m & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{\theta_5}\mathcal{M}_{\widetilde{\boldsymbol{y}}}^p & \frac{1}{\theta_7}\mathcal{K}_{\widetilde{\boldsymbol{z}}}\mathfrak{D}_{\mathcal{S}_p}^{-1} & \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_{10}}I_p
\end{bmatrix}.
$$

Proof. Since we are not preserving the sparsity pattern, the proof follows by considering $\Theta_A = \mathbf{1}_{n\times n}$, $\Theta_B = \mathbf{1}_{m\times n}$, $\Theta_D = \mathbf{1}_{m\times m}$, $\Theta_C = \mathbf{1}_{p\times m}$, and $\Theta_E = \mathbf{1}_{p\times p}$ in Theorem 4.3.5. ■

Remark 4.3.6. *The structure-preserving minimal perturbation matrices* $\widehat{\Delta A}$, $\widehat{\Delta B}$, $\widehat{\Delta C}$, $\widehat{\Delta D}$, $\widehat{\Delta E}$, $\widehat{\Delta f}$, $\widehat{\Delta g}$, *and* $\widehat{\Delta h}$, *for which* $\boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ *is attained are given by formulae presented in Theorem 4.3.5 with* $\mathcal{J}_{\mathbb{S}_1} = \widehat{\mathcal{J}}_{\mathbb{S}_1}$.

In the next result, we present the formula of structured BE when $D = \mathbf{0}$ and $E = \mathbf{0}$.

Corollary 4.3.2. *Suppose* $[\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ *is an approximate solution of the DSPP* (4.3.1) *with* $A \in \mathcal{S}_n$, $D = \mathbf{0}$, $E = \mathbf{0}$, *and* $\theta_8, \theta_9, \theta_{10} \neq 0$. *Then, we have*

$$\boldsymbol{\eta}^{\mathbb{S}_1}_{\mathbf{sps}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \left\| \widetilde{\mathcal{J}}^T_{\mathbb{S}_1} (\widetilde{\mathcal{J}}_{\mathbb{S}_1} \widetilde{\mathcal{J}}^T_{\mathbb{S}_1})^{-1} R_{\mathbf{d}} \right\|_2, \tag{4.3.24}$$

where $\widetilde{\mathcal{J}}_{\mathbb{S}_1} \in \mathbb{R}^{(n+m+p)\times \boldsymbol{l}}$ *is given by*

$$\widetilde{\mathcal{J}}_{\mathbb{S}_1} = \begin{bmatrix} \frac{1}{\theta_1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\Phi_A\mathfrak{D}^{-1}_{\mathcal{S}_n} & \frac{1}{\theta_2}\mathcal{N}^n_{\widetilde{\boldsymbol{y}}}\Phi_B & \mathbf{0} & -\frac{1}{\theta_8}I_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\theta_2}M^m_{\widetilde{\boldsymbol{x}}}\Phi_B & \frac{1}{\theta_5}N^m_{\widetilde{\boldsymbol{z}}}\Phi_C & \mathbf{0} & -\frac{1}{\theta_9}I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\theta_5}\mathcal{M}^p_{\widetilde{\boldsymbol{y}}}\Phi_C & \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_{10}}I_p \end{bmatrix},$$

$R_f = f - A\boldsymbol{x} - B^T\boldsymbol{y}$, $R_g = g - B\boldsymbol{x} - C^T\boldsymbol{z}$, $R_h = h - C\boldsymbol{y}$, *and* $\boldsymbol{l} = \mu + mn + mp + m + n + p$.

Proof. Since $D = \mathbf{0}$ and $E = \mathbf{0}$, the proof follows by considering $\theta_4 = \theta_7 = 0$. ■

Remark 4.3.7. *When* $D = \mathbf{0}$ *and* $E = \mathbf{0}$, *Lv and Zheng [96] derive the structured BE for the DSPP* (4.3.1). *However, their investigations do not take into account the sparsity pattern of the coefficient matrices.*

4.3.3. Derivation of Structured BEs for Case (ii)

In this subsection, we derive explicit formulae for the structured BEs for the DSPP for the case (ii), i.e., $A \in \mathcal{S}_n$, $B \neq F$, $D \in \mathcal{S}_m$, $C = G$ and $E \in \mathcal{S}_p$. We use the Lemmas 4.3.4, 4.1.9 and 1.3.1, and apply a similar methodology used in Section 4.3.2 to derive the formulae for the structured BEs. In the next result, we present computable formulae for the structured BE $\boldsymbol{\eta}^{\mathbb{S}_2}_{\mathbf{sps}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ by preserving sparsity pattern of the original matrices to the perturbation matrices. Before proceeding, we set the following notation:

$$\Phi_F = \mathrm{diag}(\mathrm{vec}(\Theta_F)),$$

$\Phi_A, \Phi_B\Phi_C, \Phi_D$ and Φ_E are same as defined in Subsection 4.3.2.

Theorem 4.3.8. *Let* $[\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ *be an approximate solution of the DSPP* (4.3.1) *with* $A \in \mathcal{S}_n$, $B \neq F$, $D \in \mathcal{S}_m$, $C = G$, $E \in \mathcal{S}_p$ *and* $\theta_8, \theta_9, \theta_{10} \neq 0$. *Then, we have*

$$\boldsymbol{\eta}_{\mathbf{sps}}^{\mathbb{S}_2}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \left\| \mathcal{J}_{\mathbb{S}_2}^T (\mathcal{J}_{\mathbb{S}_2} \mathcal{J}_{\mathbb{S}_2}^T)^{-1} R_{\mathbf{d}} \right\|_2, \tag{4.3.25}$$

where $\mathcal{J}_{\mathbb{S}_2} = [\widetilde{\mathcal{J}}_{\mathbb{S}_2} \;\; \mathcal{I}] \in \mathbb{R}^{(n+m+p)\times \boldsymbol{l}}$ *and* $\widetilde{\mathcal{J}}_{\mathbb{S}_2}$ *is given by*

$$\widetilde{\mathcal{J}}_{\mathbb{S}_2} = \begin{bmatrix} \frac{1}{\theta_1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\Phi_A\mathfrak{D}_{\mathcal{S}_n}^{-1} & \frac{1}{\theta_2}\mathcal{N}_{\widetilde{\boldsymbol{y}}}^n\Phi_B & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\theta_3}M_{\widetilde{\boldsymbol{x}}}^m\Phi_F & -\frac{1}{\theta_4}\mathcal{K}_{\widetilde{\boldsymbol{y}}}\Phi_D\mathfrak{D}_{\mathcal{S}_m}^{-1} & \frac{1}{\theta_5}N_{\widetilde{\boldsymbol{z}}}^m\Phi_C & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{\theta_5}\mathcal{M}_{\widetilde{\boldsymbol{y}}}^p\Phi_C & \frac{1}{\theta_7}\mathcal{K}_{\widetilde{\boldsymbol{z}}}\Phi_E\mathfrak{D}_{\mathcal{S}_p}^{-1} \end{bmatrix},$$

$R_f = f - A\widetilde{\boldsymbol{x}} - B^T\widetilde{\boldsymbol{y}}$, $R_g = g - F\widetilde{\boldsymbol{x}} + D\widetilde{\boldsymbol{y}} - C^T\widetilde{\boldsymbol{z}}$, $R_h = h - C\widetilde{\boldsymbol{y}} - E\widetilde{\boldsymbol{z}}$, $R_{\mathbf{d}} = [R_f^T, R_g^T, R_h^T]^T$, *and* $\boldsymbol{l} = \mu + \sigma + \tau + 2mn + mp + m + n + p$.

Proof. Given that $[\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ is an approximate solution of the DSPP (4.3.1) for the case (ii). Now, we are required to construct perturbation matrices $\Delta A \in \mathcal{S}_n$, $\Delta B, \Delta F \in \mathbb{R}^{m\times n}$, $\Delta D \in \mathcal{S}_m$, $\Delta C \in \mathbb{R}^{p\times m}$, $\Delta E \in \mathcal{S}_p$, which maintain the sparsity pattern of A, B, F, C, D, E, respectively, and the perturbations $\Delta f \in \mathbb{R}^n$, $\Delta g \in \mathbb{R}^m$ and $\Delta h \in \mathbb{R}^p$. Using (4.3.5), $\begin{pmatrix} \Delta A, \Delta B, \Delta C, \\ \Delta D, \Delta E, \Delta F, \\ \Delta f, \Delta g, \Delta h \end{pmatrix} \in \mathbb{S}_2$ if and only if $\Delta A, \Delta B, \Delta C, \Delta D, \Delta E, \Delta F, \Delta f, \Delta g$ and Δh satisfy the following equations:

$$\left.\begin{aligned} &\Delta A\widetilde{\boldsymbol{x}} + \Delta B^T\widetilde{\boldsymbol{y}} - \Delta f = R_f, \\ &\Delta F\widetilde{\boldsymbol{x}} - \Delta D\widetilde{\boldsymbol{y}} + \Delta C^T\widetilde{\boldsymbol{z}} - \Delta g = R_g, \\ &\Delta C\widetilde{\boldsymbol{y}} + \Delta E\widetilde{\boldsymbol{z}} - \Delta h = R_h, \end{aligned}\right\} \tag{4.3.26}$$

and $\Delta A \in \mathcal{S}_n, \Delta D \in \mathcal{S}_m$, $\Delta E \in \mathcal{S}_p$.

By following a similar the proof methodology of Theorem 4.3.5 and applying Lemma 4.3.4, we get:

$$\mathcal{J}_{\mathbb{S}_2}^1 \Delta X = R_f, \;\; \mathcal{J}_{\mathbb{S}_2}^2 \Delta X = R_g \text{ and } \mathcal{J}_{\mathbb{S}_2}^3 \Delta X = R_h, \tag{4.3.27}$$

where

$$\begin{aligned} \mathcal{J}_{\mathbb{S}_2}^1 &= \begin{bmatrix} \theta_1^{-1}\mathcal{K}_{\widetilde{\boldsymbol{x}}}\Phi_A\mathfrak{D}_{\mathcal{S}_n}^{-1} & \theta_2^{-1}N_{\widetilde{\boldsymbol{y}}}^n\Phi_B & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\theta_8^{-1}I_n & \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{n\times \boldsymbol{l}}, \\ \mathcal{J}_{\mathbb{S}_2}^2 &= \begin{bmatrix} \mathbf{0} & \mathbf{0} & \theta_3^{-1}M_{\widetilde{\boldsymbol{x}}}^m\Phi_F & -\theta_4^{-1}\mathcal{K}_{\widetilde{\boldsymbol{y}}}\Phi_C\mathfrak{D}_{\mathcal{S}_m}^{-1} & \theta_5^{-1}N_{\widetilde{\boldsymbol{y}}}^m\Phi_D & \mathbf{0} & \mathbf{0} & -\theta_9^{-1}I_m & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{m\times \boldsymbol{l}}, \\ \mathcal{J}_{\mathbb{S}_2}^3 &= \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \theta_5^{-1}\mathcal{M}_{\widetilde{\boldsymbol{y}}}^p\Phi_D & \theta_7^{-1}\mathcal{K}_{\widetilde{\boldsymbol{z}}}\Phi_E\mathfrak{D}_{\mathcal{S}_p}^{-1} & \mathbf{0} & \mathbf{0} & -\theta_{10}^{-1}I_p \end{bmatrix} \in \mathbb{R}^{p\times \boldsymbol{l}}, \end{aligned}$$

and $\Delta X = [\theta_1 \mathfrak{D}_{\mathcal{S}_n} \text{vec}_{\mathcal{S}}(\Delta A \odot \Theta_A)^T,\ \theta_2 \text{vec}(\Delta B \odot \Theta_B)^T,\ \theta_3 \text{vec}(\Delta F \odot \Theta_F)^T,\ \theta_4 \mathfrak{D}_{\mathcal{S}_m} \text{vec}_{\mathcal{S}}(\Delta D \odot \Theta_D)^T,\ \theta_5 \text{vec}(\Delta C \odot \Theta_C)^T,\ \theta_7 \mathfrak{D}_{\mathcal{S}_p} \text{vec}_{\mathcal{S}}(\Delta E \odot \Theta_E)^T,\ \theta_8 \Delta f^T,\ \theta_9 \Delta g^T,\ \theta_{10} \Delta h^T]^T \in \mathbb{R}^{\boldsymbol{l}}$. Combining the three equations in (4.3.27), we obtain

$$\mathcal{J}_{\mathbb{S}_2} \Delta X = R_{\mathbf{d}}. \tag{4.3.28}$$

Since, $\mathcal{J}_{\mathbb{S}_2}$ has full row rank for $\theta_8, \theta_9, \theta_{10} \neq 0$. Therefore, (4.3.28) is consistent and by Lemma 1.3.1, its minimum norm solution is given by

$$\Delta X_{\min} = \mathcal{J}_{\mathbb{S}_2}^T (\mathcal{J}_{\mathbb{S}_2} \mathcal{J}_{\mathbb{S}_2}^T)^{-1} R_{\mathbf{d}}. \tag{4.3.29}$$

Now, applying a similar argument to the proof method of Theorem 4.3.5, the required structured BE is

$$\boldsymbol{\eta}_{\mathbf{sps}}^{\mathbb{S}_2}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \|\Delta X_{\min}\|_2 = \left\| \mathcal{J}_{\mathbb{S}_2}^T (\mathcal{J}_{\mathbb{S}_2} \mathcal{J}_{\mathbb{S}_2}^T)^{-1} R_{\mathbf{d}} \right\|_2.$$

Hence, the proof is completed. ■

Remark 4.3.9. *The minimal perturbation matrices* $\widehat{\Delta A}_{\mathtt{sps}}$, $\widehat{\Delta C}_{\mathtt{sps}}$, $\widehat{\Delta D}_{\mathtt{sps}}$, $\widehat{\Delta E}_{\mathtt{sps}}$, $\widehat{\Delta f}_{\mathtt{sps}}$, $\widehat{\Delta g}_{\mathtt{sps}}$ *and* $\widehat{\Delta h}_{\mathtt{sps}}$ *for the Problem 4.3.1 can be computed using the formulae provided in Theorem 4.3.5 by replacing* $\mathcal{J}_{\mathbb{S}_1}$ *with* $\mathcal{J}_{\mathbb{S}_2}$. *The generating vectors for the minimal perturbation matrices* $\widehat{\Delta B}_{\mathtt{sps}}$ *and* $\widehat{\Delta F}_{\mathtt{sps}}$ *are given by*

$$\text{vec}(\widehat{\Delta B}_{\mathtt{sps}}) = \frac{1}{\theta_2} \begin{bmatrix} \mathbf{0}_\mu & I_{mn} & \mathbf{0}_{\boldsymbol{l}-\mu-mn} \end{bmatrix} \mathcal{J}_{\mathbb{S}_2}^T (\mathcal{J}_{\mathbb{S}_2} \mathcal{J}_{\mathbb{S}_2}^T)^{-1} R_{\mathbf{d}},$$

$$\text{vec}(\widehat{\Delta F}_{\mathtt{sps}}) = \frac{1}{\theta_3} \begin{bmatrix} \mathbf{0}_{\mu+mn} & I_{mn} & \mathbf{0}_{\boldsymbol{l}-\mu-2mn} \end{bmatrix} \mathcal{J}_{\mathbb{S}_2}^T (\mathcal{J}_{\mathbb{S}_2} \mathcal{J}_{\mathbb{S}_2}^T)^{-1} R_{\mathbf{d}}.$$

The following result gives structured BE while the sparsity pattern is not preserved.

Corollary 4.3.3. *Let* $[\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ *be an approximate solution of the DSPP (4.3.1) with* $A \in \mathcal{S}_n$, $D \in \mathcal{S}_m$, $E \in \mathcal{S}_p$, *and* $\theta_8, \theta_9, \theta_{10} \neq 0$. *Then, we have*

$$\boldsymbol{\eta}^{\mathbb{S}_2}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \left\| \widehat{\mathcal{J}}_{\mathbb{S}_2}^T (\widehat{\mathcal{J}}_{\mathbb{S}_2} \widehat{\mathcal{J}}_{\mathbb{S}_2}^T)^{-1} R_{\mathbf{d}} \right\|_2, \tag{4.3.30}$$

where $\widehat{\mathcal{J}}_{\mathbb{S}_2} \in \mathbb{R}^{(n+m+p)\times \boldsymbol{l}}$ *is given by*

$$\widehat{\mathcal{J}}_{\mathbb{S}_2} = \begin{bmatrix} \frac{1}{\theta_1}\mathcal{M}_{\widetilde{\boldsymbol{x}}}^n & \frac{1}{\theta_2}\mathcal{N}_{\widetilde{\boldsymbol{y}}}^n & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_8} I_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\theta_3} M_{\widetilde{\boldsymbol{x}}}^m & -\frac{1}{\theta_4}\mathcal{K}_{\widetilde{\boldsymbol{y}}} \mathfrak{D}_{\mathcal{S}_m}^{-1} & \frac{1}{\theta_5} N_{\widetilde{\boldsymbol{z}}}^m & \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_9} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{\theta_5}\mathcal{M}_{\widetilde{\boldsymbol{y}}}^p & \frac{1}{\theta_7}\mathcal{K}_{\widetilde{\boldsymbol{z}}} \mathfrak{D}_{\mathcal{S}_p}^{-1} & \mathbf{0} & \mathbf{0} & -\frac{1}{\theta_{10}} I_p \end{bmatrix}.$$

Proof. The proof proceeds by choosing $\Theta_A = \mathbf{1}_{n\times n}$, $\Theta_B = \Theta_F = \mathbf{1}_{m\times n}$, $\Theta_D = \mathbf{1}_{m\times m}$, $\Theta_C = \mathbf{1}_{p\times m}$, and $\Theta_E = \mathbf{1}_{p\times p}$ in the expression of the structured BE presented in Theorem 4.3.8. ■

Remark 4.3.10. *Similar to Corollary 4.3.2, we can compute the structured BE for the case (ii) with $D = \mathbf{0}$ and $E = \mathbf{0}$. This specific instance of structured BE has also been addressed in [98]. However, our investigation additionally ensures the preservation of the sparsity pattern.*

4.3.4. Derivation of Structured BEs for Case (*iii*)

This subsection deals with the structured BE of the DSPP for the case (*iii*), i.e., $A \neq A^T$, $D \in \mathcal{S}_m$, $E \in \mathcal{S}_p$, $B = F$, and $C \neq G$. Using a similar technique as in Section 4.3.2, in the following theorem, we present the computable formula of the structured BE when sparsity pattern of the original matrices are preserved in the perturbation matrices. Before continuing, we define $\Phi_G = \text{diag}(\text{vec}(\Theta_G))$, along with Φ_A, Φ_B, Φ_C, Φ_D and Φ_E as defined in Subsection 4.3.2.

Theorem 4.3.11. *Let $[\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ be an approximate solution of the DSPP (4.3.1) with $C \in \mathcal{S}_m$, $E \in \mathcal{S}_p$, and $\theta_8, \theta_9, \theta_{10} \neq 0$, $B = F$ and $C \neq G$. Then, we have*

$$\boldsymbol{\eta}_{\mathbf{sps}}^{\mathbb{S}_3}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = \left\| \mathcal{J}_{\mathbb{S}_3}^T (\mathcal{J}_{\mathbb{S}_3} \mathcal{J}_{\mathbb{S}_3}^T)^{-1} R_{\mathbf{d}} \right\|_2, \tag{4.3.31}$$

where $\mathcal{J}_{\mathbb{S}_3} = [\widetilde{\mathcal{J}}_{\mathbb{S}_3} \;\; \mathcal{I}] \in \mathbb{R}^{(n+m+p)\times \boldsymbol{l}}$, $\widetilde{\mathcal{J}}_{\mathbb{S}_3}$ is given by

$$\widetilde{\mathcal{J}}_{\mathbb{S}_3} = \begin{bmatrix} \frac{1}{\theta_1}\mathcal{M}_{\widetilde{\boldsymbol{x}}}^n \Phi_A & \frac{1}{\theta_2}\mathcal{N}_{\widetilde{\boldsymbol{y}}}^n \Phi_B & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\theta_2} M_{\widetilde{\boldsymbol{x}}}^m \Phi_B & -\frac{1}{\theta_4}\mathcal{K}_{\widetilde{\boldsymbol{y}}} \Phi_D \mathfrak{D}_{\mathcal{S}_m}^{-1} & \frac{1}{\theta_5} N_{\widetilde{\boldsymbol{z}}}^m \Phi_C & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \frac{1}{\theta_6}\mathcal{M}_{\widetilde{\boldsymbol{y}}}^p \Phi_G & \frac{1}{\theta_7}\mathcal{K}_{\widetilde{\boldsymbol{z}}} \Phi_E \mathfrak{D}_{\mathcal{S}_p}^{-1} \end{bmatrix},$$

$R_f = f - A\widetilde{\boldsymbol{x}} - B^T\widetilde{\boldsymbol{y}}$, $R_g = g - B\widetilde{\boldsymbol{x}} + D\widetilde{\boldsymbol{y}} - C^T\widetilde{\boldsymbol{z}}$, $R_h = h - G\widetilde{\boldsymbol{y}} - E\widetilde{\boldsymbol{z}}$, $R_{\mathbf{d}} = [R_f^T, R_g^T, R_h^T]^T$, *and* $\boldsymbol{l} = n^2 + \sigma + \tau + mn + 2mp + m + n + p$.

Proof. The proof follows by using a similar proof methodology of Theorem 4.3.5. ■

Remark 4.3.12. *The minimal perturbation matrices $\widehat{\Delta B}_{\mathbf{sps}}$, $\widehat{\Delta D}_{\mathbf{sps}}$, $\widehat{\Delta E}_{\mathbf{sps}}$, $\widehat{\Delta f}_{\mathbf{sps}}$, $\widehat{\Delta g}_{\mathbf{sps}}$ and $\widehat{\Delta h}_{\mathbf{sps}}$ can be computed using the formulae provided in Theorem 4.3.5 with $\mathcal{J}_{\mathbb{S}_1} = \mathcal{J}_{\mathbb{S}_3}$. The generating vector for the minimal perturbation matrices $\widehat{\Delta A}_{\mathbf{sps}}$, $\widehat{\Delta C}_{\mathbf{sps}}$ and $\widehat{\Delta G}_{\mathbf{sps}}$ are given by*

$$\text{vec}(\widehat{\Delta A}_{\mathbf{sps}}) = \frac{1}{\theta_1}\begin{bmatrix} I_{n^2} & \mathbf{0}_{\boldsymbol{l}-n^2} \end{bmatrix} \mathcal{J}_{\mathbb{S}_3}^T (\mathcal{J}_{\mathbb{S}_3} \mathcal{J}_{\mathbb{S}_3}^T)^{-1} R_{\mathbf{d}}.$$

$$\text{vec}(\widehat{\Delta C}_{\mathbf{sps}}) = \frac{1}{\theta_5}\begin{bmatrix} \mathbf{0}_{n^2+\mathbf{m}+mn} & I_{mp} & \mathbf{0}_{\tau+mp+n+m+p} \end{bmatrix} \mathcal{J}_{\mathbb{S}_3}^T (\mathcal{J}_{\mathbb{S}_3} \mathcal{J}_{\mathbb{S}_3}^T)^{-1} R_{\mathbf{d}},$$

$$\text{vec}(\widehat{\Delta G}_{\mathbf{sps}}) = \frac{1}{\theta_6}\begin{bmatrix} \mathbf{0}_{n^2+\mathbf{m}+mn+mp} & I_{mp} & \mathbf{0}_{\tau+n+m+p} \end{bmatrix} \mathcal{J}_{\mathbb{S}_3}^T (\mathcal{J}_{\mathbb{S}_3} \mathcal{J}_{\mathbb{S}_3}^T)^{-1} R_{\mathbf{d}}.$$

By taking $\Theta_A = \mathbf{1}_{n\times n}$, $\Theta_B = \mathbf{1}_{m\times n}$, $\Theta_D = \mathbf{1}_{m\times m}$, $\Theta_C = \Theta_G = \mathbf{1}_{p\times m}$, and $\Theta_E = \mathbf{1}_{p\times p}$ in the BE expression provided in Theorem 4.3.11, we obtain the structured BE when the sparsity pattern is not considered.

Remark 4.3.13. *The structured BE for the DSPP (4.3.1) when $A \in \mathcal{S}_n$, $D \in \mathcal{S}_m$, $E \in \mathcal{S}_p$, $B \neq F$, $C \neq G$, or, $A \neq A^T$, $D \in \mathcal{S}_m$, $E \in \mathcal{S}_p$, $B \neq F$, $C \neq G$ can be derived in a similar technique used in this section and in Subsections 4.3.2 and 4.3.3. As the derivation process is similar, we have not studied them here in detail.*

4.4. Numerical Experiments

In this section, we conduct a few numerical experiments to validate the findings of this chapter. All numerical experiments are conducted on MATLAB R2023b on an Intel(R) Core(TM) $i7$-10700 CPU, 2.90GHz, 16 GB. We denote

$$\widehat{\Delta\mathcal{M}}_{\text{sps}} = \begin{bmatrix} \widehat{\Delta A}_{\text{sps}} & \widehat{\Delta B}_{\text{sps}}^T \\ \widehat{\Delta B}_{\text{sps}} & \widehat{\Delta D}_{\text{sps}} \end{bmatrix} \text{ and } \widehat{\Delta\mathbf{b}}_{\text{sps}} = \begin{bmatrix} \widehat{\Delta f}_{\text{sps}} \\ \widehat{\Delta g}_{\text{sps}} \end{bmatrix}.$$

Example 4.4.1. Consider the circulant structured GSPP (4.1.1), where the circulant block matrices $A, B, D \in \mathcal{C}_3$ and $f, g \in \mathbb{R}^3$ are given by

$$A = \begin{bmatrix} 1.02 & 0 & 5.3 \\ 5.3 & 1.02 & 0 \\ 0 & 5.3 & 1.02 \end{bmatrix}, \quad B = \begin{bmatrix} -12.78 & 6.38 & 0 \\ 0 & -12.78 & 6.38 \\ 6.38 & 0 & -12.78 \end{bmatrix},$$

$$D = \begin{bmatrix} 59 & 1 & 0 \\ 0 & 59 & 1 \\ 1 & 0 & 59 \end{bmatrix}, f = \begin{bmatrix} 78.01 \\ 2 \\ 10 \end{bmatrix}, \text{ and } g = \begin{bmatrix} 56 \\ 3 \\ 1 \end{bmatrix}.$$

Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be an approximate solution of the GSPP, where $\widetilde{\boldsymbol{u}} = [-0.85, 6.04, 11.91]^T$, $\widetilde{\boldsymbol{p}} = [0.11, 0.026, 2.69]^T$ and $\|\mathcal{M}\widetilde{\boldsymbol{v}} - \mathbf{b}\|_2 = 0.2147$.

By applying formula (4.1.3), the computed unstructured BE $\boldsymbol{\eta}(\widetilde{\boldsymbol{v}})$ is 1.44790×10^{-04}. Again, using the formula outlined in Theorem 4.1.5, we obtained the structured BE with preserving the sparsity pattern is $\boldsymbol{\eta}_{\text{sps}}^{\mathcal{S}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = 5.5541$. In this case, the minimal perturbation matrices, which preserve the sparsity pattern as well as the circulant structure, are given by

$$\widehat{\Delta A}_{\text{sps}} = \begin{bmatrix} 0.00517 & 0 & 0.00026 \\ 0.00026 & 0.00517 & 0 \\ 0 & 0.00026 & 0.00517 \end{bmatrix}, \quad \widehat{\Delta B}_{\text{sps}} = \begin{bmatrix} -0.01439 & 0.00684 & 0 \\ 0 & -0.01439 & 0.00684 \\ 0.00684 & 0 & -0.01439 \end{bmatrix},$$

$$\widehat{\Delta D}_{\texttt{sps}} = \begin{bmatrix} -0.00133 & -0.00205 & 0 \\ 0 & -0.00133 & -0.00205 \\ -0.00205 & 0 & -0.00133 \end{bmatrix}, \widehat{\Delta f}_{\texttt{sps}} = \begin{bmatrix} 0.01799 \\ 0.05670 \\ -0.02749 \end{bmatrix}$$

$$\text{and } \widehat{\Delta g}_{\texttt{sps}} = \begin{bmatrix} -0.03242 \\ 0.00764 \\ 0.00616 \end{bmatrix}.$$

Moreover, without preserving the sparsity pattern, the structured BE is $\boldsymbol{\eta}^{\mathcal{S}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = 0.78540$. Further, in this case, the minimal perturbation matrices for which $\boldsymbol{\eta}^{\mathcal{S}_1}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ is attained and preserve circulant structure are given by

$$\widehat{\Delta A} = \begin{bmatrix} 0.00665 & -0.00595 & 0.00188 \\ 0.00188 & 0.00665 & -0.00595 \\ -0.00595 & 0.00188 & 0.00665 \end{bmatrix}, \quad \widehat{\Delta B} = \begin{bmatrix} -0.01636 & 0.00691 & 0.00246 \\ 0.00246 & -0.01636 & 0.00691 \\ 0.00691 & 0.00246 & -0.01636 \end{bmatrix},$$

$$\widehat{\Delta D} = \begin{bmatrix} -0.00021 & 0 & 0.00014 \\ 0.00014 & -0.00021 & 0 \\ 0 & 0.00014 & -0.00021 \end{bmatrix}, \widehat{\Delta f} = 10^{-7} \times \begin{bmatrix} 0.96212 \\ 1.41279 \\ -3.01440 \end{bmatrix}$$

$$\text{and } \widehat{\Delta g} = \begin{bmatrix} -0.00053 \\ -0.00002 \\ 0.00079 \end{bmatrix}.$$

We can observe that, in both the cases

$$(\mathcal{M} + \widehat{\Delta\mathcal{M}}_{\texttt{sps}})\widetilde{\boldsymbol{v}} = \mathbf{b} + \widehat{\Delta\mathbf{b}}_{\texttt{sps}} \text{ and } (\mathcal{M} + \widehat{\Delta\mathcal{M}})\widetilde{\boldsymbol{v}} = \mathbf{b} + \widehat{\Delta\mathbf{b}},$$

i.e., $\widetilde{\boldsymbol{v}}$ is the exact solution of the above circulant structured GSPP.

Example 4.4.2. Consider a Toeplitz structured GSPP where the coefficient matrices are given by:

$$A = \begin{bmatrix} 10^{-6} & 0 & 10^{3} & 0 \\ 10^{8} & 10^{-6} & 0 & 10^{3} \\ 10 & 10^{8} & 10^{-6} & 0 \\ 0 & 10 & 10^{8} & 10^{-6} \end{bmatrix}, \quad B = \begin{bmatrix} 10^{-5} & 10^{7} & 0 & 0 \\ 10^{5} & 10^{-5} & 10^{7} & 0 \\ 0 & 10^{5} & 10^{-5} & 10^{7} \\ 0 & 0 & 10^{5} & 10^{-5} \end{bmatrix},$$

$$D = \begin{bmatrix} 0 & 10^{8} & -60 & 0 \\ -0.5 & 0 & 10^{8} & -60 \\ 0 & -0.5 & 0 & 10^{8} \\ 0 & 0 & -0.5 & 0 \end{bmatrix}, f = \begin{bmatrix} 10^{8} \\ 0 \\ 10^{3} \\ 0 \end{bmatrix} \text{ and } g = \begin{bmatrix} 10^{-8} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The approximate solution of the GSPP, computed using Gaussian elimination with partial pivoting (GEP) is $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$, where

$$\widetilde{\boldsymbol{u}} = \begin{bmatrix} 6.0278 \times 10^{3} \\ -1.0000 \times 10^{4} \\ 9.8995 \times 10^{-3} \\ -9.9000 \times 10^{7} \end{bmatrix} \text{ and } \widetilde{\boldsymbol{p}} = \begin{bmatrix} -5.0378 \times 10^{4} \\ 1.0000 \times 10^{3} \\ -8.8995 \times 10^{-2} \\ 9.9000 \times 10^{6} \end{bmatrix}.$$

We choose $w_1 = 1/\|A\|_F$, $w_2 = 1/\|B\|_F$, $w_3 = 1/\|D\|_F$, $w_4 = 1/\|f\|_2$, and $w_5 = 1/\|g\|_2$. Using the formula in (4.1.3), the unstructured BE is $\boldsymbol{\eta}(\widetilde{\boldsymbol{v}}) = 6.2617 \times 10^{-18}$. However, the obtained structured BEs using Theorem 4.1.10 and Corollary 4.1.3 are $\boldsymbol{\eta}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = 2.1761 \times 10^{-9}$ and $\boldsymbol{\eta}^{\mathcal{S}_2}_{\mathtt{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}}) = 4.3070 \times 10^{-5}$, respectively. We can observe that the $\boldsymbol{\eta}(\widetilde{\boldsymbol{v}})$ in the order of $\mathcal{O}(10^{-18})$, whereas the structured BEs are significantly larger. This demonstrates that the GEP for solving this tested Toeplitz structured GSPP is backward stable but not strongly backward stable. That is, the computed approximate solution does not satisfy a nearby (sparsity preserving) Toeplitz structured GSPP.

Example 4.4.3. Consider the Toeplitz structured GSPP (4.1.1) with block matrices

$$A = toeplitz(\boldsymbol{a}_1, \boldsymbol{a}_2) \in \mathcal{T}_{n\times n}, \quad B = toeplitz(\boldsymbol{b}_1, \boldsymbol{b}_2) \in \mathcal{T}_{n\times n}, \quad D = toeplitz(\boldsymbol{c}_1, \boldsymbol{c}_2) \in \mathcal{T}_{n\times n},$$

where $\boldsymbol{a}_1 = \mathtt{sprand}(n, 1, 0.4)$, $\boldsymbol{a}_2 = randn(n, 1)$, $\boldsymbol{b}_1 = \mathtt{sprand}(n, 1, 0.1)$, $\boldsymbol{b}_2 = \mathtt{randn}(n, 1)$, $\boldsymbol{c}_1 = \mathtt{sprand}(n, 1, 0.1)$ and $\boldsymbol{c}_2 = \mathtt{randn}(n, 1)$ so that $\boldsymbol{a}_1(1) = \boldsymbol{a}_2(1), \boldsymbol{b}_1(1) = \boldsymbol{b}_2(1)$ and $\boldsymbol{c}_1(1) = \boldsymbol{c}_2(1)$. Moreover, we choose $f = \mathtt{randn}(n, 1)$ and $g = \mathtt{randn}(n, 1)$. We choose the parameters $w_i = 1$, for $i = 1, 2, \ldots, 5$.

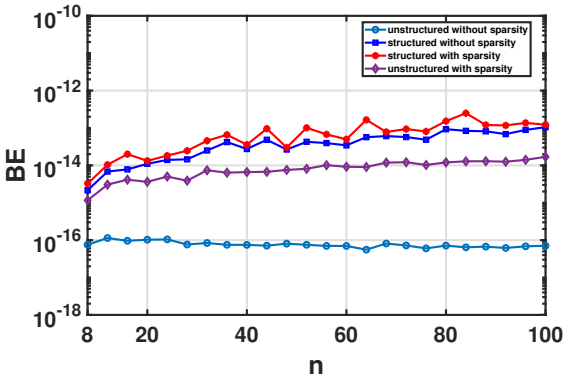


Figure 4.4.1: Different structured and unstructured BEs for $n = 8 : 4 : 100$.

We apply the GMRES method [123] with the initial guess vector zero and tolerance 10^{-7}. Let $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$ be the computed solution of the GSPP. For $n = 8 :$

4 : 100 in Figure 4.4.1, we plot the unstructured BE $\boldsymbol{\eta}(\widetilde{\boldsymbol{v}})$ using the formula (4.1.3) (denoted as '**unstructured without sparsity**'), structured BE $\boldsymbol{\eta}_{\text{sps}}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ (denoted as '**structured with sparsity**') using Theorem 4.1.7, $\boldsymbol{\eta}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ (denoted as '**structured without sparsity**') using Corollary 4.1.2 and $\boldsymbol{\eta}_{\text{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ (denoted as '**unstructured without sparsity**') using Theorem 4.1.12. From Figure 4.4.1, it can be observed that, for all values of n, the unstructured BE $\boldsymbol{\eta}(\widetilde{\boldsymbol{v}})$ around of order $\mathcal{O}(10^{-16})$ and all other BEs $\boldsymbol{\eta}_{\text{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$, $\boldsymbol{\eta}_{\text{sps}}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$, and $\boldsymbol{\eta}^{\mathcal{S}_2}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ are around of $\mathcal{O}(10^{-13})$ or less than that, which are very small. Therefore, the approximate solution computed using GMRES method for each generated Toeplitz structured GSPP effectively solves a nearby perturbed unstructured linear system as well as a nearby perturbed (sparsity preserving) Toeplitz structured GSPP.

Table 4.4.1: Unstructured and structured BEs for different values of n for Example 4.4.4.

n	$\boldsymbol{\eta}(\widetilde{\boldsymbol{v}})$	$\boldsymbol{\eta}_{\text{sps}}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$	$\boldsymbol{\eta}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$	$\boldsymbol{\eta}_{\text{sps}}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$
8	$1.5522e-16$	$5.1676e-16$	$1.3612e-15$	$4.0978e-15$
16	$7.9177e-16$	$2.6522e-15$	$1.1505e-15$	$8.3191e-15$
32	$4.2992e-15$	$1.1547e-13$	$1.3009e-13$	$6.2542e-13$
64	$8.4321e-16$	$4.8161e-14$	$1.2379e-13$	$4.2252e-13$
128	$1.7815e-15$	$3.4069e-14$	$1.5579e-13$	$6.1305e-13$

Example 4.4.4. In this example, we consider the symmetric-Toeplitz structured GSPP (4.1.1) with the block matrices $A = I_n$, $B = [b_{ij}] \in \mathcal{ST}_n$, where $b_{ij} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(i-j)^2}{2}}$, $i, j = 1, \ldots, n$, $D = -\mu I_n$, $f = \texttt{randn}(n, 1) \in \mathbb{R}^n$, and $g = \mathbf{0} \in \mathbb{R}^n$. We choose $\mu = 0.01$. To solve the GSPP, we use the CNAGSOR preconditioned GMRES (PGMRES) method [157]. We choose $w_1 = 1/\|A\|_F$, $w_2 = 1/\|B\|_F$, $w_3 = 1/\|D\|_F$, $w_4 = 1/\|f\|_2$, and $w_5 = 0$. For the computed approximate solution $\widetilde{\boldsymbol{v}} = [\widetilde{\boldsymbol{u}}^T, \widetilde{\boldsymbol{p}}^T]^T$, we compute the unstructured BE $\boldsymbol{\eta}(\widetilde{\boldsymbol{v}})$ using the formula (4.1.3), structured BE $\boldsymbol{\eta}_{\text{sps}}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ using Theorem 4.1.10, and the structured BE without preserving sparsity $\boldsymbol{\eta}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ using Corollary 4.1.3. The computed values are reported in Table 4.4.1.

We observe that the structured BEs $\boldsymbol{\eta}_{\text{sps}}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ and $\boldsymbol{\eta}^{\mathcal{S}_3}(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{p}})$ are all most all cases remains one or two order larger than the unstructured ones and remains within an order of $\mathcal{O}(10^{-13})$. Hence, we can conclude that the approximate solution $\widetilde{\boldsymbol{v}}$ obtained using the CNAGSOR PGMRES method for the tested GSPPs serves as an exact solution to

a nearly perturbed symmetric-Toeplitz structured GSPP, while preserving the sparsity pattern of the original problem.

Next, we carry out several numerical experiments to test the strong backward stability of numerical algorithms for solving the DSPP. We consider $\theta_1 = \frac{1}{\|A\|_F}$, $\theta_2 = \frac{1}{\|B\|_F}$, $\theta_3 = \frac{1}{\|F\|_F}$, $\theta_4 = \frac{1}{\|D\|_F}$, $\theta_5 = \frac{1}{\|C\|_F}$, $\theta_6 = \frac{1}{\|G\|_2}$, $\theta_7 = \frac{1}{\|E\|_2}$, $\theta_8 = \frac{1}{\|f\|_2}$, $\theta_9 = \frac{1}{\|g\|_2}$ and $\theta_{10} = \frac{1}{\|h\|_2}$.

Example 4.4.5. To test the strong backward stability of the GEP, we consider the DSPP (4.3.1) with

$$A = GPG(1:3, 1:3) \in \mathcal{S}_3,\ B = D = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 10^4 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{3\times 3},\ C = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 6 & 0 \\ 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{3\times 3},$$

$$E = GPG(4:6, 4:6) \in \mathcal{S}_3,\ f = \begin{bmatrix} 10^8 \\ 10 \\ 0 \end{bmatrix} \in \mathbb{R}^3 \text{ and } g = h = \begin{bmatrix} 10^{-8} \\ 0 \\ 0 \end{bmatrix} \in \mathbb{R}^3,$$

where $G = 10^6 \times \text{diag}(1, 5, 10, 50, 100, 500)$ and $P = [p_{ij}] \in \mathbb{R}^{6\times 6}$, $p_{ij} = \dfrac{(i+j-1)!}{(i-1)!(j-1)!}$. The approximate solution $\widetilde{\boldsymbol{w}} = [\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$ of this DSPP is obtained using the GEP, where

$$\widetilde{\boldsymbol{x}} = 10^{-05} \times \begin{bmatrix} 60.0120 \\ -8.0016 \\ 1.0002 \end{bmatrix},\ \widetilde{\boldsymbol{y}} = \begin{bmatrix} 6.0012 \\ 2.0004 \\ -2.0004 \end{bmatrix} \text{ and } \widetilde{\boldsymbol{z}} = 10^{-13} \times \begin{bmatrix} -1.7109 \\ 0.8556 \\ -0.0475 \end{bmatrix}.$$

We compute the unstructured BE $\boldsymbol{\eta}(\widetilde{\boldsymbol{w}})$, structured BEs $\boldsymbol{\eta}_{\mathbf{sps}}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ and $\boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ using the formulae given in (4.3.3), Theorem 4.3.5 and Corollary 4.3.1, respectively. The obtained BEs are given by

$$\begin{aligned} \boldsymbol{\eta}(\widetilde{\boldsymbol{w}}) =& 6.9314 \times 10^{-27},\ \boldsymbol{\eta}_{\mathbf{sps}}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = 5.4649 \times 10^{-06}, \\ &\text{and } \boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}}) = 4.7907 \times 10^{-06}. \end{aligned} \tag{4.4.1}$$

From (4.4.1), we can observe that $\boldsymbol{\eta}(\widetilde{\boldsymbol{w}})$ of $\mathcal{O}(10^{-27})$ indicates that GEP is backward stable for solving this DSPP. On the other side $\boldsymbol{\eta}_{\mathbf{sps}}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ and $\boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ is much smaller than $\boldsymbol{\eta}(\widetilde{\boldsymbol{w}})$ implies that GEP for solving this DSPP is not strongly backward stable. This shows that a backward stable iterative algorithm for solving the DSPP may not be strongly backward stable.

Example 4.4.6. In this example, we perform a comparison among our obtained structured BEs and the structured BE considered in [98]. For this, we consider the DSPP (4.3.1) with

$$A = \begin{bmatrix} 0.0968 & 0 & -0.2438 & -0.2823 \\ 0 & 0 & 1.1180 & -1.1611 \\ -0.2438 & 1.1180 & 1.6014 & -0.8693 \\ -0.2823 & -1.1611 & -0.8693 & -0.4914 \end{bmatrix}, \; B = \begin{bmatrix} 0 & 0 & 0.7090 & 0 \\ 1.9046 & 0.0928 & -0.0430 & 0.0508 \end{bmatrix},$$

$$F = \begin{bmatrix} -0.2592 & 0 & 0.2543 & 0.1248 \\ 0.0876 & 1.1375 & 0 & 0.0766 \end{bmatrix}, \; C = \mathbf{0}_{2\times 2}, \; D = \begin{bmatrix} 0 & 1.8070 \\ 1.0365 & -1.5516 \end{bmatrix} \text{ and } E = \mathbf{0}_{2\times 2}.$$

Here, $n = 4, m = 2$ and $p = 2$. Further, we consider the right-hand side vector $\mathbf{d} = [f^T, g^T, h^T]^T \in \mathbb{R}^8$, where

$$f = \begin{bmatrix} -1.1251 \\ -1.9000 \\ -0.4320 \\ -1.1422 \end{bmatrix}, \; g = \begin{bmatrix} -0.5516 \\ 1.8738 \end{bmatrix} \text{ and } h = \begin{bmatrix} 0.4982 \\ 0.8347 \end{bmatrix}.$$

The computed solution using the the MATLAB '*blackshash*' command is $\widetilde{\boldsymbol{w}} = [\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}^T]^T$, where

$$\widetilde{\boldsymbol{x}} = \begin{bmatrix} -1.6927 \\ -1.5778 \\ 1.9746 \\ 3.5598 \end{bmatrix}, \; \widetilde{\boldsymbol{y}} = \begin{bmatrix} 1.2180 \\ 0.2757 \end{bmatrix} \text{ and } \widetilde{\boldsymbol{z}} = \begin{bmatrix} 0.3571 \\ -1.8683 \end{bmatrix}.$$

The computed solution $\widetilde{\boldsymbol{w}}$ has residue $\|\mathfrak{B}\widetilde{\boldsymbol{w}} - \mathbf{d}\| = 1.4864 \times 10^{-15}$. The unstructured BE computed using the formula (4.3.3) is 5.2700×10^{-17}, structured BE using the Theorem 3.2 of [98] is 4.1137×10^{-16}, structured BE with sparsity using Theorem 4.3.8 is 2.7992×10^{-16} and the structured BE without sparsity using Corollary 4.3.3 is 2.5525×10^{-16}. We observe that all the computed BEs are in unit round-off error and the structured BEs are only one order larger than the unstructured ones. Furthermore, the structured BEs derived in our work and those obtained in the reference [98] exhibit uniform order. This shows the reliability of our derived structured BEs formulae. One notable advantage of our derived formulae lies in our ability to preserve the sparsity pattern within the perturbation matrices.

Example 4.4.7. To test the strong backward stability of the GMRES method, in this example, we consider the DSPP (4.3.1) [75] with the block matrices

$$A = \begin{bmatrix} I \otimes Z + Z \otimes I & \mathbf{0} \\ \mathbf{0} & I \otimes Z + Z \otimes I \end{bmatrix} \in \mathbb{R}^{2r^2 \times 2r^2}, \; B = \begin{bmatrix} I \otimes H & H \otimes I \end{bmatrix} \in \mathbb{R}^{r^2 \times 2r^2},$$

$D = G \otimes H \in \mathbb{R}^{r^2 \times r^2}$ and $C = E = \mathbf{0}_{r^2 \times r^2}$, where $Z = \frac{1}{(r+1)^2}\,\text{tridiag}(-1, 2, -1) \in \mathbb{R}^{r \times r}$, $H = \frac{1}{r+1}\,\text{tridiag}(0, 1, -1) \in \mathbb{R}^{r \times r}$ and $G = \text{diag}(1, r+1, \ldots, r^2 - r + 1) \in \mathbb{R}^{r \times r}$. For this problem, the dimension of $\mathfrak{B}$ is $4r^2$. We use GMRES method to solve this DSPP with termination criteria $\frac{\|\mathfrak{B}\boldsymbol{w}_k - \mathbf{d}\|_2}{\|\mathbf{d}\|_2} < \boldsymbol{tol}$, where $\boldsymbol{w}_k$ is solution at each iterate and $\boldsymbol{tol} = 10^{-13}$ and the initial guess vector zero. We compute the structured and unstructured BEs for the solution at the final iteration. The computed BEs for different values of r are listed in Table 4.4.2.

Table 4.4.2: Values of structured and unstructured BEs of the approximate solution obtained using GMRES for Example 4.4.7.

r	$\frac{\|\mathfrak{B}\boldsymbol{w}_k - \mathbf{d}\|_2}{\|\mathbf{d}\|_2}$	$\boldsymbol{\eta}(\tilde{w})$	$\boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$	$\boldsymbol{\eta}^{\mathbb{S}_1}_{\mathbf{sps}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$
4	1.0593e-15	4.1823e-17	1.3757e-16	4.9831e-16
6	2.4960e-14	5.3825e-16	1.8436e-15	7.3871e-15
8	2.0868e-14	3.0086e-16	9.6476e-16	5.4875e-15
10	3.2981e-14	3.4862e-16	1.3781e-15	9.1775e-15

From Table 4.4.2, we observe that unstructured BE $\boldsymbol{\eta}(\widetilde{\boldsymbol{w}})$, structured BE with preserving sparsity $\boldsymbol{\eta}^{\mathbb{S}_1}_{\mathbf{sps}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ and structured BE $\boldsymbol{\eta}^{\mathbb{S}_1}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{z}})$ are all around order of unit round-off error. Using our obtained structured BEs, we successfully demonstrate that the GMRES method for solving this DSPP exhibits strong backward stability.

4.5. Summary

In this chapter, we investigated the structured BEs for circulant, Toeplitz, symmetric-Toeplitz, and Hermitian structured GSPPs with and without preserving the sparsity pattern of block matrices. Moreover, we study structured BEs for DSPP in three cases when the diagonal block matrices preserve symmetric structure. Additionally, we provide minimal perturbation matrices for which an approximate solution becomes the exact solution of a nearly perturbed GSPP or DSPP, which preserves their inherent block structure and sparsity pattern. Furthermore, unstructured BE is obtained when the block matrices of

the GSPP only preserve the sparsity pattern. Our obtained results are used to derive structured BE for WRLS problems with Toeplitz or symmetric-Toeplitz coefficient matrices. Numerical experiments are performed to validate our theoretical findings and to examine the backward stability and the strong backward stability of numerical algorithms to solve structured GSPPs and DSPPs.

CHAPTER 5

Partial Condition Numbers for Saddle Point Problems*†

This chapter addresses structured normwise condition number (NCN), mixed condition number (MCN), and componentwise condition number (CCN) for a linear function of the solution (or the partial NCN, MCN, and CCN) of the GSPP and DSPP. Firstly, we present a general framework that enables us to measure the structured CNs of the individual components of the solution of the GSPP. Then, we derive their explicit formulae when the input matrices have symmetric, Toeplitz, or some general linear structures. In addition, compact formulae for the unstructured CNs are obtained, which recover previous results on CNs for GSPPs for specific choices of the linear function. Moreover, we investigate unstructured partial unified CN and structured partial NCN, MCN, and CCN for DSPPs. Furthermore, applications of the derived structured CNs are provided to determine the structured CNs for the WRLS problems, Tikhonov regularization problems and EILS problems, which retrieves some previous studies in the literature.

5.1. Partial Condition Numbers for the Generalized Saddle Point Problem

In this section, we consider the following GSPP:

$$\mathcal{M}z := \begin{bmatrix} A & B^T \\ C & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} =: \mathbf{b}, \tag{5.1.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B, C \in \mathbb{R}^{m\times n}$, $D \in \mathbb{R}^{n\times n}$, $x, f \in \mathbb{R}^n$, and $y, g \in \mathbb{R}^m$. The block matrices A, B, C and D satisfy some special structures, such as $B = C$, symmetric, Toeplitz, or have some other linear structures [26]. Recently, a large amount of efficient iteration methods have been proposed to solve the linear system (5.1.1), such as inexact Uzawa schemes [12], Krylov subspace methods [119], and so on.

* S. S. Ahmad and **P. Khatun**, "Structured condition numbers for a linear function of the solution of the generalized saddle point problems." *Electronic Transactions on Numerical Analysis,* 60:471-500, 2024.

† S. S. Ahmad and **P. Khatun**, "Partial condition numbers for double saddle point problems." *Under Revision in Numerical Algorithms.*

Perturbation analysis and CNs for the GSPP (5.1.1) have been widely studied in the literature. A brief review of the literature work of the CNs for the GSPP (5.1.1) is as follows. Wang and Liu [138] have analyzed the NCN for the solution $z = [x^T, y^T]^T$ for the KKT system, i.e., the GSPP (5.1.1) with $A = A^T$, $B = C$ and $D = \mathbf{0}$. In [147], authors have discussed perturbation bounds for the GSPP when $B = C$, and $D = \mathbf{0}$, and have derived closed formulae for the NCN, MCN, and CCN of the solutions $z = [x^T, y^T]^T$ and the individual solution components x and y. The NCN and perturbation bounds have been investigated in [151] for the solution $z = [x^T, y^T]^T$ of the GSPP (5.1.1), with the conditions $B = C$ and $D \neq \mathbf{0}$. Later, Meng and Li [100] studied the MCN and CCN for $z = [x^T, y^T]^T$. Additionally, they explored the NCN, MCN, and CCN for the individual solution components x and y. Recently, new perturbation bounds have been derived for the GSPP (5.1.1) under the condition $B \neq C$, without imposing any special structure on A and D [153].

In many applications, blocks of the coefficient matrix $\mathcal{M}$ of the system (5.1.1) exhibit linear structures (for example, symmetric, Toeplitz or symmetric-Toeplitz) [32, 60, 124, 163]. Therefore, it is reasonable to ask: how sensitive is the solution when structure-preserving perturbations are introduced to the coefficient matrix of GSPPs? To address the aforementioned query, we explore the notion of structured CNs by restricting perturbations that preserve the structures inherent in the block matrices of $\mathcal{M}$.

Furthermore, in many instances, x and y represent two distinct physical entities; for example, in the Stokes equation, x denotes the velocity vector, and y signifies the scalar pressure field [60]. Therefore, it is important to assess their individual conditioning properties. The traditional CNs lack the ability to reveal the conditioning of a specific part of the solution. To tackle this situation, CN of a linear function of the solution has been investigated in the literature. This is referred to as the partial CN. In this study, we propose a general framework for assessing the conditioning of x, y, $z = [x^T,\ y^T]^T$ and each component of z. In the proposed general framework, we consider the structured CNs of a linear function $\mathbf{L}[x^T,\ y^T]^T$ of the solution to GSPP (5.1.1), where $\mathbf{L} \in \mathbb{R}^{k\times(m+n)}$. The matrix $\mathbf{L}$ serves as a pivotal tool for the purpose of selecting solution components. For example, (i) $\mathbf{L} = I_{m+n}$ gives the CNs for $[x^T,\ y^T]^T$, (ii) $\mathbf{L} = \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}$ gives the CNs for x, and (iii) $\mathbf{L} = \begin{bmatrix} \mathbf{0} & I_m \end{bmatrix}$ gives the CNs for y.

The key contributions of this section are summarized as follows:

- We study the NCN, MCN, and CCN for the linear function $\mathbf{L}[x^T, y^T]^T$, which in turn provides a general framework, enabling us to derive CNs for the solutions $[x^T, y^T]^T$, x, y, and each component of $[x^T, y^T]^T$.
- We investigate partial unstructured CNs by considering $B = C$ and then the structured CNs when the (1,1) block A is symmetric and (1,2) block B is Toeplitz, and derive their closed form expressions. Moreover, explicit formulae for unstructured CNs are also derived. For appropriate choices of $\mathbf{L}$, we have shown that our derived unstructured CNs formulae generalize the results given in literature [100, 151].
- By considering linear structure on the block matrices A and D with $B \neq C$, we provide compact formulae of the structured partial NCN, MCN, and CCN for the GSPP (5.1.1).
- Utilizing the structured CNs formulae, we derive the structured CNs for the WRLS problem and generalize some of the previous structured CNs formulae for the Tikhonov regularization problem. This shows the generic nature of our obtained results.
- Numerical experiments demonstrate that the obtained structured CNs offer sharper bounds to the actual relative errors than their unstructured counterparts.

The organization of this section is as follows. Subsection 5.1.1 discusses notation and preliminary results about CNs. In Subsections 5.1.2-5.1.4, we investigate unstructured and structured partial NCN, MCN, and CCN for the GSPP. Furthermore, an application of our obtained structured CNs is provided in Subsection 5.1.5 for WRLS problems and Tikhonov regularization problems. In Subsection 5.1.6, numerical experiments are carried out to demonstrate the effectiveness of the proposed structured CNs. Subsection 5.1.7 presents some concluding remarks.

5.1.1. Preliminaries

In this subsection, we define some notation and review some well-known results, which play a crucial role in constructing the main findings of this section.

Following [51, 88], the entrywise division of any two vectors $x = [x_i] \in \mathbb{R}^n$ and $y = [y_i] \in \mathbb{R}^n$ is defined as $\frac{x}{y} := \left[\frac{x_i}{y_i}\right]$, where $x_i/0 = 0$ whenever $x_i = 0$ and infinity otherwise.

Throughout this section, we assume that A and $\mathcal{M}$ are nonsingular. We know that if A is nonsingular, then $\mathcal{M}$ is nonsingular if and only if its Schur complement $S =$

$D - CA^{-1}B^T$ is nonsingular [8] and its inverse is expressed as follows:

$$\mathcal{M}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B^T S^{-1} C A^{-1} & -A^{-1}B^T S^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{bmatrix}. \tag{5.1.2}$$

First, consider the case when $B = C$, i.e., the following GSPP:

$$\mathcal{M} \begin{bmatrix} x \\ y \end{bmatrix} := \begin{bmatrix} A & B^T \\ B & D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} := \mathbf{b}, \tag{5.1.3}$$

and let ΔA, $\Delta B, \Delta D, \Delta f$ and Δg be the perturbations in A, B, D, f and C, respectively. Then, we have the following perturbed problem of (5.1.3):

$$(\mathcal{M} + \Delta\mathcal{M}) \begin{bmatrix} x + \Delta x \\ y + \Delta y \end{bmatrix} = \begin{bmatrix} A + \Delta A & (B + \Delta B)^T \\ B + \Delta B & D + \Delta D \end{bmatrix} \begin{bmatrix} x + \Delta x \\ y + \Delta y \end{bmatrix} = \begin{bmatrix} f + \Delta f \\ g + \Delta g \end{bmatrix}, \tag{5.1.4}$$

which has the unique solution $\begin{bmatrix} x + \Delta x \\ y + \Delta y \end{bmatrix}$ when $\|\mathcal{M}^{-1}\|_2 \|\Delta\mathcal{M}\|_2 < 1$. Now, from (5.1.4) omitting the higher order term, we obtain

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \approx \mathcal{M}^{-1} \begin{bmatrix} \Delta f \\ \Delta g \end{bmatrix} - \mathcal{M}^{-1} \begin{bmatrix} \Delta A & \Delta B^T \\ \Delta B & \Delta D \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{5.1.5}$$

Using the properties Kronecker product and vec operation in (1.3.2), we have the following important lemma.

Lemma 5.1.1. *Let* $\begin{bmatrix} x \\ y \end{bmatrix}$ *and* $\begin{bmatrix} x + \Delta x \\ y + \Delta y \end{bmatrix}$ *be the unique solutions of the GSPP* (5.1.3) *and* (5.1.4)*, respectively. Then, we have the following perturbation expression:*

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \approx -\mathcal{M}^{-1} \begin{bmatrix} \mathcal{R} & -I_{m+n} \end{bmatrix} \begin{bmatrix} \mathrm{vec}(\Delta A) \\ \mathrm{vec}(\Delta B) \\ \mathrm{vec}(\Delta D) \\ \Delta f \\ \Delta g \end{bmatrix},$$

where

$$\mathcal{R} = \begin{bmatrix} x^T \otimes I_n & I_n \otimes y^T & \mathbf{0} \\ \mathbf{0} & x^T \otimes I_m & y^T \otimes I_m \end{bmatrix}. \tag{5.1.6}$$

Proof. The proof follows from (5.1.5) and using the properties in (1.3.2). ■

Denote $\mathbf{H} = \begin{bmatrix} A & \mathbf{0} \\ B & D \end{bmatrix}$ and $\Delta\mathbf{H} = \begin{bmatrix} \Delta A & \mathbf{0} \\ \Delta B & \Delta D \end{bmatrix}$. Xu and Li [151] investigated unstructured NCN and Meng and Li [100] studied unstructured MCN and NCN for the solution $[x^T, y^T]^T$ to the GSPP (5.1.1) when $B = C$, which are given as follows:

$$\begin{aligned} \mathscr{K}^u([x^T, y^T]^T) &:= \limsup_{\eta\to 0} \left\{ \frac{\|[\Delta x^T, \Delta y^T]^T\|_2}{\eta\|[x^T, y^T]^T\|_2} : \left\| \begin{bmatrix} \Delta\mathbf{H} & \Delta\mathbf{d} \end{bmatrix} \right\|_F \le \eta \left\| \begin{bmatrix} \mathbf{H} & \mathbf{d} \end{bmatrix} \right\|_F \right\} \\ &= \frac{\left\| \mathcal{M}^{-1} \begin{bmatrix} \mathcal{R} & -I_{m+n} \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} \mathbf{H} & \mathbf{d} \end{bmatrix} \right\|_F}{\|[x^T, y^T]^T\|_2}, \end{aligned} \tag{5.1.7}$$

$$\begin{aligned} \mathscr{M}^u([x^T, y^T]^T) &:= \limsup_{\eta\to 0} \left\{ \frac{\|[\Delta x^T, \Delta y^T]^T\|_\infty}{\eta\|[x^T, y^T]^T\|_\infty} \left| \begin{bmatrix} \Delta\mathbf{H} & \Delta\mathbf{d} \end{bmatrix} \right| \le \eta \left| \begin{bmatrix} \mathbf{H} & \mathbf{d} \end{bmatrix} \right| \right\} \\ &= \frac{\left\| |\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} \mathrm{vec}(|A|) \\ \mathrm{vec}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} + |\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_\infty}{\|[x^T, y^T]^T\|_\infty}, \end{aligned} \tag{5.1.8}$$

and

$$\begin{aligned} \mathscr{C}^u([x^T, y^T]^T) &:= \limsup_{\eta\to 0} \left\{ \frac{1}{\eta} \left\| \frac{[\Delta x^T, \Delta y^T]^T}{[x^T, y^T]^T} \right\|_\infty : \left| \begin{bmatrix} \Delta\mathbf{H} & \Delta\mathbf{d} \end{bmatrix} \right| \le \eta \left| \begin{bmatrix} \mathbf{H} & \mathbf{d} \end{bmatrix} \right| \right\} \\ &= \left\| \mathfrak{D}^\dagger_{[x^T, y^T]^T} |\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} \mathrm{vec}(|A|) \\ \mathrm{vec}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} + \mathfrak{D}^\dagger_{[x^T, y^T]^T} |\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_\infty, \end{aligned} \tag{5.1.9}$$

where $\mathcal{R}$ is defined as in (5.1.6).

In the next subsection, we discuss the unstructured partial CNs for the GSPP (5.1.3).

5.1.2. Partial CNs for the GSPP when $B = C$

In this subsection, we consider the unstructured NCN, MCN, and CCN for the linear function $\mathbf{L}[x^T, y^T]^T$, where $\mathbf{L} \in \mathbb{R}^{k\times(m+n)}$ when $B = C$, and derive their explicit formulae. Throughout the section, we assume $[x^T, y^T]^T \neq \mathbf{0}$ for MCN and $x_i \neq 0\,(i = 1, \ldots, m)$ and $y_i \neq 0\,(i = 1, \ldots, n)$ for CCN. In the following, we define unstructured partial NCN, MCN, and CCN.

Definition 5.1.1. *Let* $[x^T,\, y^T]^T$ *and* $\left[(x+\Delta x)^T,\, (y+\Delta y)^T\right]^T$ *be the unique solutions of GSPPs* (5.1.3) *and* (5.1.4)*, respectively, and* $\mathbf{L} \in \mathbb{R}^{k\times(m+n)}$*. Then we define the unstructured partial NCN, MCN, and CCN, respectively, as follows:*

$$\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T) := \limsup_{\eta\to 0} \left\{ \frac{\|\mathbf{L}[\Delta x^T,\, \Delta y^T]^T\|_2}{\eta\|\mathbf{L}[x^T,\, y^T]^T\|_2} : \left\|\begin{bmatrix}\Delta\mathbf{H} & \Delta\mathbf{d}\end{bmatrix}\right\|_F \le \eta \left\|\begin{bmatrix}\mathbf{H} & \mathbf{d}\end{bmatrix}\right\|_F \right\},$$

$$\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T) := \limsup_{\eta\to 0} \left\{ \frac{\|\mathbf{L}[\Delta x^T,\, \Delta y^T]^T\|_\infty}{\eta\|\mathbf{L}[x^T,\, y^T]^T\|_\infty} : \left|\begin{bmatrix}\Delta\mathbf{H} & \Delta\mathbf{d}\end{bmatrix}\right| \le \eta \left|\begin{bmatrix}\mathbf{H} & \mathbf{d}\end{bmatrix}\right| \right\}, \quad \textit{and}$$

$$\mathscr{C}(\mathbf{L}[x^T,\, y^T]^T) := \limsup_{\eta\to 0} \left\{ \frac{1}{\eta} \left\| \frac{\mathbf{L}[\Delta x^T,\, \Delta y^T]^T}{\mathbf{L}[x^T,\, y^T]^T} \right\|_\infty : \left|\begin{bmatrix}\Delta\mathbf{H} & \Delta\mathbf{d}\end{bmatrix}\right| \le \eta \left|\begin{bmatrix}\mathbf{H} & \mathbf{d}\end{bmatrix}\right| \right\}.$$

Note that when $\mathbf{L} = I_{m+n}$, the above definition reduces to (5.1.7)–(5.1.9). For using Lemma 6.2.1, we construct the mapping $\boldsymbol{\psi} : \mathbb{R}^{m^2+mn+n^2} \times \mathbb{R}^{m+n} \mapsto \mathbb{R}^{m+n}$ by

$$\boldsymbol{\psi}([\Omega^T, f^T, g^T]^T) := \mathbf{L}\begin{bmatrix}x\\ y\end{bmatrix} = \mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}f\\ g\end{bmatrix}, \tag{5.1.10}$$

where $\Omega^T = [\mathrm{vec}(A)^T, \mathrm{vec}(B)^T, \mathrm{vec}(D)^T]^T$.

The following result is crucial for finding the CNs formulae.

Proposition 5.1.2. *Let* $\Omega^T = [\mathrm{vec}(A)^T, \mathrm{vec}(B)^T, \mathrm{vec}(D)^T]^T$*. Then, for the map* $\boldsymbol{\psi}$ *defined in* (5.1.10)*, we have* $\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T) = \mathscr{K}(\boldsymbol{\psi}, [\Omega^T, f^T, g^T]^T)$*,* $\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T) = \mathscr{M}(\boldsymbol{\psi}, [\Omega^T, f^T, g^T]^T)$*, and* $\mathscr{C}(\mathbf{L}[x^T,\, y^T]^T) = \mathscr{C}(\boldsymbol{\psi}, [\Omega^T, f^T, g^T]^T)$*.*

Proof. Let $\Delta\Omega^T = [\mathrm{vec}(\Delta A)^T, \mathrm{vec}(\Delta B)^T, \mathrm{vec}(\Delta D)^T]^T$. Then, from (5.1.10), we obtain

$$\begin{aligned}&\boldsymbol{\psi}\left([(\Omega^T + \Delta\Omega^T), f^T + \Delta f^T, g^T + \Delta g^T]^T\right) - \boldsymbol{\psi}\left([\Omega^T,\, f^T,\, g^T]^T\right)\\ &= \mathbf{L}\begin{bmatrix}x+\Delta x\\ y+\Delta y\end{bmatrix} - \mathbf{L}\begin{bmatrix}x\\ y\end{bmatrix} = \mathbf{L}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}.\end{aligned} \tag{5.1.11}$$

Now, in Definition 5.1.1, substituting (5.1.11) and $\boldsymbol{\psi}\left([\Omega^T,\, f^T,\, g^T]^T\right) = \mathbf{L}\left[x^T,\, y^T\right]^T$, and consequently from Definition 4.1.1, the proof follows. ■

Since the *Fréchet* derivative of $\boldsymbol{\psi}$ (denoted by $\mathbf{d}\boldsymbol{\psi}$) has a pivotal role in estimating the CNs in Definition 5.1.1, it is essential to derive simple expressions for $\mathbf{d}\boldsymbol{\psi}$. By applying Lemma 5.1.1, we obtain the following results for $\mathbf{d}\boldsymbol{\psi}$.

Lemma 5.1.3. *The map* $\boldsymbol{\psi}$ *defined above is continuous and Fréchet differentiable at* $[\Omega^T, f^T, g^T]^T$ *and its Fréchet derivative at* $[\Omega^T, f^T, g^T]^T$ *is given by*

$$\mathbf{d}\boldsymbol{\psi}([\Omega^T, f^T, g^T]^T) = -\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}.$$

Proof. Since $\mathcal{M}^{-1}$ is continuous in its elements, the linear map $\boldsymbol{\psi}$ is also continuous. Let $\Delta\Omega^T = [\mathrm{vec}(\Delta A)^T, \mathrm{vec}(\Delta B)^T, \mathrm{vec}(\Delta D)^T]^T$. Then

$$\boldsymbol{\psi}\left([(\Omega^T + \Delta\Omega^T), f^T + \Delta f^T, g^T + \Delta g^T]^T\right) - \boldsymbol{\psi}\left([\Omega^T,\ f^T,\ g^T]^T\right) = \mathbf{L}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}.$$

Hence, the rest of the proof follows from the Lemma 5.1.1. ■

Applying Lemma 5.1.3, we obtain the following closed formulae for the unstructured CNs for the linear function $\mathbf{L}[x^T,\ y^T]^T$.

Theorem 5.1.4. *Let $[x^T,\ y^T]^T$ be the unique solution of the GSPP (5.1.3). Then the unstructured partial NCN, MCN, and CCN, respectively, are given by*

$$\mathscr{K}(\mathbf{L}[x^T,\ y^T]^T) = \frac{\left\|\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\mathbf{H} & \mathbf{d}\end{bmatrix}\right\|_F}{\|\mathbf{L}[x^T,\ y^T]^T\|_2},$$

$$\mathscr{M}(\mathbf{L}[x^T,\ y^T]^T) = \frac{\left\|\,|\mathbf{L}\mathcal{M}^{-1}\mathcal{R}|\begin{bmatrix}\mathrm{vec}(|A|)\\ \mathrm{vec}(|B|)\\ \mathrm{vec}(|D|)\end{bmatrix} + |\mathbf{L}\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}\right\|_\infty}{\|\mathbf{L}[x^T,\ y^T]^T\|_\infty}, \quad \textit{and}$$

$$\mathscr{C}(\mathbf{L}[x^T,\ y^T]^T) = \left\|\mathfrak{D}^\dagger_{\mathbf{L}[x^T,\,y^T]^T}|\mathbf{L}\mathcal{M}^{-1}\mathcal{R}|\begin{bmatrix}\mathrm{vec}(|A|)\\ \mathrm{vec}(|B|)\\ \mathrm{vec}(|D|)\end{bmatrix} + \mathfrak{D}^\dagger_{\mathbf{L}[x^T,\,y^T]^T}|\mathbf{L}\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}\right\|_\infty.$$

Proof. Let $\Omega^T = [\mathrm{vec}(A)^T, \mathrm{vec}(B)^T, \mathrm{vec}(D)^T]^T$. Then, from Proposition 5.1.2 and applying the NCN formula of Lemma 1.3.2 for the map $\boldsymbol{\psi}$, we obtain

$$\mathscr{K}(\mathbf{L}[x^T,\ y^T]^T) = \mathscr{K}(\boldsymbol{\psi}, [\Omega^T, f^T, g^T]^T) = \frac{\left\|\mathbf{d}\boldsymbol{\psi}\left(\Omega^T,\ f^T,\ g^T\right]^T\right)\right\|_2 \left\|[\Omega^T, f^T, g^T]^T\right\|_2}{\left\|\boldsymbol{\psi}\left(\Omega^T,\ f^T,\ g^T\right]^T\right)\|_2}. \tag{5.1.12}$$

Now, substituting the *Fréchet* derivative expression of $\boldsymbol{\psi}$ at $[\Omega^T, f^T, g^T]^T$ provided in Lemma 5.1.3 in (5.1.12), we get

$$\mathscr{K}(\mathbf{L}[x^T,\ y^T]^T) = \frac{\left\|\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\mathbf{H} & \mathbf{d}\end{bmatrix}\right\|_F}{\|\mathbf{L}[x^T,\ y^T]^T\|_2}.$$

Similarly, applying the MCN formula provided in Lemma 1.3.2 for $\boldsymbol{\psi}$, we get

$$\mathscr{M}(\mathbf{L}[x^T,\ y^T]^T) = \mathscr{M}(\boldsymbol{\psi}, [\Omega^T, f^T, g^T]^T) = \frac{\left\|\,|\mathbf{d}\boldsymbol{\psi}\left([\Omega^T,\ f^T,\ g^T]^T\right)|\,|[\Omega^T, f^T, g^T]^T|\right\|_\infty}{\left\|\boldsymbol{\psi}\left([\Omega^T,\ f^T,\ g^T]^T\right)\right\|_\infty}. \tag{5.1.13}$$

Substituting the *Fréchet* derivative expression provided in Lemma 5.1.3 in (5.1.13), we obtain

$$\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T) = \frac{\left\| \left| \mathbf{L}\mathcal{M}^{-1} \begin{bmatrix} \mathcal{R} & -I_{m+n} \end{bmatrix} \right| \left| [\Omega^T, f^T, g^T]^T \right| \right\|_\infty}{\|\mathbf{L}[x^T,\, y^T]^T\|_\infty}$$
$$= \frac{\left\| |\mathbf{L}\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} \mathrm{vec}(|A|) \\ \mathrm{vec}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} + |\mathbf{L}\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_\infty}{\left\| \mathbf{L}[x^T,\, y^T]^T \right\|_\infty}.$$

Similarly, the rest of the proof follows. ■

Remark 5.1.5. *If we consider* $\mathbf{L} = I_{m+n}$, *then the formulae of* $\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T)$, $\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T)$ *and* $\mathscr{C}(\mathbf{L}[x^T,\, y^T]^T)$ *reduces to unstructured CNs* $\mathscr{K}^u([x^T,\, y^T]^T)$, $\mathscr{M}^u([x^T,\, y^T]^T)$ *and* $\mathscr{C}^u([x^T,\, y^T]^T)$ *given in* (5.1.7)–(5.1.9), *respectively. Moreover, if we choose* $\mathbf{L} = \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}$ *and* $\mathbf{L} = \begin{bmatrix} \mathbf{0} & I_m \end{bmatrix}$, *and after some easy calculations, we can recover the unstructured CNs formulae of [100] for* x *and* y, *respectively.*

5.1.3. Structured Partial CNs when A is Symmetric and $B = C$ is Toeplitz

In this subsection, we consider the structured partial NCN, MCN, and CCN of the GSPP (5.1.3) with $A = A^T$ and $B = C \in \mathbb{R}^{n\times m}$ is a Toeplitz matrix. We denote $\mathcal{S}_n$ and $\mathcal{T}_{n\times m}$ as set of all $n \times n$ symmetric matrices and $n \times m$ Toeplitz matrices, respectively.

As $\dim(\mathcal{T}_{n\times m}) = m + n - 1$, consider the basis $\{\mathcal{J}_i\}_{i=-n+1}^{m-1}$ for $\mathcal{T}_{n\times m}$ defined as

$$\mathcal{J}_i = \begin{cases} \mathcal{T}([(e_{n-i}^n)^T,\, \mathbf{0}]^T) & \text{for} \quad i = -n+1, \ldots, -1, 0, \\ \mathcal{T}([\mathbf{0},\, (e_i^m)^T]^T) & \text{for} \quad i = 1, \ldots, m-1. \end{cases}$$

Moreover, construct the diagonal matrix $\mathfrak{D}_{\mathcal{T}_{nm}} \in \mathbb{R}^{(m+n-1)\times(m+n-1)}$ with $\mathfrak{D}_{\mathcal{T}_{nm}}(j,j) = \boldsymbol{a}_j$, where

$$\boldsymbol{a} = [1, \sqrt{2}, \ldots, \sqrt{n-1}, \sqrt{\min\{m,n\}}, \sqrt{m-1}, \ldots, \sqrt{2}, 1]^T \in \mathbb{R}^{m+n-1}$$

such that $\|T\|_F = \|\mathfrak{D}_{\mathcal{T}_{nm}} \mathrm{vec}_{\mathcal{T}}(T)\|_2$.

Lemma 5.1.6. *Let* $T \in \mathcal{T}_{n\times m}$, *then* $\mathrm{vec}(T) = \mathbf{\Phi}_{\mathcal{T}_{nm}} \mathrm{vec}_{\mathcal{T}}(T)$, *where*

$$\mathbf{\Phi}_{\mathcal{T}_{nm}} = \left[\mathrm{vec}(\mathcal{J}_{-n+1}), \ldots, \mathrm{vec}(\mathcal{J}_{m-1}) \right] \in \mathbb{R}^{mn\times(m+n-1)}.$$

Proof. Assume that $\text{vec}_{\mathcal{T}}(T) = [t_{-n+1}, \dots t_0, \dots, t_{m-1}]^T$, then

$$T = \sum_{i=-n+1}^{m-1} t_i \mathcal{J}_i \iff \text{vec}(T) = \mathbf{\Phi}_{\mathcal{T}_{nm}} \text{vec}_{\mathcal{T}}(T).$$

Hence, the proof follows. ■

Let $A \in \mathcal{S}_n$, then $A = A^T$. Moreover, we have $\dim(\mathcal{S}_n) = \frac{n(n+1)}{2} =: \boldsymbol{p}$. We denote the generator vector for A as

$$\text{vec}_{\mathcal{S}}(A) := [a_{11}, \dots, a_{1n}, a_{22}, \dots, a_{2n}, \dots, a_{(n-1)(n-1)}, a_{(n-1)n}, a_{nn}]^T \in \mathbb{R}^{\boldsymbol{p}}.$$

Consider the basis $\left\{S_{ij}^n\right\}$ for $\mathcal{S}_n$ defined as

$$S_{ij}^n = \begin{cases} e_i^n (e_j^n)^T + (e_j^n e_i^n)^T & \text{for} \quad i \neq j, \\ e_i^n (e_i^n)^T & \text{for} \quad i = j, \end{cases}$$

where $1 \leq i \leq j \leq n$. Then, we have the following immediate result for vec-structure of A.

Lemma 5.1.7. *Let $A \in \mathcal{S}_n$, then $\text{vec}(A) = \mathbf{\Phi}_{\mathcal{S}_n} \text{vec}_{\mathcal{S}}(A)$, where $\mathbf{\Phi}_{\mathcal{S}_n} \in \mathbb{R}^{n^2 \times \boldsymbol{p}}$ is given by*

$$\mathbf{\Phi}_{\mathcal{S}_n} = \begin{bmatrix} \text{vec}(S_{11}^n) & \cdots & \text{vec}(S_{1n}^n) & \text{vec}(S_{22}^n) & \cdots & \text{vec}(S_{2n}^n) & \cdots & \text{vec}(S_{(n-1)n}^n) & \text{vec}(S_{nn}^n) \end{bmatrix}.$$

Proof. The proof follows by using the similar proof method of Lemma 5.1.6. ■

We construct the diagonal matrix $\mathfrak{D}_{\mathcal{S}_n} \in \mathbb{R}^{\boldsymbol{p} \times \boldsymbol{p}}$, where

$$\begin{cases} \mathfrak{D}_{\mathcal{S}_n}(j,j) = 1 & \text{for} \quad j = \frac{(2n-(i-2))(i-1)}{2} + 1,\ i = 1, 2, \dots, n, \\ \mathfrak{D}_{\mathcal{S}_n}(j,j) = \sqrt{2} & \text{for} \quad \text{otherwise.} \end{cases}$$

This matrix satisfies the property $\|A\|_F = \|\mathfrak{D}_{\mathcal{S}_n} \text{vec}_{\mathcal{S}}(A)\|_2$.

Consider the set

$$\mathcal{E} = \left\{ \mathbf{H} = \begin{bmatrix} A & \mathbf{0} \\ B & D \end{bmatrix} : A \in \mathcal{S}_n, B \in \mathcal{T}_{m \times n}, D \in \mathbb{R}^{m \times m} \right\},$$

and let $\Delta\mathbf{H} = \begin{bmatrix} \Delta A & \mathbf{0} \\ \Delta B & \Delta D \end{bmatrix} \in \mathcal{E}$, i.e., $\Delta A \in \mathcal{S}_n$, $\Delta B \in \mathcal{T}_{m \times n}$, and $\Delta D \in \mathbb{R}^{m \times m}$.

Next, we define the structured CNs for the solution of the GSPP (5.1.3).

Definition 5.1.2. *Let $[x^T, y^T]^T$ and $\left[(x + \Delta x)^T, (y + \Delta y)^T\right]^T$ be the unique solutions of GSPPs (5.1.3) and (5.1.4), respectively, with the structure $\mathcal{E}$ and $\mathbf{L} \in \mathbb{R}^{k \times (m+n)}$. Then,*

the structured partial NCN, MCN, and CCN are defined as follows:

$$\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) := \limsup_{\eta\to 0}\left\{\frac{\|\mathbf{L}[\Delta x^T,\, \Delta y^T]^T\|_2}{\eta\|\mathbf{L}[x^T,\, y^T]^T\|_2} : \left\|\begin{bmatrix}\Delta\mathbf{H} & \Delta\mathbf{b}\end{bmatrix}\right\|_F \leq \eta\left\|\begin{bmatrix}\mathbf{H} & \mathbf{d}\end{bmatrix}\right\|_F, \Delta\mathbf{H}\in\mathcal{E}\right\},$$

$$\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) := \limsup_{\eta\to 0}\left\{\frac{\|\mathbf{L}[\Delta x^T,\, \Delta y^T]^T\|_\infty}{\eta\|\mathbf{L}[x^T,\, y^T]^T\|_\infty} : \left|\begin{bmatrix}\Delta\mathbf{H} & \Delta\mathbf{b}\end{bmatrix}\right| \leq \eta\left|\begin{bmatrix}\mathbf{H} & \mathbf{b}\end{bmatrix}\right|, \Delta\mathbf{H}\in\mathcal{E}\right\},$$

$$\mathscr{C}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) := \limsup_{\eta\to 0}\left\{\frac{1}{\eta}\left\|\frac{\mathbf{L}[\Delta x^T,\, \Delta y^T]^T}{\mathbf{L}[x^T,\, y^T]^T}\right\|_\infty : \left|\begin{bmatrix}\Delta\mathbf{H} & \Delta\mathbf{b}\end{bmatrix}\right| \leq \eta\left|\begin{bmatrix}\mathbf{H} & \mathbf{b}\end{bmatrix}\right|, \Delta\mathbf{H}\in\mathcal{E}\right\}.$$

To find the structured CNs formulae by employing Lemma 1.3.2, we define the following mapping

$$\boldsymbol{\zeta} :\, \mathbb{R}^{\boldsymbol{l}}\times\mathbb{R}^n\times\mathbb{R}^m \mapsto \mathbb{R}^{m+n} \quad \text{by} \tag{5.1.14}$$

$$\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right) = \mathbf{L}\begin{bmatrix}x\\ y\end{bmatrix} = \mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}f\\ g\end{bmatrix},$$

where $\boldsymbol{l} = \boldsymbol{p} + m^2 + m + n - 1$, $\boldsymbol{w} = \begin{bmatrix}\mathrm{vec}_{\mathcal{S}}(A)\\ \mathrm{vec}_{\mathcal{T}}(B)\\ \mathrm{vec}(D)\end{bmatrix}$ and $\mathfrak{D}_{\mathcal{E}} = \begin{bmatrix}\mathfrak{D}_{\mathcal{S}_n} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathfrak{D}_{\mathcal{T}_{mn}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & I_{m^2}\end{bmatrix}$.

In the next lemma, we provide *Fréchet* derivative of the map $\boldsymbol{\zeta}$ at $\left[\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T\right]^T$.

Lemma 5.1.8. *The mapping $\boldsymbol{\zeta}$ defined in (5.1.14) is continuously Fréchet differentiable at $\left[\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T\right]^T$ and the Fréchet derivative is given by*

$$\mathbf{d}\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right) = -\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\boldsymbol{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n}\end{bmatrix},$$

where $\boldsymbol{\Phi}_{\mathcal{E}} = \begin{bmatrix}\boldsymbol{\Phi}_{\mathcal{S}_n} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Phi}_{\mathcal{T}_{mn}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & I_{m^2}\end{bmatrix}$.

Proof. The continuity of the linear map $\boldsymbol{\zeta}$ follows from the continuity of $\mathcal{M}^{-1}$. For the second part, let $\Delta\boldsymbol{w} = \begin{bmatrix}\mathrm{vec}_{\mathcal{S}}(\Delta A)\\ \mathrm{vec}_{\mathcal{T}}(\Delta B)\\ \mathrm{vec}(\Delta D)\end{bmatrix}$ and consider

$$\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}(\boldsymbol{w}^T+\Delta\boldsymbol{w}^T), f^T+\Delta f^T, g^T+\Delta g^T]\right) - \boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right) = \mathbf{L}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}. \tag{5.1.15}$$

Then from Lemma 1.3.2, we obtain

$$\begin{aligned}
\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} &\approx -\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\Delta A)\\ \mathrm{vec}(\Delta B)\\ \mathrm{vec}(\Delta D)\\ \Delta f\\ \Delta g\end{bmatrix}\\
&= -\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_{\mathcal{S}_n} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Phi}_{\mathcal{T}_{mn}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & I_{m^2+m+n}\end{bmatrix}\begin{bmatrix}\mathrm{vec}_{\mathcal{S}}(\Delta A)\\ \mathrm{vec}_{\mathcal{T}}(\Delta B)\\ \mathrm{vec}(\Delta D)\\ \Delta f\\ \Delta g\end{bmatrix}\\
&= -\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}} & -I_{m+n}\end{bmatrix}\begin{bmatrix}\mathfrak{D}_{\mathcal{E}}^{-1}\mathfrak{D}_{\mathcal{E}}\Delta\boldsymbol{w}\\ \Delta f\\ \Delta g\end{bmatrix}\\
&= -\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n}\end{bmatrix}\begin{bmatrix}\mathfrak{D}_{\mathcal{E}}\Delta\boldsymbol{w}\\ \Delta f\\ \Delta g\end{bmatrix}. \qquad (5.1.16)
\end{aligned}$$

Combining (5.1.16) and (5.1.15), the *Fréchet* derivative of $\boldsymbol{\zeta}$ at $\begin{bmatrix}\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}\\ f\\ g\end{bmatrix}$ is

$$\mathbf{d}\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right) = -\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n}\end{bmatrix}.$$

Hence, the proof follows. ■

Using the Lemma 5.1.8 and Lemma 1.3.2, we next derive the compact formulae for the structured CNs defined in Definition 5.1.2.

Theorem 5.1.9. *Let $[x^T,\, y^T]^T$ be the unique solution of the GSPP (5.1.3) with the structure $\mathcal{E}$. Then, the structured partial NCN, MCN, and CCN, respectively, are given by*

$$\mathscr{K}(\mathbf{L}\left[x^T,\, y^T\right]^T;\, \mathcal{E}) = \frac{\left\|\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\mathbf{H} & \mathbf{b}\end{bmatrix}\right\|_F}{\|\mathbf{L}[x^T,\, y^T]^T\|_2},$$

$$\mathscr{M}(\mathbf{L}\left[x^T,\, y^T\right]^T;\, \mathcal{E}) = \frac{\left\||\mathbf{L}\mathcal{M}^{-1}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}|\begin{bmatrix}\mathrm{vec}_{\mathcal{S}}(|A|)\\ \mathrm{vec}_{\mathcal{T}}(|B|)\\ \mathrm{vec}(|D|)\end{bmatrix} + |\mathbf{L}\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}\right\|_\infty}{\left\|\mathbf{L}[x^T,\, y^T]^T\right\|_\infty}, \quad \textit{and}$$

$$\mathscr{C}(\mathbf{L}\left[x^T,\, y^T\right]^T;\, \mathcal{E}) = \left\|\mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,y^T]^T}|\mathbf{L}\mathcal{M}^{-1}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}|\begin{bmatrix}\mathrm{vec}_{\mathcal{S}}(|A|)\\ \mathrm{vec}_{\mathcal{T}}(|B|)\\ \mathrm{vec}(|D|)\end{bmatrix} + \mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,y^T]^T}|\mathbf{L}\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}\right\|_\infty.$$

Proof. Let $\boldsymbol{w}^T = [\mathrm{vec}_{\mathcal{S}}^T(A), \mathrm{vec}_{\mathcal{T}}(B)^T, \mathrm{vec}(D)^T]^T$. Following the proof method of Proposition 5.1.2, we have

$$\begin{aligned}
&\mathscr{K}(\mathbf{L}\left[x^T,\, y^T\right]^T;\, \mathcal{E}) = \mathscr{K}(\boldsymbol{\zeta}, [\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T),\\
&\mathscr{M}(\mathbf{L}\left[x^T,\, y^T\right]^T;\, \mathcal{E}) = \mathscr{M}(\boldsymbol{\zeta}, [\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T),\\
&\text{and} \quad \mathscr{C}(\mathbf{L}\left[x^T,\, y^T\right]^T;\, \mathcal{E}) = \mathscr{C}(\boldsymbol{\zeta}, [\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T).
\end{aligned}$$

Applying the NCN formula given in Lemma 1.3.2 for the map $\boldsymbol{\zeta}$, we obtain

$$\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) = \frac{\left\|\mathbf{d}\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right)\right\|_2 \left\|\begin{bmatrix}\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}\\ f\\ g\end{bmatrix}\right\|_2}{\|\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right)\|_2}. \tag{5.1.17}$$

Now, substituting *Fréchet* derivative of $\boldsymbol{\zeta}$ provided in Lemma 5.1.3 in (5.1.17), we have

$$\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) = \frac{\left\|\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\mathbf{H} & \mathbf{b}\end{bmatrix}\right\|_F}{\|\mathbf{L}[x^T,\, y^T]^T\|_2}.$$

Similarly, applying the MCN formula provided in Lemma 1.3.2 for $\boldsymbol{\zeta}$, we get

$$\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) = \frac{\left\||\mathbf{d}\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right)|\begin{bmatrix}\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}\\ f\\ g\end{bmatrix}\right\|_\infty}{\|\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right)\|_\infty}. \tag{5.1.18}$$

Now, using Lemma 5.1.8 in (5.1.18), we obtain

$$\begin{aligned}\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) &= \frac{\left\| \left| \mathbf{L}\mathcal{M}^{-1} \begin{bmatrix} \mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n} \end{bmatrix} \right| \left| \begin{bmatrix} \mathfrak{D}_{\mathcal{E}}\boldsymbol{w} \\ f \\ g \end{bmatrix} \right| \right\|_{\infty}}{\|\mathbf{L}[x^T,\, y^T]^T\|_{\infty}} \\ &= \frac{\left\| |\mathbf{L}\mathcal{M}^{-1}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1}||\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}| + |\mathbf{L}\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_{\infty}}{\|\mathbf{L}[x^T,\, y^T]^T\|_{\infty}} \\ &= \frac{\left\| |\mathbf{L}\mathcal{M}^{-1}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}| \begin{bmatrix} \mathrm{vec}_{\mathcal{S}}(|A|) \\ \mathrm{vec}_{\mathcal{T}}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} + |\mathbf{L}\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_{\infty}}{\left\|\mathbf{L}[x^T,\, y^T]^T\right\|_{\infty}}.\end{aligned}$$

In an analogous method, we get

$$\begin{aligned}\mathscr{C}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{E}) &= \left\| \mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,y^T]^T} |\mathbf{d}\boldsymbol{\zeta}\left([\mathfrak{D}_{\mathcal{E}}\boldsymbol{w}^T,\, f^T,\, g^T]^T\right)| \left| \begin{bmatrix} \mathfrak{D}_{\mathcal{E}}\boldsymbol{w} \\ f \\ g \end{bmatrix} \right| \right\|_{\infty} \\ &= \left\| \mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,y^T]^T} \mathbf{L}|\mathcal{M}^{-1}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}| \begin{bmatrix} \mathrm{vec}_{\mathcal{S}}(|A|) \\ \mathrm{vec}_{\mathcal{T}}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} + \mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,y^T]^T} |\mathbf{L}\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_{\infty}.\end{aligned}$$

Hence, the proof is completed. ■

Remark 5.1.10. *Note that the structured MCN and CCN formulae presented in Theorem 5.1.9 involve computing the inverse of the matrix $\mathcal{M} \in \mathbb{R}^{(m+n)\times(m+n)}$, while the structured NCN formula involves computing the inverse of both matrices $\mathcal{M}$ and $\mathfrak{D}_{\mathcal{E}} \in \mathbb{R}^{\boldsymbol{l}\times\boldsymbol{l}}$. However, $\mathfrak{D}_{\mathcal{E}}$ is a diagonal matrix. Therefore, its inverse can be computed using only $\mathcal{O}(\boldsymbol{l})$ operations. On the other hand, to avoid computing $\mathcal{M}^{-1}$ explicitly, motivated by [87], we adopt the following procedure. Notably, the computation of $\mathcal{M}^{-1}$ is coming in the following form:*

$$\mathbf{L}\mathcal{M}^{-1} \begin{bmatrix} \mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n} \end{bmatrix} \quad or \quad \mathbf{L}\mathcal{M}^{-1}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}, \quad or \quad \mathbf{L}\mathcal{M}^{-1}.$$

Thus, first, we solve the system $\mathcal{M}X = Y$, where $Y = \begin{bmatrix} \mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n} \end{bmatrix}$ or $\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}$ and then compute $\mathbf{L}X$. The system $\mathcal{M}X = Y$ can be solved efficiently by LU decomposition. To compute $\mathbf{L}\mathcal{M}^{-1}$, we can solve $\mathbf{L} = XM$. It is worth noting that we only need to perform the LU decomposition once for all cases; this makes the procedure efficient and reliable.

Remark 5.1.11. *The Toeplitz matrix B is symmetric-Toeplitz (a special case of Toeplitz matrix) if $n = m$ and $b_{-n+1} = b_{n-1}, \ldots, b_{-1} = b_1$, where*

$$\mathrm{vec}_{\mathcal{T}}(B) = [b_{-n+1}, \ldots, b_1, b_0, \ldots, b_{n-1}]^T.$$

In this case, the basis for the set of symmetric-Toeplitz matrices is defined as $\left\{\widetilde{\mathcal{J}}_i\right\}_{i=1}^{n}$, where

$$\begin{aligned} \widetilde{\mathcal{J}}_1 &= \mathcal{T}([(e_n^n)^T, \mathbf{0}]^T) && \text{and} \\ \widetilde{\mathcal{J}}_{i+1} &= \mathcal{T}([(e_{n-i}^n)^T, (e_i^{(n-1)})^T]^T), && \text{for } i = 1, \ldots, n-1. \end{aligned}$$

Hence, the structured CNs for the GSPP (5.1.1) when A is symmetric, B is symmetric-Toeplitz is given by the formulae as in Theorem 5.1.9, with

$$\mathbf{\Phi}_{\mathcal{T}_{mn}} = \left[\mathrm{vec}(\widetilde{\mathcal{J}}_1), \ldots, \mathrm{vec}(\widetilde{\mathcal{J}}_n)\right] \in \mathbb{R}^{n^2 \times n}$$

and $\mathfrak{D}_{\mathcal{T}_{mn}} \in \mathbb{R}^{n \times n}$ with $\mathfrak{D}_{\mathcal{T}_{mn}}(j,j) = \hat{\boldsymbol{a}}_j$, where

$$\hat{\boldsymbol{a}} = [\sqrt{n}, \sqrt{2(n-1)}, \sqrt{2(n-2)}, \ldots, \sqrt{2}]^T \in \mathbb{R}^n.$$

Next, we compare the structured CNs with the unstructured ones given in (5.1.7)–(5.1.9).

Theorem 5.1.12. *With the above notation, when $\mathbf{L} = I_{m+n}$, we have the following relations:*

$$\begin{gathered} \mathscr{K}([x^T, y^T]^T; \mathcal{E}) \leq \mathscr{K}^u([x^T, y^T]^T), \ \mathscr{M}([x^T, y^T]^T; \mathcal{E}) \leq \mathscr{M}^u([x^T, y^T]^T) \\ \text{and} \quad \mathscr{C}([x^T, y^T]^T; \mathcal{E}) \leq \mathscr{C}^u([x^T, y^T]^T). \end{gathered}$$

Proof. Since $\mathbf{L} = I_{m+n}$, for the NCN, using the properties of the spectral norm, we obtain

$$\begin{aligned} \left\|\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n}\end{bmatrix}\right\|_2 &\leq \left\|\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & \mathbf{0} \\ \mathbf{0} & I_{m+n}\end{bmatrix}\right\|_2 \\ &= \left\|\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}\right\|_2. \end{aligned}$$

The last equality is obtained by using the fact that $\|\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1}\|_2 = 1$. Hence, the first claim is achieved.

Since $\mathbf{\Phi}_{\mathcal{E}}$ has at most one nonzero entry in each row, we obtain

$$\begin{aligned}
|\mathcal{M}^{-1}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}| \begin{bmatrix} \mathrm{vec}_{\mathcal{S}}(|A|) \\ \mathrm{vec}_{\mathcal{T}}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} &\leq |\mathcal{M}^{-1}\mathcal{R}||\mathbf{\Phi}_{\mathcal{E}}| \begin{bmatrix} \mathrm{vec}_{\mathcal{S}}(|A|) \\ \mathrm{vec}_{\mathcal{T}}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} \\
&= |\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} |\mathbf{\Phi}_{\mathcal{S}_n}|\mathrm{vec}_{\mathcal{S}}(|A|) \\ |\mathbf{\Phi}_{\mathcal{T}_{mn}}|\mathrm{vec}_{\mathcal{T}}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} \\
&= |\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} |\mathbf{\Phi}_{\mathcal{S}_n}\mathrm{vec}_{\mathcal{S}}(A)| \\ |\mathbf{\Phi}_{\mathcal{T}_{mn}}\mathrm{vec}_{\mathcal{T}}(B)| \\ \mathrm{vec}(|D|) \end{bmatrix} \\
&= |\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} \mathrm{vec}(|A|) \\ \mathrm{vec}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix}.
\end{aligned}$$

Therefore, from Theorem 5.1.4, we obtain

$$\mathscr{M}([x^T,\, y^T]^T;\, \mathcal{E}) \leq \frac{\left\| |\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} \mathrm{vec}(|A|) \\ \mathrm{vec}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} + |\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_\infty}{\|[x^T,\, y^T]^T\|_\infty} = \mathscr{M}^u([x^T,\, y^T]^T)$$

and

$$\begin{aligned}
\mathscr{C}([x^T,\, y^T]^T;\, \mathcal{E}) &\leq \left\| \mathfrak{D}^\dagger_{[x^T,\, y^T]^T} |\mathcal{M}^{-1}\mathcal{R}| \begin{bmatrix} \mathrm{vec}(|A|) \\ \mathrm{vec}(|B|) \\ \mathrm{vec}(|D|) \end{bmatrix} + \mathfrak{D}^\dagger_{[x^T,\, y^T]^T} |\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_\infty \\
&= \mathscr{C}^u([x^T,\, y^T]^T).
\end{aligned}$$

Hence, the proof is completed. ■

5.1.4. Structured Partial CNs when A and D have Linear Structures

In this subsection, we consider $\mathcal{L}_1 \subseteq \mathbb{R}^{n\times n}$ and $\mathcal{L}_2 \subseteq \mathbb{R}^{m\times m}$ are two distinct linear subspaces containing different classes of structured matrices. Suppose that the $\dim(\mathcal{L}_1) = p$ and $\dim(\mathcal{L}_2) = s$ and the corresponding bases are $\{E_i\}_{i=1}^p$ and $\{F_i\}_{i=1}^s$, respectively. Let $A \in \mathcal{L}_1$ and $D \in \mathcal{L}_2$. Then there are unique vectors

$$\mathrm{vec}_{\mathcal{L}_1}(A) = [a_1, a_2, \ldots, a_p]^T \in \mathbb{R}^p \quad \text{and} \quad \mathrm{vec}_{\mathcal{L}_2}(D) = [d_1, d_2, \ldots, d_s]^T \in \mathbb{R}^s$$

such that

$$A = \sum_{i=1}^{p} a_i E_i \quad \text{and} \quad D = \sum_{i=1}^{s} d_i F_i. \tag{5.1.19}$$

Subsequently, we obtain the following for the vec-structure of the matrices A and D.

Lemma 5.1.13. *Let $A \in \mathcal{L}_1$ and $D \in \mathcal{L}_2$, then $\text{vec}(A) = \mathbf{\Phi}_{\mathcal{L}_1}\text{vec}_{\mathcal{L}_1}(A)$ and $\text{vec}(D) = \mathbf{\Phi}_{\mathcal{L}_2}\text{vec}_{\mathcal{L}_2}(D)$, where*

$$\mathbf{\Phi}_{\mathcal{L}_1} = \begin{bmatrix}\text{vec}(E_1) & \text{vec}(E_2) & \cdots & \text{vec}(E_p)\end{bmatrix} \in \mathbb{R}^{n^2\times p}$$
$$\text{and} \quad \mathbf{\Phi}_{\mathcal{L}_2} = \begin{bmatrix}\text{vec}(F_1) & \text{vec}(F_2) & \cdots & \text{vec}(F_s)\end{bmatrix} \in \mathbb{R}^{m^2\times s}.$$

Proof. Assume that $\text{vec}_{\mathcal{L}_1}(A) = [a_1, a_2, \ldots, a_p]^T \in \mathbb{R}^p$, then from (5.1.19), we obtain

$$\text{vec}(A) = \sum_{i=1}^{p} a_i \text{vec}(E_i) = \mathbf{\Phi}_{\mathcal{L}_1}\text{vec}_{\mathcal{L}_1}(A).$$

Similarly, we can obtain $\text{vec}(D) = \mathbf{\Phi}_{\mathcal{L}_2}\text{vec}_{\mathcal{L}_2}(D)$. ■

The matrices $\mathbf{\Phi}_{\mathcal{L}_1}$ and $\mathbf{\Phi}_{\mathcal{L}_2}$ contains the information about the structures of A and D consisting with the linear subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. For unstructured matrices, $\mathbf{\Phi}_{\mathcal{L}_1} = I_{n^2}$ and $\mathbf{\Phi}_{\mathcal{L}_2} = I_{m^2}$. On the other hand, there exist diagonal matrices $\mathfrak{D}_{\mathcal{L}_1} \in \mathbb{R}^{p\times p}$ and $\mathfrak{D}_{\mathcal{L}_2} \in \mathbb{R}^{s\times s}$ with the diagonal entries $\mathfrak{D}_{\mathcal{L}_j}(i,i) = \|\mathbf{\Phi}_{\mathcal{L}_j}(:,i)\|_2$, for $j = 1, 2$, such that

$$\|A\|_F = \|\mathfrak{D}_{\mathcal{L}_1} a\|_2 \quad \text{and} \quad \|D\|_F = \|\mathfrak{D}_{\mathcal{L}_2} d\|_2. \tag{5.1.20}$$

To perform structured perturbation analysis, we restrict the perturbation ΔA on A and ΔD on D to the linear subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Then, there are unique vectors $\text{vec}_{\mathcal{L}_1}(\Delta A) \in \mathbb{R}^p$ and $\text{vec}_{\mathcal{L}_2}(\Delta D) \in \mathbb{R}^s$ such that

$$\text{vec}(\Delta A) = \mathbf{\Phi}_{\mathcal{L}_1}\text{vec}_{\mathcal{L}_1}(\Delta A) \text{ and } \text{vec}(\Delta D) = \mathbf{\Phi}_{\mathcal{L}_2}\text{vec}_{\mathcal{L}_1}(\Delta D). \tag{5.1.21}$$

Now, consider the following set:

$$\mathcal{L} = \left\{ \mathcal{M} = \begin{bmatrix} A & B^T \\ C & D \end{bmatrix} : A \in \mathcal{L}_1, B, C \in \mathbb{R}^{m\times n}, D \in \mathcal{L}_2 \right\}. \tag{5.1.22}$$

Consider the perturbations ΔA, ΔB, ΔC, ΔD, Δf, and Δg on the matrices A, B, C, D, f, and g, respectively. Then, the perturbed counterpart of the system (5.1.1)

$$(\mathcal{M} + \Delta\mathcal{M}) \begin{bmatrix} x + \Delta x \\ y + \Delta y \end{bmatrix} = \begin{bmatrix} A + \Delta A & (B + \Delta B)^T \\ C + \Delta C & D + \Delta D \end{bmatrix} \begin{bmatrix} x + \Delta x \\ y + \Delta y \end{bmatrix} = \begin{bmatrix} f + \Delta f \\ g + \Delta g \end{bmatrix} \tag{5.1.23}$$

has a unique solution $\begin{bmatrix} x + \Delta x \\ y + \Delta y \end{bmatrix}$ when $\|\mathcal{M}\|_2 \|\Delta\mathcal{M}\|_2 < 1$. Consequently, neglecting higher-order terms, we can rewrite (5.1.23) as

$$\mathcal{M}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}=\begin{bmatrix}A & B^T\\ C & D\end{bmatrix}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}\approx\begin{bmatrix}\Delta f\\ \Delta g\end{bmatrix}-\begin{bmatrix}\Delta A & \Delta B^T\\ \Delta C & \Delta D\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}. \tag{5.1.24}$$

Using the properties of the Kronecker product mentioned in (1.3.2), we have the following lemma.

Lemma 5.1.14. *Let $[x^T,\, y^T]^T$ and $[(x+\Delta x)^T,\, (y+\Delta y)^T]^T$ be the unique solutions of the GSPP (5.1.1) and (5.1.23), respectively, with structure $\mathcal{L}$. Then, we have the following perturbation expression*

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}\approx-\mathcal{M}^{-1}\begin{bmatrix}\mathcal{H} & -I_{m+n}\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\Delta A)\\ \mathrm{vec}(\Delta B)\\ \mathrm{vec}(\Delta C)\\ \mathrm{vec}(\Delta D)\\ \Delta f\\ \Delta g\end{bmatrix}, \tag{5.1.25}$$

where

$$\mathcal{H}=\begin{bmatrix}x^T\otimes I_n & I_n\otimes y^T & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & x^T\otimes I_m & y^T\otimes I_m\end{bmatrix}. \tag{5.1.26}$$

Next, we define the structured partial NCN, MCN, and CCN for the solution of the GSPP (5.1.1) with the structure $\mathcal{L}$.

Definition 5.1.3. *Let $[x^T,\, y^T]^T$ and $[(x+\Delta x)^T,\, (y+\Delta y)^T]^T$ be the unique solutions of GSPPs (5.1.1) and (5.1.23), respectively, with the structure $\mathcal{L}$. Suppose $\mathbf{L}\in\mathbb{R}^{k\times(m+n)}$, then the structured partial NCN, MCN, and CCN, respectively, are defined as follows:*

$$\begin{aligned}&\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T;\,\mathcal{L})\\ &\quad:=\limsup_{\eta\to0}\left\{\frac{\left\|\mathbf{L}[\Delta x^T,\,\Delta y^T]^T\right\|_2}{\eta\left\|\mathbf{L}[x^T,\,y^T]^T\right\|_2}:\left\|\begin{bmatrix}\Delta\mathcal{M} & \Delta\mathbf{d}\end{bmatrix}\right\|_F\le\eta\left\|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right\|_F,\,\Delta\mathcal{M}\in\mathcal{L}\right\},\\ &\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T;\,\mathcal{L})\\ &\quad:=\limsup_{\eta\to0}\left\{\frac{\left\|\mathbf{L}[\Delta x^T,\,\Delta y^T]^T\right\|_\infty}{\eta\left\|\mathbf{L}[x^T,\,y^T]^T\right\|_\infty}:\left|\begin{bmatrix}\Delta\mathcal{M} & \Delta\mathbf{d}\end{bmatrix}\right|\le\eta\left|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right|,\,\Delta\mathcal{M}\in\mathcal{L}\right\},\\ &\mathscr{C}(\mathbf{L}[x^T,\, y^T]^T;\,\mathcal{L})\\ &\quad::=\limsup_{\eta\to0}\left\{\frac{1}{\eta}\left\|\frac{\mathbf{L}[\Delta x^T,\,\Delta y^T]^T}{\mathbf{L}[x^T,\,y^T]^T}\right\|_\infty:\left|\begin{bmatrix}\Delta\mathcal{M} & \Delta\mathbf{d}\end{bmatrix}\right|\le\eta\left|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right|,\,\Delta\mathcal{M}\in\mathcal{L}\right\}.\end{aligned}$$

The main objective of this section is to develop explicit formulae for the structured CNs defined above. To accomplish these, let $\boldsymbol{v}$ be a vector in $\mathbb{R}^{p+2mn+s}$ defined as

$$\boldsymbol{v} = \left[\text{vec}_{\mathcal{L}_1}^T(A), \text{vec}(B)^T, \text{vec}(C)^T, \text{vec}_{\mathcal{L}_2}^T(D)\right]^T. \tag{5.1.27}$$

To apply the Lemma 1.3.2, we define the mapping

$$\begin{aligned} &\Upsilon : \mathbb{R}^{p+2mn+s} \times \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}^k \quad \text{by} \\ &\Upsilon\left([\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T,\, f^T,\, g^T]^T\right) = \mathbf{L}\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{L}\mathcal{M}^{-1}\begin{bmatrix} f \\ g \end{bmatrix}, \end{aligned} \tag{5.1.28}$$

where

$$\mathfrak{D}_{\mathcal{L}} = \begin{bmatrix} \mathfrak{D}_{\mathcal{L}_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{2mn} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathfrak{D}_{\mathcal{L}_2} \end{bmatrix} \tag{5.1.29}$$

such that $\|\mathcal{M}\|_F = \|\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}\|_2$.

In the following lemma, we present explicit formulations of $\mathbf{d}\Upsilon$.

Lemma 5.1.15. *The mapping Υ defined in (5.1.28) is continuously Fréchet differentiable at $[\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T,\, f^T,\, g^T]^T$ and the Fréchet derivative is given by*

$$\mathbf{d}\Upsilon\left([\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T,\, f^T,\, g^T]^T\right) = -\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix} \mathcal{H}\mathbf{\Phi}_{\mathcal{L}}\mathfrak{D}_{\mathcal{L}} & -I_{m+n} \end{bmatrix}, \tag{5.1.30}$$

where $\mathbf{\Phi}_{\mathcal{L}} = \begin{bmatrix} \mathbf{\Phi}_{\mathcal{L}_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{2mn} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Phi}_{\mathcal{L}_2} \end{bmatrix}$, *$\mathcal{H}$ and $\mathfrak{D}_{\mathcal{L}}$ are defined as in (5.1.26) and (5.1.29), respectively.*

Proof. The proof follows in a similar way to the proof method of Lemma 5.1.8. ■

We now present compact formulae of the structured partial NCN, MCN, and CCN introduced in Definition 4.1.1. We use the Lemmas 1.3.2 and 5.1.15 to prove the following theorem.

Theorem 5.1.16. *The structured partial NCN, MCN, and CCN of the GSPP (5.1.1) with the structure $\mathcal{L}$, respectively, are given by*

$$\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{L}) = \frac{\left\|\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{H}\mathbf{\Phi}_{\mathcal{L}}\mathfrak{D}_{\mathcal{L}}^{-1} & -I_{m+n}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right\|_F}{\|\mathbf{L}[x^T,\, y^T]^T\|_2},$$

$$\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{L}) = \frac{\left\|\left|\mathbf{L}\mathcal{M}^{-1}\mathcal{H}\mathbf{\Phi}_{\mathcal{L}}\right|\begin{bmatrix}\mathrm{vec}_{\mathcal{L}_1}(|A|)\\ \mathrm{vec}(|B|)\\ \mathrm{vec}(|C|)\\ \mathrm{vec}_{\mathcal{L}_2}(|D|)\end{bmatrix} + \left|\mathbf{L}\mathcal{M}^{-1}\right|\begin{bmatrix}|f|\\ |g|\end{bmatrix}\right\|_\infty}{\left\|\mathbf{L}[x^T,\, y^T]^T\right\|_\infty}, \text{ and}$$

$$\mathscr{C}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{L}) = \left\|\mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,y^T]^T}|\mathbf{L}\mathcal{M}^{-1}\mathcal{H}\mathbf{\Phi}_{\mathcal{L}}|\begin{bmatrix}\mathrm{vec}_{\mathcal{L}_1}(|A|)\\ \mathrm{vec}(|B|)\\ \mathrm{vec}(|C|)\\ \mathrm{vec}_{\mathcal{L}_2}(|D|)\end{bmatrix} + \mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,y^T]^T}|\mathbf{L}\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}\right\|_\infty.$$

Proof. Similar to the proof method of Proposition 5.1.2, we have

$$\begin{aligned}
\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{L}) &= \mathscr{K}(\Upsilon, \left[\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T, f^T, g^T\right]^T),\\
\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T;\ \ \mathcal{L}) &= \mathscr{M}(\Upsilon, \left[\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T, f^T, g^T\right]^T),\\
\text{and } \mathscr{C}(\mathbf{L}[x^T,\, y^T]^T;\, \mathcal{L}) &= \mathscr{C}(\Upsilon, \left[\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T, f^T, g^T\right]^T).
\end{aligned}$$

Using Lemma 5.1.15 and NCN formula provided in Lemma 1.3.2, we have

$$\begin{aligned}
\mathscr{K}(\mathbf{L}[x^T,\, y^T]^T;\mathcal{L}) &= \frac{\left\|\mathbf{d}\Upsilon\left([\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T,\, f^T,\, g^T]^T\right)\right\|_2 \left\|\begin{bmatrix}\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}\\ f\\ g\end{bmatrix}\right\|_2}{\left\|\Upsilon\left([\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T,\, f^T,\, g^T]^T\right)\right\|_2}\\
&= \frac{\left\|\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{H}\mathbf{\Phi}_{\mathcal{L}}\mathfrak{D}_{\mathcal{L}}^{-1} & -I_{m+n}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right\|_F}{\|\mathbf{L}[x^T,\, y^T]^T\|_2}.
\end{aligned}$$

For structured MCN, again using Lemmas 1.3.2 and 5.1.15, we obtain

$$\mathscr{M}(\mathbf{L}[x^T,\, y^T]^T;\mathcal{L}) = \frac{\left\|\left|\mathbf{d}\Upsilon\left([\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T,\, f^T,\, g^T]^T\right)\right|\left|\begin{bmatrix}\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}\\ f\\ g\end{bmatrix}\right|\right\|_\infty}{\left\|\Upsilon\left([\mathfrak{D}_{\mathcal{L}}\boldsymbol{v}^T,\, f^T,\, g^T]^T\right)\right\|_\infty}$$

$$= \frac{\left\| \left| \mathbf{L}\mathcal{M}^{-1} \begin{bmatrix} \mathcal{H}\mathbf{\Phi}_{\mathcal{L}}\mathfrak{D}_{\mathcal{L}}^{-1} & -I_{m+n} \end{bmatrix} \right| \left| \begin{bmatrix} \mathfrak{D}_{\mathcal{L}}\boldsymbol{v} \\ f \\ g \end{bmatrix} \right| \right\|_{\infty}}{\|\mathbf{L}[x^T,\, y^T]^T\|_{\infty}}$$

$$= \frac{\left\| |\mathbf{L}\mathcal{M}^{-1}\mathcal{H}\mathbf{\Phi}_{\mathcal{L}}| \begin{bmatrix} \mathrm{vec}_{\mathcal{L}_1}(|A|) \\ \mathrm{vec}(|B|) \\ \mathrm{vec}(|C|) \\ \mathrm{vec}_{\mathcal{L}_2}(|D|) \end{bmatrix} + |\mathbf{L}\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_{\infty}}{\left\|\mathbf{L}[x^T,\, y^T]^T\right\|_{\infty}}.$$

Similarly, the rest of the proof follows. ■

Remark 5.1.17. *To compute the inverses of $\mathcal{M}$ and $\mathfrak{D}_{\mathcal{L}}$, one can follow a similar procedure as discussed in Remark 5.1.10.*

Remark 5.1.18. *Considering $\mathbf{L} = I_{m+n}$, $\begin{bmatrix} I_m & \mathbf{0} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{0} & I_n \end{bmatrix}$ in Theorem 5.1.16, we obtain the structured NCN, MCN, and CCN for the solution $[x^T,\, y^T]^T, x$ and y, respectively.*

Remark 5.1.19. *For $A \in \mathcal{S}_n$ and $D \in \mathcal{S}_m$, set*

$$\mathbf{\Phi}_{\mathcal{S}} = \begin{bmatrix} \mathbf{\Phi}_{\mathcal{S}_n} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{2mn} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Phi}_{\mathcal{S}_m} \end{bmatrix} \quad \text{and} \quad \mathfrak{D}_{\mathcal{S}} = \begin{bmatrix} \mathfrak{D}_{\mathcal{S}_n} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{2mn} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathfrak{D}_{\mathcal{S}_m} \end{bmatrix},$$

where $\mathbf{\Phi}_{\mathcal{S}_m}, \mathbf{\Phi}_{\mathcal{S}_n}, \mathfrak{D}_{\mathcal{S}_m}$ and $\mathfrak{D}_{\mathcal{S}_n}$ are defined as in Subsection 5.1.3. Then, the structured partial NCN, MCN, and CCN when $\mathcal{L}_1 = \mathcal{S}_n$ and $\mathcal{L}_2 = \mathcal{S}_m$ are obtained by substituting $\mathbf{\Phi}_{\mathcal{L}} = \mathbf{\Phi}_{\mathcal{S}}$, $\mathfrak{D}_{\mathcal{L}} = \mathfrak{D}_{\mathcal{S}}$, $\mathrm{vec}_{\mathcal{L}_1}(A) = \mathrm{vec}_{\mathcal{S}_n}(A)$ and $\mathrm{vec}_{\mathcal{L}_2}(D) = \mathrm{vec}_{\mathcal{S}_m}(D)$ in Theorem 5.1.16.

Next, consider the linear system $\mathcal{M}z = \mathbf{b}$, where $\mathcal{M} \in \mathbb{R}^{l\times l}$ being any nonsingular matrix and $\mathbf{b} \in \mathbb{R}^l$. Then, this system can be partitioned as GSPP (5.1.1) by setting $l = m + n$. Let $\Delta\mathcal{M} \in \mathbb{R}^{l\times l}$ and $\Delta\mathbf{b} \in \mathbb{R}^l$, then the perturbed system is given by

$$(\mathcal{M} + \Delta\mathcal{M})(z + \Delta z) = (\mathbf{b} + \Delta\mathbf{b}).$$

Skeel [127] and Rohn [118] propose the following formulae for the unstructured MCN and CCN for the solution of the above linear system:

$$\begin{aligned} \widetilde{\mathscr{M}}(z) &:= \limsup_{\eta\to 0} \left\{ \frac{\|\Delta z\|_{\infty}}{\eta\|z\|_{\infty}} : |\Delta\mathcal{M}| \le \eta|\mathcal{M}|, |\Delta\mathbf{b}| \le \eta|\mathbf{b}| \right\} \\ &= \frac{\||\mathcal{M}^{-1}||\mathcal{M}||z| + |\mathcal{M}^{-1}||\mathbf{b}|\|_{\infty}}{\|z\|_{\infty}}, \end{aligned} \tag{5.1.31}$$

$$\widetilde{\mathscr{C}}(z) := \limsup_{\eta\to 0}\left\{\frac{1}{\eta}\left\|\frac{\Delta z}{z}\right\|_\infty : |\Delta\mathcal{M}| \le \eta|\mathcal{M}|, |\Delta\mathbf{b}| \le \eta|\mathbf{b}|\right\}$$
$$= \left\|\frac{|\mathcal{M}^{-1}||\mathcal{M}||z| + |\mathcal{M}^{-1}||\mathbf{b}|}{|z|}\right\|_\infty. \tag{5.1.32}$$

Remark 5.1.20. *Considering $\mathbf{\Phi}_{\mathcal{S}_m} = I_{m^2}$ and $\mathbf{\Phi}_{\mathcal{S}_n} = I_{n^2}$ on the formula of $\mathscr{K}([x^T, y^T]^T; \mathcal{L})$, we obtain the unstructured NCN for $\mathcal{M}z = \mathbf{b}$, where $\mathcal{M} \in \mathbb{R}^{l\times l}$, $\mathbf{b} \in \mathbb{R}^l$ and $l = (m+n)$, which is given by*

$$\widetilde{\mathscr{K}}(z) := \limsup_{\eta\to 0}\left\{\frac{\|\Delta z\|_2}{\eta\|z\|_2} : \left\|\begin{bmatrix}\Delta\mathcal{M} & \Delta\mathbf{b}\end{bmatrix}\right\|_F \le \eta\left\|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right\|_F\right\}$$
$$= \frac{\left\|\mathcal{M}^{-1}\begin{bmatrix}\mathcal{H} & -I_{m+n}\end{bmatrix}\right\|_2\left\|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right\|_F}{\|z\|_2}.$$

The following theorem compares the structured NCN, MCN, and CCN obtained in Theorem 5.1.16 and the unstructured counterparts defined above.

Theorem 5.1.21. *Let $z = [x^T, y^T]^T$ and $\mathbf{L} = I_{m+n}$. Then, for the GSPP (5.1.1) with the structure $\mathcal{L}$, following relations holds:*

$$\mathscr{K}([x^T, y^T]^T; \mathcal{L}) \le \widetilde{\mathscr{K}}([x^T, y^T]^T), \quad \mathscr{M}([x^T, y^T]^T; \mathcal{L}) \le \widetilde{\mathscr{M}}([x^T, y^T]^T)$$
$$\text{and} \quad \mathscr{C}([x^T, y^T]^T; \mathcal{L}) \le \widetilde{\mathscr{C}}([x^T, y^T]^T).$$

Proof. Since $\|\mathbf{\Phi}_{\mathcal{L}}\mathfrak{D}_{\mathcal{L}}^{-1}\|_2 = 1$, the proof follows similar to the proof method of Theorem 5.1.12. Hence, from Theorem 5.1.16 and Remark 5.1.20, we have $\mathscr{K}([x^T, y^T]^T; \mathcal{L}) \le \widetilde{\mathscr{K}}([x^T, y^T]^T)$. Now, using the property that the matrices $\mathbf{\Phi}_{\mathcal{L}_i}$, $i = 1, 2$, have at most one nonzero entry in each row [86], and similar to Theorem 5.1.12, we obtain $|\mathbf{\Phi}_{\mathcal{L}_1}\text{vec}_{\mathcal{L}_1}(A)| = |\mathbf{\Phi}_{\mathcal{L}_1}|\text{vec}_{\mathcal{L}_1}(|A|)$ and $|\mathbf{\Phi}_{\mathcal{L}_2}\text{vec}_{\mathcal{L}_2}(D)| = |\mathbf{\Phi}_{\mathcal{L}_2}|\text{vec}_{\mathcal{L}_2}(|D|)$. Then

$$|\mathcal{M}^{-1}\mathcal{H}\mathbf{\Phi}_{\mathcal{L}}|\begin{bmatrix}\text{vec}_{\mathcal{L}_1}(|A|)\\ \text{vec}(|B|)\\ \text{vec}(|C|)\\ \text{vec}_{\mathcal{L}_2}(|D|)\end{bmatrix} + |\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix} \le |\mathcal{M}^{-1}||\mathcal{H}|\begin{bmatrix}\text{vec}(|A|)\\ \text{vec}(|B|)\\ \text{vec}(|C|)\\ \text{vec}(|D|)\end{bmatrix} + |\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}$$
$$\le |\mathcal{M}^{-1}|\begin{bmatrix}|x^T|\otimes I_m & I_m\otimes|y^T| & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & |x^T|\otimes I_n & |y^T|\otimes I_n\end{bmatrix}\begin{bmatrix}\text{vec}(|A|)\\ \text{vec}(|B|)\\ \text{vec}(|C|)\\ \text{vec}(|D|)\end{bmatrix} + |\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}$$
$$= |\mathcal{M}^{-1}||\mathcal{M}|\begin{bmatrix}|x|\\ |y|\end{bmatrix} + |\mathcal{M}^{-1}|\begin{bmatrix}|f|\\ |g|\end{bmatrix}.$$

Consequently, by Theorem 5.1.16, we have

$$\mathscr{M}([x^T,\, y^T]^T;\, \mathcal{L}) \leq \frac{\left\| |\mathcal{M}^{-1}||\mathcal{M}| \begin{bmatrix} |x| \\ |y| \end{bmatrix} + |\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix} \right\|_\infty}{\|[x^T,\, y^T]^T\|_\infty}, \tag{5.1.33}$$

$$\mathscr{C}([x^T,\, y^T]^T;\, \mathcal{L}) \leq \left\| \frac{|\mathcal{M}^{-1}||\mathcal{M}| \begin{bmatrix} |x| \\ |y| \end{bmatrix} + |\mathcal{M}^{-1}| \begin{bmatrix} |f| \\ |g| \end{bmatrix}}{|\begin{bmatrix} x^T,\, y^T \end{bmatrix}^T|} \right\|_\infty. \tag{5.1.34}$$

Now, considering $l = m + n$, $z = [x^T,\, y^T]^T$ and $\mathbf{b} = [f^T,\, g^T]^T$ in (5.1.31) and (5.1.32), and from above, we obtain the following relations:

$$\mathscr{M}([x^T,\, y^T]^T;\, \mathcal{L}) \leq \widetilde{\mathscr{M}}([x^T,\, y^T]^T) \quad \text{and} \quad \mathscr{C}([x^T,\, y^T]^T;\, \mathcal{L}) \leq \widetilde{\mathscr{C}}([x^T,\, y^T]^T).$$

Hence, the proof follows. ■

5.1.5. Application to WRLS Problems

Consider the WRLS problem (1.1.3) with $K = Q^T$ and let $\mathbf{r} = W(f - Qy)$, then the minimization problem (1.1.3) can be expressed as the following augmented linear system:

$$\widehat{\mathcal{M}} \begin{bmatrix} \mathbf{r} \\ y \end{bmatrix} := \begin{bmatrix} W^{-1} & Q \\ Q^T & -\lambda I_m \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ y \end{bmatrix} = \begin{bmatrix} f \\ \mathbf{0} \end{bmatrix}. \tag{5.1.35}$$

Identifying $A = W^{-1}$, $B = Q^T$, $D = -\lambda I_m$, $x = \mathbf{r}$, and $g = \mathbf{0}$, we can see that the augmented system (5.1.35) is in the form of the GSPP (5.1.3). Therefore, finding the CNs of the WRLS problem (1.1.3) is equivalent to the CNs of the GSPP (5.1.3) for y with $g = \mathbf{0}$. This accomplish by Theorem 5.1.4. Before that, we reformulate (5.1.2) (with $B = C$) as

$$\mathcal{M}^{-1} = \begin{bmatrix} M & N \\ K & S^{-1} \end{bmatrix}, \tag{5.1.36}$$

where $M = A^{-1} + A^{-1}B^TS^{-1}BA^{-1}$, $N = -A^{-1}B^TS^{-1}$, $K = -S^{-1}BA^{-1}$ and $S = D - BA^{-1}B^T$.

Theorem 5.1.22. *Let y be the unique solution of the problem (1.1.3) and $\mathbf{r} = W(f - Qy)$. Then, the structured NCN, MCN, and CCN for y, respectively, are given by*

$$\mathscr{K}^{rls}(y;\, \mathcal{E}) = \frac{\|\mathcal{X}\|_2 \left\| \begin{bmatrix} \widehat{\mathcal{M}} & \mathbf{d} \end{bmatrix} \right\|_F}{\|y\|_2}, \quad \mathscr{M}^{rls}(y;\, \mathcal{E}) = \frac{\|\mathcal{N}_y\|_\infty}{\|y\|_\infty}, \quad \text{and } \mathscr{C}^{rls}(y;\, \mathcal{E}) = \left\| \mathfrak{D}_y^\dagger \mathcal{N}_y \right\|_\infty,$$

where

$$\mathcal{X} = \left[(\mathbf{r}^T \otimes \widetilde{K})\mathbf{\Phi}_{\mathcal{S}_m}\mathfrak{D}_{\mathcal{S}_m}^{-1} \quad (K \otimes y^T + \mathbf{r}^T \otimes \widetilde{S}^{-1})\mathbf{\Phi}_{\mathcal{T}_{nm}}\mathfrak{D}_{\mathcal{E}}^{-1} \quad y^T \otimes \widetilde{S}^{-1} \quad -\widetilde{K} \quad -\widetilde{S}^{-1}\right],$$
$$\mathcal{N}_y = |(\mathbf{r}^T \otimes \widetilde{K})\mathbf{\Phi}_{\mathcal{S}_m}|\mathrm{vec}_{\mathcal{S}}(|A|) + |((K \otimes y^T) + (\mathbf{r}^T \otimes \widetilde{S}^{-1}))\mathbf{\Phi}_{\mathcal{T}_{nm}}|\mathrm{vec}_{\mathcal{T}}(|Q^T|)$$
$$+ |\widetilde{S}^{-1}||D||y| + |K||f|, \ \widetilde{K} = -\widetilde{S}^{-1}Q^TW, \ \textit{and} \ \widetilde{S} = -(\lambda I_n + Q^TWQ).$$

Proof. Let $\mathbf{L} = \begin{bmatrix}\mathbf{0} & I_m\end{bmatrix} \in \mathbb{R}^{m\times(m+n)}$, $A = W^{-1}, B = Q^T, D = -\lambda I_m$, $x = \mathbf{r}$, and $g = \mathbf{0}$. Then from Theorem 5.1.4, we have

$$\begin{aligned}
&\mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}\mathcal{R}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & -I_{m+n}\end{bmatrix} \\
&= \begin{bmatrix}\widetilde{K} & \widetilde{S}^{-1}\end{bmatrix}\begin{bmatrix}\mathcal{R} & -I_{m+n}\end{bmatrix}\begin{bmatrix}\mathbf{\Phi}_{\mathcal{E}}\mathfrak{D}_{\mathcal{E}}^{-1} & \mathbf{0} \\ \mathbf{0} & I_{m+n}\end{bmatrix} \\
&= \begin{bmatrix}(\mathbf{r}^T \otimes K)\mathbf{\Phi}_{\mathcal{S}_n}\mathfrak{D}_{\mathcal{S}_n}^{-1} & (\widetilde{K} \otimes y^T + \mathbf{r}^T \otimes \widetilde{S}^{-1})\mathbf{\Phi}_{\mathcal{T}_{mn}}\mathfrak{D}_{\mathcal{T}_{mn}}^{-1} & y^T \otimes \widetilde{S}^{-1} & -\widetilde{K} & -\widetilde{S}^{-1}\end{bmatrix}.
\end{aligned}$$

Hence, the expression for $\mathscr{K}^{rls}(y;\mathcal{E})$ is obtained from Theorem 5.1.4. The rest of the proof follows in a similar manner. ■

Since, in most cases of the WRLS problem, the weighted matrix W and regularization matrix $D = -\lambda I_m$ has no perturbation, we consider $\Delta A = \mathbf{0}$ and $\Delta D = \mathbf{0}$. Moreover, as $g = \mathbf{0}$, we assume $\Delta g = \mathbf{0}$. Then, perturbation expansion in Lemma 5.1.1 can be reformulated as

$$\begin{aligned}
\begin{bmatrix}\Delta x \\ \Delta y\end{bmatrix} &= -\mathcal{M}^{-1}\begin{bmatrix}I_n \otimes y^T & -I_n \\ x^T \otimes I_m & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\Delta B) \\ \Delta f\end{bmatrix} \\
&= -\left[\mathcal{R}_{rls} \quad -\begin{bmatrix}M \\ K\end{bmatrix}\right]\begin{bmatrix}\mathrm{vec}(\Delta B) \\ \Delta f\end{bmatrix},
\end{aligned} \tag{5.1.37}$$

where

$$\mathcal{R}_{rls} = \begin{bmatrix}M \otimes y^T + x^T \otimes N \\ K \otimes y^T + x^T \otimes S^{-1}\end{bmatrix}.$$

Now, applying a similar method to Subsection 5.1.3, we obtain the following expressions for the NCN, MCN, and CCN for $\mathbf{L}[x^T, y^T]^T$ when $B = C$ and $g = \mathbf{0}$.

Theorem 5.1.23. *Let $\Delta B \in \mathcal{T}_{m\times n}$ and with the above notations, structured partial NCN, MCN, and CCN for the GSPP (5.1.3), respectively, are given by*

$$\begin{aligned}
\widehat{\mathscr{K}}(\mathbf{L}[x^T,\, y^T]^T) &:= \limsup_{\eta\to 0}\left\{\frac{\|\mathbf{L}[\Delta x^T,\, \Delta y^T]^T\|_2}{\eta\|\mathbf{L}[x^T,\, y^T]^T\|_2} : \left\|\begin{bmatrix}\Delta B & \Delta f\end{bmatrix}\right\|_F \le \eta \left\|\begin{bmatrix} B & f\end{bmatrix}\right\|_F\right\}\\
&= \frac{\left\|\mathbf{L}\begin{bmatrix}\mathcal{R}_{rls}\mathbf{\Phi}_{\mathcal{T}_{mn}}\mathfrak{D}_{\mathcal{T}_{mn}}^{-1} & -\begin{bmatrix}M\\ K\end{bmatrix}\end{bmatrix}\right\|_2 \left\|\begin{bmatrix} B & f\end{bmatrix}\right\|_F}{\|\mathbf{L}[x^T,\, y^T]^T\|_2},\\
\widehat{\mathscr{M}}(\mathbf{L}[x^T,\, y^T]^T) &:= \limsup_{\eta\to 0}\left\{\frac{\|\mathbf{L}[\Delta x^T,\, \Delta y^T]^T\|_\infty}{\eta\|\mathbf{L}[x^T,\, y^T]^T\|_\infty} : \left|\begin{bmatrix}\Delta B & \Delta f\end{bmatrix}\right| \le \eta \left|\begin{bmatrix} B & f\end{bmatrix}\right|\right\}\\
&= \frac{\left\||\mathbf{L}\mathcal{R}_{rls}\mathbf{\Phi}_{\mathcal{T}_{mn}}|\mathrm{vec}_{\mathcal{T}}(|B|) + \left|\mathbf{L}\begin{bmatrix}M\\ K\end{bmatrix}\right| |f|\right\|_\infty}{\|\mathbf{L}[x^T,\, y^T]^T\|_\infty},\\
\widehat{\mathscr{C}}(\mathbf{L}[x^T,\, y^T]^T) &:= \limsup_{\eta\to 0}\left\{\frac{1}{\eta}\left\|\frac{\mathbf{L}[\Delta x^T,\, \Delta y^T]^T}{\mathbf{L}[x^T,\, y^T]^T}\right\|_\infty : \left|\begin{bmatrix}\Delta B & \Delta f\end{bmatrix}\right| \le \eta \left|\begin{bmatrix} B & f\end{bmatrix}\right|\right\}\\
&= \left\|\mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,\, y^T]^T}|\mathbf{L}\mathcal{R}_{rls}\mathbf{\Phi}_{\mathcal{T}_{mn}}|\mathrm{vec}_{\mathcal{T}}(|B|) + \mathfrak{D}^{\dagger}_{\mathbf{L}[x^T,\, y^T]^T}\left|\mathbf{L}\begin{bmatrix}M\\ K\end{bmatrix}\right| |f|\right\|_\infty.
\end{aligned}$$

Proof. For applying Lemma 1.3.2, we define

$$\begin{aligned}
&\widehat{\boldsymbol{\zeta}} : \mathbb{R}^{m+n-1}\times\mathbb{R}^n \mapsto \mathbb{R}^{m+n} \quad \text{by}\\
&\widehat{\boldsymbol{\zeta}}\left([\mathfrak{D}_{\mathcal{T}_{mn}}\mathrm{vec}_{\mathcal{T}}(B)^T,\, f^T]^T\right) = \mathbf{L}\begin{bmatrix}x\\ y\end{bmatrix} = \mathbf{L}\mathcal{M}^{-1}\begin{bmatrix}f\\ \mathbf{0}\end{bmatrix}.
\end{aligned}$$

Then, the map $\widehat{\boldsymbol{\zeta}}$ is continuously *Fréchet* differentiable at $[\mathfrak{D}_{\mathcal{T}_{mn}}\mathrm{vec}_{\mathcal{T}}(B)^T,\, f^T]^T$ with

$$\mathbf{d}\widehat{\boldsymbol{\zeta}}\left([\mathfrak{D}_{\mathcal{T}_{mn}}\mathrm{vec}_{\mathcal{T}}(B)^T,\, f^T]^T\right) = -\mathbf{L}\begin{bmatrix}\mathcal{R}_{rls}\mathbf{\Phi}_{\mathcal{T}_{mn}}\mathfrak{D}_{\mathcal{T}_{mn}}^{-1} & -\begin{bmatrix}M\\ K\end{bmatrix}\end{bmatrix}.$$

The rest of the proof follows similarly to Theorem 5.1.9. ■

Using the above result, we can derive the following structured CNs for the problem (1.1.3), when the weighted matrix and regularization matrix have no perturbation.

Corollary 5.1.1. *The structured NCN, MCN, and CCN for the solution y of the WRLS problem (1.1.3), respectively, are given by*

$$\widehat{\mathscr{K}^{rls}}(y) = \frac{\left\| \left[(\widetilde{K} \otimes y^T + \mathbf{r}^T \otimes \widetilde{S}^{-1}) \mathbf{\Phi}_{\mathcal{T}_{nm}} \mathfrak{D}_{\mathcal{T}_{nm}}^{-1} \quad -\widetilde{K} \right] \right\|_2 \left\| \begin{bmatrix} Q & f \end{bmatrix} \right\|_F}{\|y\|_2},$$

$$\widehat{\mathscr{M}^{rls}}(y) = \frac{\left\| |(\widetilde{K} \otimes y^T + \mathbf{r}^T \otimes \widetilde{S}^{-1}) \mathbf{\Phi}_{\mathcal{T}_{nm}} |\mathrm{vec}_{\mathcal{T}}(|Q^T|) + |\widetilde{K}||f| \right\|_\infty}{\|y\|_\infty},$$

$$\widehat{\mathscr{C}^{rls}}(y) = \|\mathfrak{D}_y^\dagger |(\widetilde{K} \otimes y^T + \mathbf{r}^T \otimes \widetilde{S}^{-1}) \mathbf{\Phi}_{\mathcal{T}_{mn}} |\mathrm{vec}_{\mathcal{T}}(|Q^T|) + \mathfrak{D}_y^\dagger |\widetilde{K}||f|\|_\infty,$$

where $\widetilde{K} = \widetilde{S}^{-1} Q^T W$ *and* $\widetilde{S} = -(\lambda I_m + Q^T W Q)$.

Proof. Substituting $\mathbf{L} = \begin{bmatrix} \mathbf{0} & I_m \end{bmatrix} \in \mathbb{R}^{m \times (m+n)}$, $B = Q^T, A = W^{-1}, D = -\lambda I_m$ and $x = W(f - Qy)$ in Theorem 5.1.23, the proof follows. ■

Remark 5.1.24. *We consider the Tikhonov regularization problem*

$$\min_{w \in \mathbb{R}^m} \left\{ \|B^T w - f\|_2^2 + \lambda \|Rw\|_2^2 \right\},$$

where R is the regularization matrix and $\lambda > 0$ regularization parameter. Let w be the unique solution of the Tikhonov regularization problem. Then, substituting $\mathbf{L} = \begin{bmatrix} \mathbf{0} & I_m \end{bmatrix} \in \mathbb{R}^{m \times (m+n)}$, $A = I_n, D = -\lambda R^T R$, $x = (f - B^T w)$ and $y = w$, in Theorem 5.1.23, we can recover the structured NCN, MCN, and CCN formulae discussed in [101] for Toeplitz structure.

5.1.6. Numerical Experiments

In order to check the reliability of the proposed structured CNs, we perform several numerical experiments in this section. We construct the perturbations to the input data as follows:

$$\Delta A = 10^{-q} \cdot \Delta A_1 \odot A, \quad \Delta B = 10^{-q} \cdot \Delta B_1 \odot B, \quad \Delta C = 10^{-q} \cdot \Delta C_1 \odot C, \tag{5.1.38}$$

$$\Delta D = 10^{-q} \cdot \Delta D_1 \odot D, \, \Delta f = 10^{-q} \cdot \Delta f_1 \odot f, \text{ and } \Delta g = 10^{-q} \cdot \Delta g_1 \odot g, \tag{5.1.39}$$

where $\Delta A_1 \in \mathbb{R}^{n \times n}, \Delta B_1, \Delta C_1 \in \mathbb{R}^{m \times n}$ and $\Delta D_1 \in \mathbb{R}^{m \times m}$ are the random matrices, preserving the structures of original matrices. Suppose that $[x^T, y^T]^T$ and $[\tilde{x}^T, \tilde{y}^T]^T$ are the unique solutions of the original GSPP and the perturbed GSPP, respectively. To estimate an upper bound for the forward error in the solution, the normwise, mixed, and

componentwise relative errors in $\mathbf{L}[x^T, y^T]^T$, respectively, are defined by:

$$relk = \frac{\|\mathbf{L}[\tilde{x}^T,\, \tilde{y}^T]^T - \mathbf{L}[x^T,\, y^T]^T\|_2}{\|\mathbf{L}[x^T,\, y^T]^T\|_2}, \quad relm = \frac{\|\mathbf{L}[\tilde{x}^T,\, \tilde{y}^T]^T - \mathbf{L}[x^T,\, y^T]^T\|_\infty}{\|\mathbf{L}[x^T,\, y^T]^T\|_\infty},$$
$$\text{and} \quad relc = \left\| \frac{\mathbf{L}[\tilde{x}^T,\, \tilde{y}^T]^T - \mathbf{L}[x^T,\, y^T]^T}{\mathbf{L}[x^T,\, y^T]^T} \right\|_\infty.$$

The following quantities

$$\eta_1 \cdot \mathscr{K}(\mathbf{L}[x^T,\, y^T]^T), \quad \eta_2 \cdot \mathscr{M}(\mathbf{L}[x^T,\, y^T]^T), \quad \eta_2 \cdot \mathbb{C}(\mathbf{L}[x^T,\, y^T]^T) \quad \text{and}$$
$$\eta_1 \cdot \mathscr{K}(\mathbf{L}[x^T,\, y^T]^T; \mathbb{S}), \quad \eta_2 \cdot \mathscr{M}(\mathbf{L}[x^T,\, y^T]^T; \mathbb{S}), \quad \eta_2 \cdot \mathscr{C}(\mathbf{L}[x^T,\, y^T]^T; \mathbb{S}),$$

where $\mathbb{S} = \{\mathcal{E}, \mathcal{L}\}$, are the estimated upper bounds of *relk*, *relm*, and *relc* obtained by the CNs in unstructured and structured cases, respectively. Here, the quantities η_1 and η_2 are defined as [94]:

$$\eta_1 = \begin{cases} \dfrac{\left\|\begin{bmatrix}\Delta\mathbf{H} & \Delta\mathbf{d}\end{bmatrix}\right\|_F}{\left\|\begin{bmatrix}\mathbf{H} & \mathbf{d}\end{bmatrix}\right\|_F}, & \text{when } \mathbb{S} = \mathcal{E}, \\ \dfrac{\left\|\begin{bmatrix}\Delta\mathcal{M} & \Delta\mathbf{d}\end{bmatrix}\right\|_F}{\left\|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right\|_F}, & \text{when } \mathbb{S} = \mathcal{L}, \end{cases}$$

and $\eta_2 = \min\{\eta : \left|\begin{bmatrix}\Delta\mathcal{M} & \Delta\mathbf{d}\end{bmatrix}\right| \leq \eta \left|\begin{bmatrix}\mathcal{M} & \mathbf{d}\end{bmatrix}\right|\}$. We choose the matrix $\mathbf{L}$ as I_{m+n}, $\begin{bmatrix}I_n & \mathbf{0}\end{bmatrix}$, and $\begin{bmatrix}\mathbf{0} & I_m\end{bmatrix}$, so that the CNs for $[x^T,\, y^T]^T$, x and y, respectively, are obtained.

Example 5.1.1. In this example, we consider the GSPP (5.1.3) arising from the WRLS problem [24]. Here $m = n$ and the Toeplitz matrix B is given as follows:

$$B = [b_{ij}] \in \mathcal{T}_{n\times n} \qquad \text{with} \quad b_{ij} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(i-j)^2}{2\sigma^2}},$$

$A \in \mathbb{R}^{n\times n}$ is set to be a positive diagonal random matrix, and $D = -\nu I_n$ ($\nu > 0$). The right hand side vector is taken as $\mathbf{d} = randn(2n, 1) \in \mathbb{R}^{2n}$.

We select $\sigma = 2$ and $\nu = 0.001$ as in [10]. We set $q = 8$ and construct perturbation matrices as in (5.1.38)-(5.1.39) with $\Delta B_1 \in \mathcal{T}^{n\times m}$ and $\Delta A_1 = \frac{1}{2}(\widehat{A} + \widehat{A}^T)$, $\widehat{A}$ is a random matrix. In all cases, we observed $\eta_1 \approx \mathcal{O}(10^{-9})$ and $\eta_2 \approx \mathcal{O}(10^{-8})$. The numerical results for structured and unstructured NCN, MCN, and CCN, and the exact relative errors are reported in Tables 5.1.1-5.1.3 for different values of n. We use Theorem 5.1.9 and Remark 5.1.11 to compute the structured CNs and Theorem 5.1.4 to compute unstructured CNs. The results presented in Tables 5.1.1-5.1.3 reveal that the structured NCN, MCN, and CCN are much smaller than the unstructured ones for all values n. Specifically, for large matrices (with dimensions of $\mathcal{M}$ taken up to 400), the structured CNs are approximately

Table 5.1.1: Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = I_{2n}$ for Example 5.1.1.

$n = m$	$relk$	$\mathscr{K}([x^T, y^T]^T)$	$\mathscr{K}([x^T, y^T]^T; \mathcal{E})$	$relm$	$\mathscr{M}([x^T, y^T]^T)$	$\mathscr{M}([x^T, y^T]^T; \mathcal{E})$	$relc$	$\mathscr{C}([x^T, y^T]^T)$	$\mathscr{C}([x^T, y^T]^T; \mathcal{E})$
50	4.1808e-07	2.8177e+04	2.4798e+04	4.6643e-07	1.4438e+03	5.3588e+02	3.3769e-05	5.7501e+04	2.0978e+04
100	2.4188e-07	4.8911e+03	4.6982e+03	2.5583e-07	1.4305e+02	4.1253e+01	1.4497e-05	1.1440e+04	2.4661e+03
150	5.3749e-07	1.9378e+04	1.7985e+04	6.1184e-07	5.5108e+02	1.3986e+02	2.4998e-04	3.6099e+05	8.1050e+04
200	7.5206e-07	3.2373e+04	9.4706e+03	8.8297e-07	1.0386e+03	4.5302e+02	9.7741e-05	2.0373e+05	4.4730e+04

Table 5.1.2: Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}$ for Example 5.1.1.

$n = m$	$relk$	$\mathscr{K}([x^T, y^T]^T)$	$\mathscr{K}([x^T, y^T]^T; \mathcal{E})$	$relm$	$\mathscr{M}([x^T, y^T]^T)$	$\mathscr{M}([x^T, y^T]^T; \mathcal{E})$	$relc$	$\mathscr{C}([x^T, y^T]^T)$	$\mathscr{C}([x^T, y^T]^T; \mathcal{E})$
50	3.8496e-07	2.7536e+04	2.4281e+04	4.1735e-07	1.2042e+03	4.6611e+02	3.2289e-06	1.0158e+04	3.1020e+03
100	3.0293e-07	7.0491e+03	6.7372e+03	3.7151e-07	2.7486e+02	7.2328e+01	6.2591e-06	6.5226e+03	1.2953e+03
150	7.2376e-07	2.7944e+04	2.5692e+04	8.4422e-07	7.5098e+02	2.0082e+02	6.3056e-05	3.6099e+05	8.1050e+04
200	8.0141e-07	3.6034e+04	3.2041e+04	7.4664e-07	1.0283e+03	4.1526e+02	9.7741e-05	2.0067e+05	4.4730e+04

Table 5.1.3: Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = \begin{bmatrix} \mathbf{0} & I_n \end{bmatrix}$ for Example 5.1.1.

$n = m$	$relk$	$\mathscr{K}([x^T, y^T]^T)$	$\mathscr{K}([x^T, y^T]^T; \mathcal{E})$	$relm$	$\mathscr{M}([x^T, y^T]^T)$	$\mathscr{M}([x^T, y^T]^T; \mathcal{E})$	$relc$	$\mathscr{C}([x^T, y^T]^T)$	$\mathscr{C}([x^T, y^T]^T; \mathcal{E})$
50	4.2087e-07	2.8235e+04	7.2137e+03	4.6643e-07	1.4438e+03	5.3588e+02	3.3769e-05	5.7501e+04	2.0978e+04
100	2.3999e-07	4.8471e+03	1.1173e+03	2.5583e-07	1.4305e+02	4.1253e+01	1.4497e-05	1.1440e+04	2.4661e+03
150	5.2664e-07	1.8878e+04	5.6962e+03	6.1184e-07	5.5108e+02	1.3986e+02	2.4998e-04	9.0978e+04	3.1977e+04
200	7.4779e-07	3.2089e+04	9.2643e+03	8.8297e-07	1.0386e+03	4.5302e+02	5.4363e-05	2.0373e+05	3.4820e+04

an order of magnitude smaller than unstructured ones, showcasing the superiority of proposed structured CNs.

Example 5.1.2. In this example, we consider the GSPP arising from the discretization of the following Stokes equation by upwind scheme [14]:

$$\begin{cases} -\mu\Delta\mathbf{u} + \nabla p = \tilde{f}, & \text{in } \mathbf{\Omega}, \\ \nabla \cdot \mathbf{u} = \tilde{g} & \text{in } \mathbf{\Omega}, \\ \mathbf{u} = \mathbf{0}, & \text{on } \partial\mathbf{\Omega}, \\ \int_{\mathbf{\Omega}} p(x)dx = 0, \end{cases} \tag{5.1.40}$$

where $\mathbf{\Omega} = (0,1) \times (0,1) \in \mathbb{R}^2$, $\partial\mathbf{\Omega}$ is the boundary of $\mathbf{\Omega}$, μ is the viscosity parameter, Δ is the Laplace operator, ∇ represents the gradient, $\nabla\cdot$ is the divergence, $\mathbf{u}$ is the velocity vector, and p is the scalar function representing the pressure. By discretizing (5.1.40), we obtain the GSPP (5.1.1) with

$$A = \begin{bmatrix} I_r \otimes T + T \otimes I_r & \mathbf{0} \\ \mathbf{0} & I_r \otimes T + T \otimes I_r \end{bmatrix} \in \mathbb{R}^{2r^2 \times 2r^2}, \; B^T = \begin{bmatrix} I_r \otimes G \\ G \otimes I_r \end{bmatrix} \in \mathbb{R}^{2r^2 \times r^2},$$

$C = -B$, and $D = \mathbf{0}$, where

$$T = \frac{\mu}{h^2} \text{ tridiag}(-1, 2, -1) \in \mathbb{R}^{r \times r} \text{ and } G = \frac{1}{h} \text{ tridiag}(-1, 1, 0) \in \mathbb{R}^{r \times r}.$$

Note that, for this test problem $\mu = 0.1$, $n = 2r^2$ and $m = r^2$, and we choose $\mathbf{b} = [f^T, g^T]^T$ so that the exact solution is $z = [1, 1, \ldots, 1]^T \in \mathbb{R}^{m+n}$. To avoid making A too sparse, we add $X = 0.5(X_1 + X_1^T)$ to A, where $X_1 = sprandn(m, n, 0.1)$.

Table 5.1.4: Comparison of unstructured and structured NCN, MCN, and CCN with their corresponding relative errors when $\mathbf{L} = I_{m+n}$ for Example 5.1.2.

r	$relk$	$\widetilde{\mathscr{K}}(z)$	$\mathscr{K}([x^T, y^T]^T; \mathcal{L})$	$relm$	$\widetilde{\mathscr{M}}(z)$	$\mathscr{M}([x^T, y^T]^T; \mathcal{L})$	$relc$	$\widetilde{\mathscr{C}}(z)$	$\mathscr{C}([x^T, y^T]^T; \mathcal{L})$
3	4.6396e-08	1.0866e+02	1.0325e+02	9.2530e-08	1.0315e+02	8.3160e+01	9.2530e-08	1.0315e+02	8.3160e+01
4	1.0295e-07	1.2567e+03	1.1946e+03	4.0283e-07	1.1158e+03	8.9754e+02	4.0283e-07	1.1158e+03	8.9754e+02
5	1.3490e-07	1.1905e+03	1.1256e+03	5.3926e-07	1.2062e+03	9.4423e+02	5.3926e-07	1.2062e+03	9.4423e+02
6	1.1442e-07	1.4744e+03	1.3738e+03	3.8692e-07	1.1110e+03	8.5833e+02	3.8692e-07	1.1110e+03	8.5833e+02
7	1.4617e-07	2.5853e+03	2.5366e+03	5.2901e-07	1.1384e+03	9.1026e+02	5.2901e-07	1.1384e+03	8.1026e+02
8	5.1493e-08	2.6605e+03	2.1679e+03	2.0993e-07	1.0634e+03	8.8527e+02	2.0993e-07	1.0634e+03	8.8527e+02
9	7.7302e-08	1.2791e+03	1.0043e+03	2.5382e-07	1.0339e+03	8.2775e+02	2.5382e-07	1.0339e+03	8.2775e+02
10	1.2621e-07	1.5807e+04	1.5205e+04	4.5006e-07	1.0406e+04	8.3004e+03	4.5006e-07	1.0406e+04	8.3004e+03

The perturbations in the input data constructed as in (5.1.38)-(5.1.39) with $q = 8$, $\Delta A_1 = \frac{1}{2}(\widehat{A} + \widehat{A}^T)$, where $\widehat{A} \in \mathbb{R}^{n \times n}$ is random matrix. The numerical result for the structured and unstructured NCN, MCN, and CCN with $\mathbf{L} = I_{m+n}$ are presented in Table 5.1.4 for $r = 3, 4, \ldots, 10$. Since the block matrix A is symmetric, we compute the structured NCN, MCN and CCN using Theorem 5.1.16 and Remark 5.1.19 with $D = \mathbf{0}$. Unstructured CNs are computed using (5.1.31), (5.1.32), and Remark 5.1.20. We observed $\eta_1 \approx \mathcal{O}(10^{-9})$ and $\eta_2 \approx \mathcal{O}(10^{-8})$ in all cases. Results reported in Table 5.1.4 demonstrate that for all values of r, structured MCN and CCN are almost one order smaller than the unstructured MCN and CCN. Moreover, the estimated upper bounds of the relative error

of the solution produced by the structured CNs are sharper than those obtained by the unstructured CNs irrespective of the increasing size of $\mathcal{M}$ (taken up to 300).

5.1.7. Summary

In this section, by considering structure-preserving perturbations on the block matrices, we have investigated structured partial NCN, MCN, and CCN for the GSPP. We derive compact formulae of structured partial CNs in two cases. First, when $B = C$ is Toeplitz and A is symmetric. Second, when $B \neq C$ and the matrices A and D possess linear structures. Furthermore, we have obtained unstructured CNs' formulae for $B = C$, which generalizes the previous results on CNs of GSPP when $\mathbf{L}$ is I_{m+n}, $\begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{0} & I_m \end{bmatrix}$. Additionally, the relations between structured and unstructured CNs are obtained. It is found that the structured CNs are always smaller than their unstructured counterparts. An application of obtained structured CNs formulae is provided to find the structured CNs for WRLS problems, and they are also used to retrieve some prior found results for Tikhonov regularization problems. Numerical experiments are performed to validate the theoretical findings pertaining to proposed structured CNs. Moreover, empirical investigations indicate that the proposed structured MCN and CCN give much more accurate error estimations to the solution of GSPPs compared to unstructured CNs.

5.2. Partial Condition Numbers for Double Saddle Point Problems

This section presents a unified framework for investigating the partial CN for the solution of DSPPs and provides closed-form expressions for it. This unified framework encompasses the well-known partial NCN, MCN and CCN as special cases. Furthermore, we derive sharp upper bounds for the partial NCN, MCN, and CCN, which are computationally efficient and free of expensive Kronecker products. By applying perturbations that preserve the structure of the block matrices of the DSPPs, we analyze the structured partial NCN, MCN and CCN when the block matrices exhibit linear structures. By leveraging the relationship between DSPP and EILS problems, we recover the partial CNs for the EILS problem. Numerical results confirm the sharpness of the derived upper bounds and demonstrate their effectiveness in estimating the partial CNs.

5.2.1. Background

We consider the following linear system with the double saddle point structure:

$$\mathfrak{B}\boldsymbol{w} = \mathbf{d}, \tag{5.2.1}$$

where

$$\mathfrak{B} = \begin{bmatrix} A & B^T & \mathbf{0} \\ B & -D & C^T \\ \mathbf{0} & C & E \end{bmatrix}; \ \boldsymbol{w} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}; \ \mathbf{d} = \begin{bmatrix} f \\ g \\ h \end{bmatrix}; \tag{5.2.2}$$

$A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{p\times m}$, $D \in \mathbb{R}^{m\times m}$, $E \in \mathbb{R}^{p\times p}$, $\mathbf{b} \in \mathbb{R}^l$ and $l = n + m + p$. Conditions on the invertability for the matrix $\mathfrak{B}$ have been studied in [20, 69]. To ensure a unique solution to (5.2.1), throughout the section, we assume that $\mathfrak{B}$ is nonsingular.

Perturbation analysis and CNs for standard SPPs have been extensively studied in the literature; see [100, 147, 151]. However, these studies do not take advantage of the three-by-three block structure of the coefficient matrix $\mathfrak{B}$. Furthermore, they do not provide sensitivity analysis for the individual solution components $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, or for each component of $\boldsymbol{w}$. For the first time, this class of CN was investigated in [38] for the system of linear equations and later extensively studied for various problems in recent years, for instance, in linear least squares (LS) problems [6], weighted LS problems [55], the indefinite LS problems [87, 137], total LS problems [7], and GSPPs [4].

In recent years, the structured CNs of various problems have been studied, emphasizing the preservation of the linear structure of the original matrices in the perturbation matrices, such as linear systems [120, 121], linear LS problems [50], GSPPs [4]. The block matrices A, D, and E often exhibit particular linear structures in various applications; see [111, 115]. This makes it compelling to explore the structured partial CNs for the DSPP (5.2.1) by preserving the linear structures of the diagonal block matrices to their corresponding perturbation matrices.

In this section, we consider the CN of the linear function $\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ of the solution $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$, where $\mathbf{L} \in \mathbb{R}^{k\times l}$ $(k \leq l)$ or the partial CN of the solution $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ of the DSPP (5.2.1). Furthermore, our investigation presents a general framework that encompasses well-known CNs, such as NCN, MCN, and CCN, as special cases.

The key contributions of the section are highlighted as follows:

- In this work, we explore a general form of partial CN, referred to as partial unified CN, which has a versatile nature and provides a comprehensive framework encompassing the NCN, MCN, and CCN of the solution of the DSPP.
- By considering structure-preserving perturbations on A, D, and E, when they retain some linear structures, we derive structured partial CNs for the DSPP.
- By exploring the connection between DSPPs and EILS problems, we demonstrate that our derived CN formula can be used to recover the partial CNs for the EILS problem.
- Numerical experiments demonstrate that the derived upper bounds provide sharp estimates of the partial CNs. Furthermore, the partial CNs offer precise estimates of the relative forward error in the solution.

The structure of the rest of the section is as follows. Subsection 5.2.2 introduces a few notations, basic definitions, and preliminaries. Subsection 5.2.3 presents a unified framework partial CN of the solution of the DSPP (5.2.1). Subsection 5.2.4 focuses on the investigation of structured partial CNs for the DSPP. In Subsection 5.2.5, we discuss the partial CNs for EILS problems. Subsection 5.2.6 consists of some numerical examples. Subsection 5.2.7 includes the concluding statements.

5.2.2. Preliminaries

Following [149], for any vector $z \in \mathbb{R}^n$, we define

$$z^{\ddagger} = [z_1^{\ddagger}, z_2^{\ddagger}, \ldots, z_n^{\ddagger}]^T, \tag{5.2.3}$$

where

$$z_i^{\ddagger} = \begin{cases} \frac{1}{z_i}, & z_i \neq 0, \\ 1, & z_i = 0. \end{cases} \tag{5.2.4}$$

Moreover, the entrywise division of two vectors $z, w \in \mathbb{R}^n$ is defined as follows:

$$\frac{z}{w} = \mathfrak{D}_{w^{\ddagger}} z.$$

Note that, for $z, w \in \mathbb{R}^n$, $\dfrac{z}{w} = w^{\ddagger} \odot z$. For given matrices $A_1, A_2, \ldots, A_n$, we use

$$\mathrm{vec}(\mathbf{X}) := [\mathrm{vec}(A_1)^T,\ \mathrm{vec}(A_2)^T,\ \ldots,\ \mathrm{vec}(A_n)^T]^T,$$

where $\mathbf{X} = (A_1, A_2, \ldots, A_n)$.

Next, we introduce the concept of the general CNs, referred to as the unified CN.

Definition 5.2.1. *[87] Let $\Upsilon : \mathbb{R}^p \mapsto \mathbb{R}^q$ be a continuous mapping defined on an open set $\Omega_\Upsilon \subseteq \mathbb{R}^p$. Then, the unified CN of Υ at $\boldsymbol{v} \in \Omega_\Upsilon$ is defined by*

$$\mathfrak{K}_\Upsilon(\boldsymbol{v}) = \lim_{\boldsymbol{\epsilon}\to 0} \sup_{0<\|\chi^\ddagger \odot \Delta \boldsymbol{v}\|_\tau \leq \boldsymbol{\epsilon}} \frac{\left\|\xi^\ddagger \odot (\Upsilon(\boldsymbol{v}+\Delta\boldsymbol{v}) - \Upsilon(\boldsymbol{v}))\right\|_\gamma}{\|\chi^\ddagger \odot \Delta \boldsymbol{v}\|_\tau}, \tag{5.2.5}$$

where $\xi \in \mathbb{R}^q$, $\chi \in \mathbb{R}^p$ are the parameters such that if some entry of χ is zero, then the corresponding entry of $\Delta\boldsymbol{v}$ must be zero, and $\|\cdot\|_\tau$ and $\|\cdot\|_\gamma$ are two vector norms defined on $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively.

Note that, Definition 5.2.1 leads to the following bound:

$$\left\|\xi^\ddagger \odot (\Upsilon(\boldsymbol{v}+\Delta\boldsymbol{v}) - \Upsilon(\boldsymbol{v}))\right\|_\gamma \leq \mathfrak{K}_\Upsilon(\boldsymbol{v})\|\chi^\ddagger \odot \Delta\boldsymbol{v}\|_\tau + \mathcal{O}(\|\chi^\ddagger \odot \Delta\boldsymbol{v}\|_\tau^2). \tag{5.2.6}$$

Therefore, the forward error in the solution can be estimated using CNs.

Remark 5.2.1. *The unified CN described in Definition 5.2.1 represents a broad generalization of various well-known CNs that have been explored in the literature. For example:*

- ***NCN:*** *Consider $\tau = \gamma = 2$, $\chi = \|\boldsymbol{v}\|_2 \mathbf{1}_p \in \mathbb{R}^p$ with $\boldsymbol{v} \neq \mathbf{0}$ and $\xi = \|\Upsilon(\boldsymbol{v})\|_2 \mathbf{1}_q \in \mathbb{R}^q$ with $\Upsilon(\boldsymbol{v}) \neq \mathbf{0}$, then we obtain the NCN, denoted by $\mathfrak{K}_\Upsilon^{(2)}(\boldsymbol{v})$.*
- ***MCN:*** *Consider $\tau = \gamma = \infty$, $\chi = \boldsymbol{v} \neq \mathbf{0}$, $\xi = \|\Upsilon(\boldsymbol{v})\|_\infty \mathbf{1}_q \in \mathbb{R}^q$ with $\Upsilon(\boldsymbol{v}) \neq \mathbf{0}$, then we obtain the MCN, denoted by $\mathfrak{K}_{mix,\Upsilon}^{\infty}(\boldsymbol{v})$.*
- ***CCN:*** *Consider $\tau = \gamma = \infty$, $\chi = \boldsymbol{v} \neq \mathbf{0}$, and $\xi = \Upsilon(\boldsymbol{v}) \in \mathbb{R}^q$ with $\Upsilon(\boldsymbol{v}) \neq \mathbf{0}$, then we obtain the CCN, denoted by $\mathfrak{K}_{com,\Upsilon}^{\infty}(\boldsymbol{v})$.*

Next, we present a key result that is essential for the following sections. To derive this, let $\mathbf{H} = (A, B, C, D, E)$ and we set

$$\text{vec}(\mathbf{H}) = [\text{vec}(A)^T, \text{vec}(B)^T, \text{vec}(C)^T, \text{vec}(D)^T, \text{vec}(E)^T]^T.$$

Consider $\Delta A, \Delta B, \Delta C, \Delta D, \Delta E$ and $\Delta\mathbf{d}$ are the perturbations on A, B, C, D, E and $\mathbf{d}$, respectively. Further, we denote

$$\Delta\mathfrak{B} = \begin{bmatrix} \Delta A & \Delta B^T & \mathbf{0} \\ \Delta B & -\Delta D & \Delta C^T \\ \mathbf{0} & \Delta C & \Delta E \end{bmatrix},$$

and assume that $\|\Delta\mathfrak{B}\|_2 \leq \boldsymbol{\epsilon}\|\mathfrak{B}\|_2$ and $\|\Delta\mathbf{d}\|_2 \leq \boldsymbol{\epsilon}\|\mathbf{d}\|_2$. Then, we have the following perturbed DSSP:

$$\begin{bmatrix} A+\Delta A & (B+\Delta B)^T & \mathbf{0} \\ B+\Delta B & -(D+\Delta C) & (C+\Delta C)^T \\ \mathbf{0} & C+\Delta C & E+\Delta E \end{bmatrix} \begin{bmatrix} \boldsymbol{x}+\Delta\boldsymbol{x} \\ \boldsymbol{y}+\Delta\boldsymbol{y} \\ \boldsymbol{z}+\Delta\boldsymbol{z} \end{bmatrix} = \begin{bmatrix} f+\Delta f \\ g+\Delta g \\ h+\Delta h \end{bmatrix}, \tag{5.2.7}$$

which has the unique solution $\begin{bmatrix} \boldsymbol{x}+\Delta\boldsymbol{x} \\ \boldsymbol{y}+\Delta\boldsymbol{y} \\ \boldsymbol{z}+\Delta\boldsymbol{z} \end{bmatrix}$ when $\|\mathfrak{B}^{-1}\|_2\|\Delta\mathfrak{B}\|_2 < 1$.

Consequently, we obtain the following important result.

Lemma 5.2.2. *Suppose* $[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ *and* $[(\boldsymbol{x}+\Delta\boldsymbol{x})^T, (\boldsymbol{y}+\Delta\boldsymbol{y})^T, (\boldsymbol{z}+\Delta\boldsymbol{z})^T]^T$ *are the unique solutions of the original DSPP* (5.2.1) *and perturbed DSPP* (5.2.7)*, respectively. Then, the first-order perturbation expression of* $\left[\Delta\boldsymbol{x}^T, \Delta\boldsymbol{y}^T, \Delta\boldsymbol{z}^T\right]^T$ *is given by*

$$\begin{bmatrix} \Delta\boldsymbol{x} \\ \Delta\boldsymbol{y} \\ \Delta\boldsymbol{z} \end{bmatrix} = -\mathfrak{B}^{-1}\begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix}\begin{bmatrix} \mathrm{vec}(\Delta\mathbf{H}) \\ \Delta\mathbf{b} \end{bmatrix} + \mathcal{O}(\boldsymbol{\epsilon}^2), \tag{5.2.8}$$

where

$$\mathcal{G} = \begin{bmatrix} \boldsymbol{x}^T\otimes I_n & I_n\otimes\boldsymbol{y}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{x}^T\otimes I_m & I_m\otimes\boldsymbol{z}^T & -\boldsymbol{y}^T\otimes I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{y}^T\otimes I_p & \mathbf{0} & \boldsymbol{z}^T\otimes I_p \end{bmatrix} \in \mathbb{R}^{l\times\boldsymbol{s}},$$

$$\mathrm{vec}(\Delta\mathbf{H}) = [\mathrm{vec}(\Delta A)^T, \mathrm{vec}(\Delta B)^T, \mathrm{vec}(\Delta C)^T, \mathrm{vec}(\Delta D)^T, \mathrm{vec}(\Delta E)^T]^T,$$

and $\boldsymbol{s} = (n^2+m^2+p^2+nm+mp)$.

Proof. Combining (5.2.1) and (5.2.7), we obtain

$$\begin{bmatrix} A & B^T & \mathbf{0} \\ B & -D & C^T \\ \mathbf{0} & C & E \end{bmatrix}\begin{bmatrix} \Delta\boldsymbol{x} \\ \Delta\boldsymbol{y} \\ \Delta\boldsymbol{z} \end{bmatrix} = \begin{bmatrix} \Delta f \\ \Delta g \\ \Delta h \end{bmatrix} - \begin{bmatrix} \Delta A\boldsymbol{x}+\Delta B^T\boldsymbol{y} \\ \Delta B\boldsymbol{x}-\Delta D\boldsymbol{y}+\Delta C^T\boldsymbol{z} \\ \Delta C\boldsymbol{y}+\Delta E\boldsymbol{z} \end{bmatrix} + \mathcal{O}(\boldsymbol{\epsilon}^2). \tag{5.2.9}$$

Thus, the proof follows by applying the vec operator and utilizing the properties of the Kronecker product on (5.2.9). ■

5.2.3. Partial Unified CNs for the DSPP

This section primarily focuses on developing a unified framework for the partial CN for the solution $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ of the DSPP (5.2.1). As special cases, we also derive the compact formulae and computationally efficient upper bounds for the partial NCN, MCN, and CCN.

To derive the partial unified CN of the DSPP (5.2.1), we define the following mapping:

$$\boldsymbol{\varphi} : \mathbb{R}^{n\times n} \times \mathbb{R}^{m\times n} \times \mathbb{R}^{p\times m} \times \mathbb{R}^{m\times m} \times \mathbb{R}^{p\times p} \times \mathbb{R}^{l} \to \mathbb{R}^{k}$$

$$\boldsymbol{\varphi}(\mathbf{H}, \mathbf{d}) = \mathbf{L} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} = \mathbf{L}\mathfrak{B}^{-1}\mathbf{d}, \tag{5.2.10}$$

where $\mathbf{L} \in \mathbb{R}^{k\times l}(k \leq l)$. Following the Definition 5.2.1, we now define the partial unified CN for the DSPP using the mapping $\boldsymbol{\varphi}$ as follows.

Definition 5.2.2. *Suppose $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ is the unique solution of the DSPP (5.2.1) and $\mathbf{L} \in \mathbb{R}^{k\times l}$. Consider the map $\boldsymbol{\varphi}$ defined as in (5.2.10). Then, the partial unified CN of $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ with respect to (w.r.t.) $\mathbf{L}$ is defined as follows:*

$$\mathfrak{K}_{\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) := \lim_{\boldsymbol{\epsilon}\to 0} \sup_{0<\left\|\operatorname{vec}\left(\Psi^{\ddagger}\odot\Delta\mathbf{H}, \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\right)\right\|_{\tau}\leq\boldsymbol{\epsilon}} \frac{\left\|\xi_{\mathbf{L}}^{\ddagger} \odot (\boldsymbol{\varphi}(\mathbf{H}+\Delta\mathbf{H}, \mathbf{d}+\Delta\mathbf{d}) - \boldsymbol{\varphi}(\mathbf{H}, \mathbf{d}))\right\|_{\gamma}}{\|\operatorname{vec}\left(\Psi^{\ddagger}\odot\Delta\mathbf{H}, \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\right)\|_{\tau}},$$

where $\xi_{\mathbf{L}} \in \mathbb{R}^{k}$, $\Psi = (\Psi_A, \Psi_B, \Psi_C, \Psi_D, \Psi_E)$, $\Psi_A \in \mathbb{R}^{n\times n}$, $\Psi_B \in \mathbb{R}^{m\times n}$, $\Psi_C \in \mathbb{R}^{p\times m}$, $\Psi_D \in \mathbb{R}^{m\times m}$, $\Psi_E \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\chi} \in \mathbb{R}^{\boldsymbol{l}}$ are the parameters with the assumptions that if some entries of Ψ and $\boldsymbol{\chi}$ are zero, then the corresponding entry of $\Delta\mathbf{H}$ and $\Delta\mathbf{d}$, respectively, must be zero.

Remark 5.2.3. *In the context of Remark 5.2.1, to obtain the partial NCN, we consider $\xi_{\mathbf{L}} = \|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_2\mathbf{1}_l$, for the partial MCN, we consider $\xi_{\mathbf{L}} = \|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_\infty\mathbf{1}_l$, and for partial CCN, we consider $\xi_{\mathbf{L}} = \mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$.*

In the following theorem, we provide a compact and closed-form expression for the partial unified CN.

Theorem 5.2.4. *Suppose $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ is the unique solution of the DSPP (5.2.1) and $\mathbf{L} \in \mathbb{R}^{k\times l}$. Then, the partial unified CN of $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ w.r.t. $\mathbf{L}$ is given by*

$$\mathfrak{K}_{\varphi}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \left\| \mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}} \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\operatorname{vec}(\Psi)} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \right\|_{\tau,\gamma}, \tag{5.2.11}$$

where $\|\cdot\|_{\tau,\gamma}$ is the matrix norm induced by vector norms $\|\cdot\|_\tau$ and $\|\cdot\|_\gamma$.

Proof. From the definition of the mapping $\boldsymbol{\varphi}$ in (5.2.10) and Lemma 5.2.2, we get

$$\begin{aligned}\boldsymbol{\varphi}(\mathbf{H}+\Delta\mathbf{H},\mathbf{d}+\Delta\mathbf{d})-\boldsymbol{\varphi}(\mathbf{H},\mathbf{d}) &= \mathbf{L}\begin{bmatrix}\boldsymbol{x}+\Delta\boldsymbol{x}\\ \boldsymbol{y}+\Delta\boldsymbol{y}\\ \boldsymbol{z}+\Delta\boldsymbol{z}\end{bmatrix}-\mathbf{L}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{bmatrix}\\ &= \mathbf{L}\begin{bmatrix}\Delta\boldsymbol{x}\\ \Delta\boldsymbol{y}\\ \Delta\boldsymbol{z}\end{bmatrix}\\ &= -\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\mathcal{G} & -I_l\end{bmatrix}\begin{bmatrix}\operatorname{vec}(\Delta\mathbf{H})\\ \Delta\mathbf{d}\end{bmatrix}+\mathcal{O}(\boldsymbol{\epsilon}^2). \qquad (5.2.12)\end{aligned}$$

By considering the requirement on Ψ and $\boldsymbol{\chi}$, we have

$$\begin{bmatrix}\operatorname{vec}(\Delta\mathbf{H})\\ \Delta\mathbf{d}\end{bmatrix}=\begin{bmatrix}\mathfrak{D}_{\operatorname{vec}(\Psi)} & \mathbf{0}\\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}}\end{bmatrix}\begin{bmatrix}\operatorname{vec}(\Psi^{\ddagger}\odot\Delta\mathbf{H})\\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\end{bmatrix}. \qquad (5.2.13)$$

Substituting (5.2.13) into (5.2.12) and from Definition 5.2.2, we obtain

$$\begin{aligned}\mathfrak{K}_{\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) &= \sup_{\left\|\operatorname{vec}\left(\Psi^{\ddagger}\odot\Delta\mathbf{H},\boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\right)\right\|_{\tau}\neq 0}\frac{\left\|\mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}}\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\mathcal{G} & -I_l\end{bmatrix}\begin{bmatrix}\mathfrak{D}_{\operatorname{vec}(\Psi)} & \mathbf{0}\\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}}\end{bmatrix}\begin{bmatrix}\operatorname{vec}(\Psi^{\ddagger}\odot\Delta\mathbf{H})\\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\end{bmatrix}\right\|_{\gamma}}{\left\|\operatorname{vec}\left(\Psi^{\ddagger}\odot\Delta\mathbf{H},\boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\right)\right\|_{\tau}}\\ &= \left\|\mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}}\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\mathcal{G} & -I_l\end{bmatrix}\begin{bmatrix}\mathfrak{D}_{\operatorname{vec}(\Psi)} & \mathbf{0}\\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}}\end{bmatrix}\right\|_{\tau,\gamma}. \qquad (5.2.14)\end{aligned}$$

Hence, the proof is completed. ■

Next, we derive various partial CNs by considering specific norms. In the following result, we focus on when $\tau=\gamma=2$.

Theorem 5.2.5. *Consider $\tau=\gamma=2$ and assuming that Ψ, $\boldsymbol{\chi}$ and $\xi_{\mathbf{L}}$ are positive real numbers, then the partial CN has the following forms:*

$$\mathfrak{K}_{\boldsymbol{\varphi}}^{(2)}(\mathbf{H},\mathbf{d};\mathbf{L})=\frac{\left\|\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\Psi\mathcal{G} & -\boldsymbol{\chi}I_l\end{bmatrix}\right\|_2}{\xi_{\mathbf{L}}} \quad \textit{and} \qquad (5.2.15)$$

$$\widehat{\mathfrak{K}}_{\boldsymbol{\varphi}}^{(2)}(\mathbf{H},\mathbf{d};\mathbf{L})=\frac{\left\|\mathbf{L}\mathfrak{B}^{-1}(\Psi^2\mathcal{J}+\boldsymbol{\chi}^2I_l)(\mathfrak{B}^{-1})^T\mathbf{L}^T\right\|_2^{1/2}}{\xi_{\mathbf{L}}}, \qquad (5.2.16)$$

where $\mathcal{J}\in\mathbb{R}^{\boldsymbol{l}\times\boldsymbol{l}}$ is given by

$$\mathcal{J}=\begin{bmatrix}(\|\boldsymbol{x}\|_2^2+\|\boldsymbol{y}\|_2^2)I_n & \boldsymbol{x}\boldsymbol{y}^T & \mathbf{0}\\ \boldsymbol{y}\boldsymbol{x}^T & (\|\boldsymbol{x}\|_2^2+\|\boldsymbol{y}\|_2^2+\|\boldsymbol{z}\|_2^2)I_m & \boldsymbol{y}\boldsymbol{z}^T\\ \mathbf{0} & \boldsymbol{z}\boldsymbol{y}^T & (\|\boldsymbol{y}\|_2^2+\|\boldsymbol{z}\|_2^2)I_p\end{bmatrix}.$$

Proof. Since $\tau = \gamma = 2$ and Ψ, $\boldsymbol{\chi}$ and $\xi_{\mathbf{L}}$ are positive real numbers, from Theorem 5.2.4, we obtain

$$\mathfrak{K}^{(2)}_{1,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) = \frac{\left\|\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\Psi\mathcal{G} & -\boldsymbol{\chi}I_l\end{bmatrix}\right\|_2}{\xi_{\mathbf{L}}}. \tag{5.2.17}$$

Using the property that, for $Z \in \mathbb{R}^{m\times n}$, $\|Z\|_2 = \|ZZ^T\|_2^{1/2}$, we have

$$\begin{aligned}\left\|\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\Psi\mathcal{G} & -\boldsymbol{\chi}I_l\end{bmatrix}\right\|_2 &= \left\|\mathbf{L}\mathfrak{B}^{-1}(\Psi^2\mathcal{G}\mathcal{G}^T + \boldsymbol{\chi}^2 I_l)(\mathfrak{B}^{-1})^T\mathbf{L}^T\right\|_2^{1/2} \\ &= \left\|\mathbf{L}\mathfrak{B}^{-1}(\Psi^2\mathcal{J} + \boldsymbol{\chi}^2 I_l)(\mathfrak{B}^{-1})^T\mathbf{L}^T\right\|_2^{1/2}. \end{aligned}\tag{5.2.18}$$

Hence, the proof follows by substituting (5.2.18) into (5.2.17). ■

Remark 5.2.6. *Notably, the equivalent expression of* $\widehat{\mathfrak{K}}^{(2)}_{\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L})$ *in* (5.2.16) *is free of computationally expensive Kronecker products. Moreover, the matrices in* (5.2.16) *and* (5.2.15) *have dimensions* $k \times k$ *and* $k \times (l + \boldsymbol{s})$ *respectively. Hence, the expression in* (5.2.16) *significantly reduces the storage requirements.*

Remark 5.2.7. *The partial CN in Theorem 5.2.5 is a simplified version of the partial NCN for the solution of the DSPP* (5.2.1). *The NCN for* $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T, \boldsymbol{x}, \boldsymbol{y}$ *and* $\boldsymbol{z}$ *can be obtained by considering*

$$\mathbf{L} = I_{\boldsymbol{l}},\ \begin{bmatrix}I_n & \mathbf{0}_{n\times(m+p)}\end{bmatrix},\ \begin{bmatrix}\mathbf{0}_{m\times n} & I_m & \mathbf{0}_{m\times p}\end{bmatrix} \text{ and } \begin{bmatrix}\mathbf{0}_{p\times(n+m)} & I_p\end{bmatrix},$$

respectively, in Theorem 5.2.5.

In the next result, we provide an easily computable upper bound for the partial CN $\mathfrak{K}^{(2)}_{\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L})$.

Corollary 5.2.1. *Under the assumption of Theorem 5.2.5, we have following upper bound:*

$$\mathfrak{K}^{(2)}_{1,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) \le \mathfrak{K}^{(2),u}_{1,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) := \frac{\|\mathbf{L}\mathfrak{B}^{-1}\|_2}{\xi_{\mathbf{L}}}\left(\Psi\|\mathcal{J}\|_2^{1/2} + \boldsymbol{\chi}\right). \tag{5.2.19}$$

Proof. Using the properties of the spectral norm that for the matrices X and Y of appropriate sizes, $\left\|\begin{bmatrix}X & Y\end{bmatrix}\right\|_2 \le \|X\|_2 + \|Y\|_2$ and $\|XY\|_2 \le \|X\|_2\|Y\|_2$, and from (5.2.15), we obtain

$$\begin{aligned}\left\|\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\Psi\mathcal{G} & -\boldsymbol{\chi}I_l\end{bmatrix}\right\|_2 &\le \Psi\|\mathbf{L}\mathfrak{B}^{-1}\mathcal{G}\|_2 + \boldsymbol{\chi}\|\mathbf{L}\mathfrak{B}^{-1}\|_2 \\ &\le \Psi\|\mathbf{L}\mathfrak{B}^{-1}\|_2\|\mathcal{G}\|_2 + \boldsymbol{\chi}\|\mathbf{L}\mathfrak{B}^{-1}\|_2 \\ &= \Psi\|\mathbf{L}\mathfrak{B}^{-1}\|_2\|\mathcal{J}\|_2^{1/2} + \boldsymbol{\chi}\|\mathbf{L}\mathfrak{B}^{-1}\|_2. \end{aligned}\tag{5.2.20}$$

Hence, the proof follows from (5.2.15) and (5.2.20). ■

In the following theorem, we investigate the partial CN for the DSPP when $\tau = \gamma = \infty$, from which we derive the partial MCN and CCN.

Theorem 5.2.8. *When $\tau = \gamma = \infty$, the partial CN is given as follows:*

$$\mathfrak{K}^{\infty}_{\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \left\| |\mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}}| \left| \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}(|\Psi|) \\ |\boldsymbol{\chi}| \end{bmatrix} \right\|_{\infty}. \tag{5.2.21}$$

Moreover, we consider $\Psi = \mathbf{H}$, $\boldsymbol{\chi} = \mathbf{d}$. Then, if we set $\xi_{\mathbf{L}} = \|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_{\infty}\mathbf{1}_l$, the partial MCN is given by

$$\mathfrak{K}^{\infty}_{mix,\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \frac{\left\| \left| \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}(|\mathbf{H}|) \\ |\mathbf{d}| \end{bmatrix} \right\|_{\infty}}{\|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_{\infty}} \tag{5.2.22}$$

and if we set $\xi_{\mathbf{L}} = \mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$, the partial CCN is given by

$$\mathfrak{K}^{\infty}_{com,\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \left\| \frac{\left| \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}(|\mathbf{H}|) \\ |\mathbf{d}| \end{bmatrix}}{|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T|} \right\|_{\infty}. \tag{5.2.23}$$

Proof. Consider $\tau = \gamma = \infty$, then from Theorem 5.2.4, we have

$$\begin{aligned}
\mathfrak{K}^{\infty}_{\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) &= \left\| \mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}} \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathrm{vec}(\Psi)} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \right\|_{\infty} \\
&= \left\| |\mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}}| \left| \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \right| \begin{bmatrix} |\mathfrak{D}_{\mathrm{vec}(\Psi)}| & \mathbf{0} \\ \mathbf{0} & |\mathfrak{D}_{\boldsymbol{\chi}}| \end{bmatrix} \right\|_{\infty} \\
&= \left\| |\mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}}| \left| \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \right| \begin{bmatrix} |\mathfrak{D}_{\mathrm{vec}(\Psi)}| & \mathbf{0} \\ \mathbf{0} & |\mathfrak{D}_{\boldsymbol{\chi}}| \end{bmatrix} \mathbf{1}_{s+l} \right\|_{\infty} \\
&= \left\| |\mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}}| \left| \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}(|\Psi|) \\ |\boldsymbol{\chi}| \end{bmatrix} \right\|_{\infty}.
\end{aligned}$$

Rest of the proof followings considering $\Psi = \mathbf{H}$, $\boldsymbol{\chi} = \mathbf{d}$, and $\xi_{\mathbf{L}} = \|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_{\infty}\mathbf{1}_l$ (or $\xi_{\mathbf{L}} = \mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$). ■

Remark 5.2.9. *The MCN and CCN for $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T, \boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ can be obtained by considering*

$$\mathbf{L} = I_l, \ \begin{bmatrix} I_n & \mathbf{0}_{n\times(m+p)} \end{bmatrix}, \ \begin{bmatrix} \mathbf{0}_{m\times n} & I_m & \mathbf{0}_{m\times p} \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{0}_{p\times(n+m)} & I_p \end{bmatrix},$$

respectively, in (5.2.22) and (5.2.23) of Theorem 5.2.8.

In the following result, we provide sharp upper bounds for the partial MCN and CCN obtained in Theorem 5.2.8.

Corollary 5.2.2. *Assume that the conditions in Theorem 5.2.8 hold. Then*

$$\mathfrak{K}^{\infty}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) \leq \mathfrak{K}^{\infty,u}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) := \frac{\||\mathbf{L}\mathfrak{B}^{-1}|\,(|\mathcal{H}|+|\mathbf{d}|)\|_{\infty}}{\|\mathbf{L}[\boldsymbol{x}^T,\boldsymbol{y}^T,\boldsymbol{z}^T]^T\|_{\infty}}$$

and

$$\mathfrak{K}^{\infty}_{com,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) \leq \mathfrak{K}^{\infty,u}_{com,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) := \left\| \frac{|\mathbf{L}\mathfrak{B}^{-1}|\,(|\mathcal{H}|+|\mathbf{d}|)}{\mathbf{L}[\boldsymbol{x}^T,\boldsymbol{y}^T,\boldsymbol{z}^T]^T} \right\|_{\infty},$$

where $\mathcal{H} = \begin{bmatrix} |A||\boldsymbol{x}|+|B^T||\boldsymbol{y}| \\ |B||\boldsymbol{x}|+|D^T||\boldsymbol{y}|+|C^T||\boldsymbol{z}| \\ |C||\boldsymbol{y}|+|E||\boldsymbol{z}| \end{bmatrix}.$

Proof. Utilizing the properties of Kronecker product in (1.3.2), we have

$$\begin{aligned}
&\left|\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\mathcal{G} & -I_l\end{bmatrix}\right|\begin{bmatrix}\mathrm{vec}(|\mathbf{H}|)\\ |\mathbf{d}|\end{bmatrix} \leq |\mathbf{L}\mathfrak{B}^{-1}|\begin{bmatrix}|\mathcal{G}| & I_l\end{bmatrix}\begin{bmatrix}\mathrm{vec}(|\mathbf{H}|)\\ |\mathbf{d}|\end{bmatrix}\\
&= |\mathbf{L}\mathfrak{B}^{-1}|(|\mathcal{G}|\mathrm{vec}(|\mathbf{H}|)+|\mathbf{d}|)\\
&= |\mathbf{L}\mathfrak{B}^{-1}|\left(\begin{bmatrix}(|\boldsymbol{x}|^T\otimes I_n)\mathrm{vec}(|A|)+(I_n\otimes|\boldsymbol{y}|^T)\mathrm{vec}(|B|)\\ (|\boldsymbol{x}|^T\otimes I_m)\mathrm{vec}(|B|)+(I_m\otimes|\boldsymbol{z}|^T)\mathrm{vec}(|C|)+(|\boldsymbol{y}|^T\otimes I_m)\mathrm{vec}(|D|)\\ (|\boldsymbol{y}|^T\otimes I_p)\mathrm{vec}(|C|)+(|\boldsymbol{z}|^T\otimes I_p)\mathrm{vec}(|E|)\end{bmatrix}+|\mathbf{d}|\right)\\
&= |\mathbf{L}\mathfrak{B}^{-1}|\left(\left|\begin{bmatrix}|A||\boldsymbol{x}|+|B^T||\boldsymbol{y}|\\ |B||\boldsymbol{x}|+|D^T||\boldsymbol{y}|+|C^T||\boldsymbol{z}|\\ |C||\boldsymbol{y}|+|E||\boldsymbol{z}|\end{bmatrix}\right|+|\mathbf{d}|\right). \qquad (5.2.24)
\end{aligned}$$

From (5.2.24) and the expressions of partial MCN and CCN in Theorem 5.2.8, we get

$$\mathfrak{K}^{\infty}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) \leq \mathfrak{K}^{\infty,u}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) \text{ and } \mathfrak{K}^{\infty}_{com,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) \leq \mathfrak{K}^{\infty,u}_{com,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}).$$

Hence, the proof follows. ■

5.2.4. Structured Partial CNs

Consider three subspaces $\mathbb{S}_1 \subseteq \mathbb{R}^{n\times n}$, $\mathbb{S}_2 \subseteq \mathbb{R}^{m\times m}$ and $\mathbb{S}_3 \subseteq \mathbb{R}^{p\times p}$ consisting of three distinct linear structured matrices, such as symmetric and Toeplitz. Suppose that the corresponding dimensions of the linear subspaces are s, r and q, respectively. Let $A \in \mathbb{S}_1$,

$D \in \mathbb{S}_2$ and $E \in \mathbb{S}_3$, then according to [72, 86, 120], there exist unique generating vectors $\boldsymbol{a} \in \mathbb{R}^s$, $\boldsymbol{d} \in \mathbb{R}^r$ and $\boldsymbol{e} \in \mathbb{R}^q$ such that

$$\operatorname{vec}(A) = \boldsymbol{\Phi}_{\mathbb{S}_1}\boldsymbol{a}, \ \operatorname{vec}(D) = \boldsymbol{\Phi}_{\mathbb{S}_2}\boldsymbol{d} \ \text{ and } \ \operatorname{vec}(E) = \boldsymbol{\Phi}_{\mathbb{S}_3}\boldsymbol{e}, \tag{5.2.25}$$

where $\boldsymbol{\Phi}_{\mathbb{S}_1} \in \mathbb{R}^{n^2\times s}$, $\boldsymbol{\Phi}_{\mathbb{S}_2} \in \mathbb{R}^{m^2\times r}$ and $\boldsymbol{\Phi}_{\mathbb{S}_3} \in \mathbb{R}^{m^2\times q}$. These matrices are fixed for each specific structure and encapsulate the information corresponding to the linear structure of their respective subspaces.

Let $\operatorname{vec}_{\mathbb{S}}(\mathbf{H}) = [\boldsymbol{a}^T, \operatorname{vec}(B)^T, \operatorname{vec}(C)^T, \boldsymbol{d}^T, \boldsymbol{e}^T]^T$. Then the structured partial CN for solution $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ of the DSPP (5.2.1) w.r.t. $\mathbf{L}$ is given by

$$\mathfrak{K}^{\mathbb{S}}_{\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) := \lim_{\boldsymbol{\epsilon}\to 0} \sup_{\substack{0<\left\|\operatorname{vec}\left(\Psi^{\ddagger}\odot\Delta\mathbf{H}, \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\right)\right\|_{\tau}\leq\boldsymbol{\epsilon} \\ \Delta A\in\mathbb{S}_1, \Delta D\in\mathbb{S}_2, \Delta E\in\mathbb{S}_3}} \frac{\left\|\xi^{\ddagger}_{\mathbf{L}} \odot (\boldsymbol{\varphi}(\mathbf{H}+\Delta\mathbf{H}, \mathbf{d}+\Delta\mathbf{d}) - \boldsymbol{\varphi}(\mathbf{H}, \mathbf{d}))\right\|_{\gamma}}{\left\|\operatorname{vec}\left(\Psi^{\ddagger}\odot\Delta\mathbf{H}, \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\right)\right\|_{\tau}}, \tag{5.2.26}$$

where $\xi_{\mathbf{L}} \in \mathbb{R}^k$, $\Psi = (\Psi_A, \Psi_B, \Psi_C, \Psi_D, \Psi_E)$, $\Psi_A \in \mathbb{S}_1$, $\Psi_B \in \mathbb{R}^{m\times n}$, $\Psi_C \in \mathbb{R}^{p\times m}$, $\Psi_D \in \mathbb{S}_2$, $\Psi_E \in \mathbb{S}_3$ and $\boldsymbol{\chi} \in \mathbb{R}^l$.

Since the matrices $\Delta A, \Psi_A \in \mathbb{S}_1$, $\Delta C, \Psi_D \in \mathbb{S}_2$ and $\Delta E, \Psi_E \in \mathbb{S}_3$, as in (5.2.25), we have

$$\operatorname{vec}(\Delta A) = \boldsymbol{\Phi}_{\mathbb{S}_1}\Delta\boldsymbol{a}, \ \operatorname{vec}(\Psi_A) = \boldsymbol{\Phi}_{\mathbb{S}_1}\psi_A, \ \operatorname{vec}(\Delta C) = \boldsymbol{\Phi}_{\mathbb{S}_2}\Delta\boldsymbol{d}, \tag{5.2.27}$$

$$\operatorname{vec}(\Psi_D) = \boldsymbol{\Phi}_{\mathbb{S}_2}\psi_D, \ \operatorname{vec}(\Delta E) = \boldsymbol{\Phi}_{\mathbb{S}_3}\Delta\boldsymbol{e}, \ \text{and} \ \operatorname{vec}(\Psi_E) = \boldsymbol{\Phi}_{\mathbb{S}_3}\psi_E, \tag{5.2.28}$$

where $\Delta\boldsymbol{a}$, $\Delta\boldsymbol{d}$, $\Delta\boldsymbol{e}$, ψ_D, ψ_A and ψ_E are the unique generating vectors of ΔA, ΔC, ΔE, Ψ_A, Ψ_D and Ψ_E, respectively. Note that, $\Psi^{\ddagger}_A \in \mathbb{S}_1$, $\Psi^{\ddagger}_D \in \mathbb{S}_2$ and $\Psi^{\ddagger}_E \in \mathbb{S}_3$, consequently, we obtain

$$\operatorname{vec}(\Psi^{\ddagger}_A \odot \Delta A) = \boldsymbol{\Phi}_{\mathbb{S}_1}(\psi^{\ddagger}_A \odot \Delta\boldsymbol{a}), \quad \operatorname{vec}(\Psi^{\ddagger}_D \odot \Delta C) = \boldsymbol{\Phi}_{\mathbb{S}_2}(\psi^{\ddagger}_D \odot \Delta\boldsymbol{d}) \tag{5.2.29}$$

$$\text{and} \ \ \operatorname{vec}(\Psi^{\ddagger}_E \odot \Delta E) = \boldsymbol{\Phi}_{\mathbb{S}_3}(\psi^{\ddagger}_E \odot \Delta\boldsymbol{e}). \tag{5.2.30}$$

Subsequently, we obtain the following result.

Lemma 5.2.10. *Let $\Delta A, \Psi_A \in \mathbb{S}_1$, $\Delta C, \Psi_D \in \mathbb{S}_2$, $\Delta E, \Psi_E \in \mathbb{S}_3$, $B, \Psi_B \in \mathbb{R}^{m\times n}$, $C, \Psi_C \in \mathbb{R}^{p\times m}$, and $\mathbf{b}, \boldsymbol{\chi} \in \mathbb{R}^{\boldsymbol{l}}$. Then, we have*

$$\begin{bmatrix} \operatorname{vec}(\Psi^{\ddagger}\odot\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_{\boldsymbol{l}} \end{bmatrix} \begin{bmatrix} \operatorname{vec}_{\mathbb{S}}(\Psi^{\ddagger}\odot H) \\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{b} \end{bmatrix}, \tag{5.2.31}$$

where $\operatorname{vec}_{\mathbb{S}}(\Psi^{\ddagger}\odot\mathbf{H}) = [(\psi^{\ddagger}_A\odot\Delta\boldsymbol{a})^T, (\Psi^{\ddagger}_B\odot\operatorname{vec}(\Delta B))^T, (\Psi^{\ddagger}_C\odot\operatorname{vec}(\Delta C))^T, (\psi^{\ddagger}_D\odot\Delta\boldsymbol{d})^T, (\psi^{\ddagger}_D\odot\Delta\boldsymbol{e})^T]^T$ *and*

$$\boldsymbol{\Phi}_{\mathbb{S}} = \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{mn+mp} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Phi}_{\mathbb{S}_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Phi}_{\mathbb{S}_3} \end{bmatrix}.$$

Proof. The proof follows using identities in (5.2.29) and (5.2.30). ■

In the following theorem, we present closed-form expressions for the structured partial CN by considering $\tau = \gamma = 2$.

Theorem 5.2.11. *Let $A \in \mathbb{S}_1$, $D \in \mathbb{S}_2$, $E \in \mathbb{S}_3$ and $\mathbf{L} \in \mathbb{R}^{k\times l}$. Suppose that $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ is the unique solution of DSPP (5.2.1). Then the structured partial CN of $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ w.r.t. $\mathbf{L}$ is given by*

$$\mathfrak{K}_{\boldsymbol{\varphi}}^{(2),\mathbb{S}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \left\| \mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}} \mathbf{L} \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathrm{vec}(\Psi)} \boldsymbol{\Phi}_{\mathbb{S}} \mathfrak{D}_{\mathbb{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \right\|_2,$$

where

$$\mathfrak{D}_{\mathbb{S}} = \begin{bmatrix} \mathfrak{D}_{\boldsymbol{u}_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{mn+mp} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathfrak{D}_{\boldsymbol{u}_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathfrak{D}_{\boldsymbol{u}_3} \end{bmatrix},$$

$$\boldsymbol{u}_1 = [\|\boldsymbol{\Phi}_{\mathbb{S}_1}(:,1)\|_2, \ldots, \|\boldsymbol{\Phi}_{\mathbb{S}_1}(:,s)\|_2]^T, \quad \boldsymbol{u}_2 = [\|\boldsymbol{\Phi}_{\mathbb{S}_2}(:,1)\|_2, \ldots, \|\boldsymbol{\Phi}_{\mathbb{S}_2}(:,k)\|_2]^T,$$

$$\textit{and} \quad \boldsymbol{u}_3 = [\|\boldsymbol{\Phi}_{\mathbb{S}_3}(:,1)\|_2, \ldots, \|\boldsymbol{\Phi}_{\mathbb{S}_3}(:,q)\|_2]^T.$$

Proof. Taking $\tau = \gamma = 2$ in (5.2.26), and using (5.2.13) and (5.2.12), we obtain

$$\mathfrak{K}_{\boldsymbol{\varphi}}^{(2),\mathbb{S}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \sup_{\substack{\|\mathrm{vec}(\Psi^{\ddagger}\odot\Delta\mathbf{H}, \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d})\|_2 \neq 0 \\ \Delta A\in\mathbb{S}_1, \Delta D\in\mathbb{S}_2, \Delta E\in\mathbb{S}_3}} \frac{\left\| \mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}} \mathbf{L} \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathrm{vec}(\Psi)} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \begin{bmatrix} \mathrm{vec}(\Psi^{\ddagger}\odot\Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d} \end{bmatrix} \right\|_2}{\|\mathrm{vec}\left(\Psi^{\ddagger}\odot\Delta\mathbf{H}, \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d}\right)\|_2}.$$

Substituting (5.2.31) into the above equation yields

$$\mathfrak{K}_{\boldsymbol{\varphi}}^{(2),\mathbb{S}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \sup_{\substack{\left\| \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger}\odot\Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d} \end{bmatrix} \right\|_2 \neq 0, \\ \Delta A\in\mathbb{S}_1, \Delta D\in\mathbb{S}_2, \Delta E\in\mathbb{S}_3}} \frac{\left\| \mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}} \mathbf{L} \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathrm{vec}(\Psi)} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger}\odot\Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d} \end{bmatrix} \right\|_2}{\left\| \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger}\odot\Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d} \end{bmatrix} \right\|_2}. \tag{5.2.32}$$

Utilizing the fact that the matrices $\mathbf{\Phi}_{\mathbb{S}_i}$ for $i = 1, 2, 3$, are column orthogonal [86], we get $\mathbf{\Phi}_{\mathbb{S}_i}^T \mathbf{\Phi}_{\mathbb{S}_i} = \mathfrak{D}_{\boldsymbol{u}_i}^2$, where $\mathfrak{D}_{\boldsymbol{u}_i}$ for $i = 1, 2, 3$, are the diagonal matrices. Then

$$\left\| \begin{bmatrix} \mathbf{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix}^T \begin{bmatrix} \mathbf{\Phi}_{\mathbb{S}}^T \mathbf{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_2^{1/2}$$
$$= \left\| \begin{bmatrix} \mathfrak{D}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_2. \tag{5.2.33}$$

Observe that

$$\begin{bmatrix} \mathbf{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{\Phi}_{\mathbb{S}} \mathfrak{D}_{\mathbb{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix}. \tag{5.2.34}$$

Therefore, substituting (5.2.33) and (5.2.34) in (5.2.32), we obtain

$$\mathfrak{K}_{\boldsymbol{\varphi}}^{(2),\mathbb{S}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \sup_{\substack{\left\| \begin{bmatrix} \mathfrak{D}_{\mathbb{S}} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_2 \neq 0 \\ \Delta A \in \mathbb{S}_1, \Delta D \in \mathbb{S}_2, \Delta E \in \mathbb{S}_3}} \frac{\left\| \mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}} \mathbf{L} \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathrm{vec}(\Psi)} \mathbf{\Phi}_{\mathbb{S}} \mathfrak{D}_{\mathbb{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathbb{S}} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_2}{\left\| \begin{bmatrix} \mathfrak{D}_{\mathbb{S}} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_2}$$
$$= \left\| \mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}} \mathbf{L} \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathrm{vec}(\Psi)} \mathbf{\Phi}_{\mathbb{S}} \mathfrak{D}_{\mathbb{S}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \right\|_2.$$

Hence, the proof is completed. ■

Next, we consider $\tau = \gamma = \infty$, and derive the structured partial MCN and CCN for the DSPP.

Theorem 5.2.12. *Let $A \in \mathbb{S}_1$, $D \in \mathbb{S}_2$, $E \in \mathbb{S}_3$ and $\mathbf{L} \in \mathbb{R}^{k \times l}$. Suppose that $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ is the unique solution of DSPP (5.2.1). When $\tau = \gamma = \infty$, the structured partial CN of the solution $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ w.r.t. $\mathbf{L}$ is given as follows:*

$$\mathfrak{K}_{\boldsymbol{\varphi}}^{\infty,\mathbb{S}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \left\| |\mathfrak{D}_{\xi_{\mathbf{L}}^{\ddagger}}| \left| \mathbf{L} \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(|\Psi|) \\ |\boldsymbol{\chi}| \end{bmatrix} \right\|_{\infty}. \tag{5.2.35}$$

Moreover, we consider $\Psi = \mathbf{H}$ and $\boldsymbol{\chi} = \mathbf{d}$. Then, if we set $\xi_{\mathbf{L}} = \|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_{\infty} \mathbf{1}_l$, the structured partial MCN is given by

$$\mathfrak{K}_{mix,\boldsymbol{\varphi}}^{\infty,\mathbb{S}}(\mathbf{H}, \mathbf{d}; \mathbf{L}) = \frac{\left\| \left| \mathbf{L} \mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \mathbf{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(|\mathbf{H}|) \\ |\mathbf{d}| \end{bmatrix} \right\|_{\infty}}{\|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_{\infty}} \tag{5.2.36}$$

and if we set $\xi_{\mathbf{L}} = \mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$*, the structured partial CCN is given by*

$$\mathfrak{K}^{\infty,\mathbb{S}}_{com,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) = \left\| \frac{\left| \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(|\mathbf{H}|) \\ |\mathbf{d}| \end{bmatrix}}{|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T|} \right\|_{\infty}. \tag{5.2.37}$$

Proof. By construction of the matrices $\boldsymbol{\Phi}_{\mathbb{S}_1}$, $\boldsymbol{\Phi}_{\mathbb{S}_2}$ and $\boldsymbol{\Phi}_{\mathbb{S}_3}$, they have at most one nonzero element in each row. Thus, we get

$$\left\| \begin{bmatrix} \mathrm{vec}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_{\infty} = \left\| \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_{\infty} = \left\| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_{\infty}. \tag{5.2.38}$$

By considering $\tau = \gamma = \infty$ on (5.2.26) and using (5.2.38), (5.2.13) and (5.2.12), we obtain

$$\begin{aligned}
\mathfrak{K}^{\infty,\mathbb{S}}_{\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L}) &= \sup_{\substack{\|\mathrm{vec}(\Psi^{\ddagger}\odot\Delta\mathbf{H},\boldsymbol{\chi}^{\ddagger}\odot\Delta\mathbf{d})\|_{\infty}\neq 0 \\ \Delta A\in\mathbb{S}_1,\,\Delta D\in\mathbb{S}_2,\,\Delta E\in\mathbb{S}_3}} \frac{\left\| \mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}} \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Delta\mathbf{H}) \\ \Delta\mathbf{d} \end{bmatrix} \right\|_{\infty}}{\|\mathrm{vec}\left(\Psi^{\ddagger} \odot \Delta\mathbf{H}, \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d}\right)\|_{\infty}} \\
&= \sup_{\substack{\left\| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_{\infty} \neq 0 \\ \Delta A\in\mathbb{S}_1,\,\Delta D\in\mathbb{S}_2,\,\Delta E\in\mathbb{S}_3}} \frac{\left\| \mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}} \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}}\mathfrak{D}_{\mathrm{vec}_{\mathbb{S}}(\Psi)} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_{\infty}}{\left\| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(\Psi^{\ddagger} \odot \Delta\mathbf{H}) \\ \boldsymbol{\chi}^{\ddagger} \odot \Delta\mathbf{d} \end{bmatrix} \right\|_{\infty}} \\
&= \left\| \mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}} \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \begin{bmatrix} \mathfrak{D}_{\mathrm{vec}_{\mathbb{S}}(\Psi)} & \mathbf{0} \\ \mathbf{0} & \mathfrak{D}_{\boldsymbol{\chi}} \end{bmatrix} \right\|_{\infty} \\
&= \left\| |\mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}}| \left| \mathbf{L}\mathfrak{B}^{-1} \begin{bmatrix} \mathcal{G} & -I_l \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{\mathbb{S}} & \mathbf{0} \\ \mathbf{0} & I_l \end{bmatrix} \right| \begin{bmatrix} \mathrm{vec}_{\mathbb{S}}(|\Psi|) \\ |\boldsymbol{\chi}| \end{bmatrix} \right\|_{\infty}.
\end{aligned}$$

The rest of the proof follows by considering $\Psi = \mathbf{H}$, $\boldsymbol{\chi} = \mathbf{d}$, and $\xi_{\mathbf{L}} = \|\mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T\|_{\infty}\mathbf{1}_l$ (or $\xi_{\mathbf{L}} = \mathbf{L}[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$). ■

5.2.5. Deduction of Partial CNs for the EILS Problem

The EILS problem is an extension of the famous linear least squares problem, having linear constraints on unknown parameters. We consider the EILS problems given (1.1.5).

The solution of the EILS problem also satisfies the following the augmented system [137]:

$$\widehat{\mathfrak{B}}\begin{bmatrix}\lambda\\ \boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} := \begin{bmatrix}\mathbf{0} & \mathbf{0} & C\\ \mathbf{0} & \mathbb{J} & M\\ C^T & M^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\lambda\\ \boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} = \begin{bmatrix}d\\ b\\ \mathbf{0}\end{bmatrix}, \tag{5.2.39}$$

where $\boldsymbol{x} = \mathbb{J}\boldsymbol{r}$, $\boldsymbol{r} = b - M\boldsymbol{y}$ and $\lambda = (CC^T)^{-1}CM^T\mathbb{J}\boldsymbol{r}$ is the vector of Lagrange multipliers [31]. Note that the system in (5.2.39) can be equivalently transformed into

$$\begin{bmatrix}\mathbb{J} & M & \mathbf{0}\\ M^T & \mathbf{0} & C^T\\ \mathbf{0} & C & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \lambda\end{bmatrix} = \begin{bmatrix}b\\ \mathbf{0}\\ d\end{bmatrix} =: \mathbf{d}. \tag{5.2.40}$$

Observe that, the above system is in the form of DSPP (5.2.1) with $A = \mathbb{J}$, $B = M^T$, $\mathbf{d} = [b^T, \mathbf{0}, d^T]^T$ and $\boldsymbol{z} = \lambda$. Therefore, the task of assessing the conditioning of the EILS problem (1.1.5) can be achieved by determining the CNs for the solution $\boldsymbol{y}$ of DSPP (5.2.40).

Generally the signature matrix $\mathbb{J}$ has no perturbation and as $D = \mathbf{0}$, $E = \mathbf{0}$ and $g = \mathbf{0}$, we consider $\Delta A = \mathbf{0}$, $\Delta D = \mathbf{0}$, $\Delta E = 0$ and $\Delta g = \mathbf{0}$ in (5.2.7). Then, the perturbation expression in (5.2.8) reduces to

$$\begin{bmatrix}\Delta\boldsymbol{x}\\ \Delta\boldsymbol{z}\\ \Delta\boldsymbol{y}\end{bmatrix} = -\mathfrak{B}^{-1}\begin{bmatrix}\widehat{\mathcal{G}} & -I_{n+p}\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\Delta B^T)\\ \mathrm{vec}(\Delta C)\\ \Delta f\\ \Delta h\end{bmatrix} + \mathcal{O}(\boldsymbol{\epsilon}^2), \tag{5.2.41}$$

where

$$\widehat{\mathcal{G}} = \begin{bmatrix}\boldsymbol{y}^T \otimes I_n & \mathbf{0}\\ I_m \otimes \boldsymbol{x}^T & I_m \otimes \boldsymbol{z}^T\\ \mathbf{0} & \boldsymbol{y}^T \otimes I_p\end{bmatrix} \in \mathbb{R}^{l\times\hat{\boldsymbol{s}}}, \tag{5.2.42}$$

$\hat{\boldsymbol{s}} = m(n+p)$.

Let $\widehat{\mathbf{H}} = (B^T, C)$, $\Delta\widehat{\mathbf{H}} = (\Delta B^T, \Delta C)$, $\widehat{\mathbf{d}} = [f^T, h^T]^T$ and $\Delta\widehat{\mathbf{d}} = [\Delta\widehat{f}^T, \Delta\widehat{h}^T]^T$. We define the following mapping:

$$\widetilde{\boldsymbol{\varphi}} : \mathbb{R}^{m\times n} \times \mathbb{R}^{p\times m} \times \mathbb{R}^{n+p} \to \mathbb{R}^k$$

$$\widetilde{\boldsymbol{\varphi}}(\widehat{\mathbf{H}}, \widehat{\mathbf{d}}) = \mathbf{L}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{bmatrix} = \mathbf{L}\mathfrak{B}^{-1}\widehat{\mathbf{d}}, \tag{5.2.43}$$

where $\mathbf{L} \in \mathbb{R}^{k\times \boldsymbol{l}}$. Using a similar method to the Theorem 5.2.4, we have the following result.

Theorem 5.2.13. *Assume that $[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ is the unique solution of the DSPP (5.2.1) with $D = \mathbf{0}$, $E = \mathbf{0}$, $g = \mathbf{0}$ and $\mathbf{L} \in \mathbb{R}^{k\times \boldsymbol{l}}$. Then the partial unified CN of $[\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ w.r.t. $\mathbf{L}$ is given by*

$$\begin{aligned}\mathfrak{K}_{\widetilde{\boldsymbol{\varphi}}}(\widehat{\mathbf{H}}, \widehat{\mathbf{d}}; \mathbf{L}) &:= \lim_{\boldsymbol{\epsilon}\to 0} \sup_{0<\left\|\mathrm{vec}\left(\widehat{\Psi}\odot\Delta\widehat{\mathbf{H}}, \widehat{\boldsymbol{\chi}}\odot\Delta\widehat{\mathbf{d}}\right)\right\|_{\tau}\leq\boldsymbol{\epsilon}} \frac{\left\|\xi^{\ddagger}_{\mathbf{L}} \odot \left(\widetilde{\boldsymbol{\varphi}}(\widehat{\mathbf{H}}+\Delta\widehat{\mathbf{H}}, \widehat{\mathbf{d}}+\Delta\widehat{\mathbf{d}}) - \widetilde{\boldsymbol{\varphi}}(\widehat{\mathbf{H}}, \widehat{\mathbf{d}})\right)\right\|_{\gamma}}{\left\|\mathrm{vec}\left(\widehat{\Psi}^{\ddagger}\odot\Delta\widehat{\mathbf{H}}, \widehat{\boldsymbol{\chi}}^{\ddagger}\odot\Delta\widehat{\mathbf{d}}\right)\right\|_{\tau}},\\ &= \left\|\mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}}\mathbf{L}\mathfrak{B}^{-1}\begin{bmatrix}\widehat{\mathcal{G}} & -I_{n+p}\end{bmatrix}\begin{bmatrix}\mathfrak{D}_{\mathrm{vec}(\widehat{\Psi})} & \mathbf{0}\\ \mathbf{0} & \mathfrak{D}_{\widehat{\boldsymbol{\chi}}}\end{bmatrix}\right\|_{\tau,\gamma},\end{aligned}$$

where $\widehat{\Psi} = (\Psi_{B^T}, \Psi_C)$, $\Psi_{B^T} \in \mathbb{R}^{n\times m}, \Psi_C \in \mathbb{R}^{p\times m}$ and $\widehat{\boldsymbol{\chi}} \in \mathbb{R}^{n+p}$.

Remark 5.2.14. *Taking $\mathbf{L} = [\mathbf{0}_{k\times n} \quad \mathbf{L}_1 \quad \mathbf{0}_{k\times p}]$, $\mathbf{L}_1 \in \mathbb{R}^{k\times m}$, $A = \mathbb{J}$ in Theorem 5.2.13, and since $\mathfrak{B} = \Sigma\widehat{\mathfrak{B}}\Sigma^{-1}$, where*

$$\Sigma = \begin{bmatrix}\mathbf{0} & I_n & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & I_m\\ I_p & \mathbf{0} & \mathbf{0}\end{bmatrix}, \tag{5.2.44}$$

using the formula for the inverse of $\widehat{\mathfrak{B}}$ given in [93], we obtain

$$\mathfrak{K}_{\widetilde{\boldsymbol{\varphi}}}(\widehat{\mathbf{H}}, \mathbf{b}; \mathbf{L}) = \left\|\mathfrak{D}_{\xi^{\ddagger}_{\mathbf{L}}}\mathbf{L}_1\begin{bmatrix}\Xi & \Lambda & -(\mathcal{Q}\mathcal{M}\mathcal{Q})^{\dagger}MJ & \mathcal{B}_M\end{bmatrix}\begin{bmatrix}\mathfrak{D}_{\mathrm{vec}(\widehat{\Psi})} & \mathbf{0}\\ \mathbf{0} & \mathfrak{D}_{\widehat{\boldsymbol{\chi}}}\end{bmatrix}\right\|_{\tau,\gamma}, \tag{5.2.45}$$

where $\Xi = \boldsymbol{y}^T\otimes(\mathcal{Q}\mathcal{M}\mathcal{Q})^{\dagger}M\mathbb{J} - (\mathcal{Q}\mathcal{M}\mathcal{Q})^{\dagger}\otimes\boldsymbol{x}^T$, $\Lambda = \boldsymbol{y}^T\otimes\mathcal{B}_M - (\mathcal{Q}\mathcal{M}\mathcal{Q})^{\dagger}\otimes\boldsymbol{z}^T$, $\mathcal{M} = M\mathbb{J}M^T$, $\mathcal{Q} = I_m - C^{\dagger}C$ and $\mathcal{B}_M = (I_m - \mathcal{Q}\mathcal{M}\mathcal{Q})^{\dagger}$. Note that the partial CN expression is the same as derived in [137].

5.2.6. Numerical Experiments

In this part, we present some numerical examples to verify the reliability of the derived partial NCN, MCN, and CCN and their upper bounds for the DSPP. Additionally, we demonstrate their effectiveness in providing tight upper bounds for the relative forward error for the solution of the DSPP. We construct the entrywise perturbation as follows:

$$\begin{aligned}&\Delta A = 10^{-s}\cdot\mathtt{randn}(n,n)\odot A, \quad \Delta B = 10^{-s}\cdot\mathtt{randn}(m,n)\odot B,\\ &\Delta C = 10^{-s}\cdot\mathtt{randn}(p,m)\odot C, \quad \Delta D = 10^{-s}\cdot\mathtt{randn}(m,m)\odot D,\\ &\Delta E = 10^{-s}\cdot\mathtt{randn}(p,p)\odot E, \quad \text{and} \quad \Delta\mathbf{d} = 10^{-s}\cdot\mathtt{randn}(l,1)\odot\mathbf{d}.\end{aligned}$$

Let $\boldsymbol{w} = [\boldsymbol{x}^T, \boldsymbol{y}^T, \boldsymbol{z}^T]^T$ and $\widetilde{\boldsymbol{w}} = [\widetilde{\boldsymbol{x}}^T, \widetilde{\boldsymbol{y}}^T, \widetilde{\boldsymbol{z}}]^T$ be the unique solutions of the original DSPP and the perturbed DSPP, respectively. To estimate an upper bound for the forward error in the solution, their normwise, mixed, and componentwise relative forward errors are defined as follows:

$$\boldsymbol{r}_k = \frac{\|\mathbf{L}\widetilde{\boldsymbol{w}} - \mathbf{L}\boldsymbol{w}\|_2}{\|\mathbf{L}\boldsymbol{w}\|_2}, \quad \boldsymbol{r}_m = \frac{\|\mathbf{L}\widetilde{\boldsymbol{w}} - \mathbf{L}\boldsymbol{w}\|_\infty}{\|\mathbf{L}\boldsymbol{w}\|_\infty} \quad \text{and} \quad \boldsymbol{r}_c = \left\| \frac{\mathbf{L}\widetilde{\boldsymbol{w}} - \mathbf{L}\boldsymbol{w}}{\mathbf{L}\boldsymbol{w}} \right\|_\infty,$$

respectively. Define the following quantities:

$$\boldsymbol{\epsilon}_1 = \frac{\left\| \begin{bmatrix} \Delta\mathfrak{B} & \Delta\mathbf{d} \end{bmatrix} \right\|_F}{\left\| \begin{bmatrix} \mathfrak{B} & \mathbf{d} \end{bmatrix} \right\|_F} \quad \text{and} \quad \boldsymbol{\epsilon}_2 = \min\left\{ \boldsymbol{\epsilon} : |\Delta\mathfrak{B}| \leq \boldsymbol{\epsilon}\,|\mathfrak{B}|\,,\ |\Delta\mathbf{d}| \leq \boldsymbol{\epsilon}\,|\mathbf{d}| \right\}. \tag{5.2.46}$$

Thus from (5.2.6), $\boldsymbol{\epsilon}_1 \mathfrak{K}^{(2)}_{\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L})$, $\boldsymbol{\epsilon}_2 \mathfrak{K}^{\infty}_{mix,\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L})$ and $\boldsymbol{\epsilon}_2 \mathfrak{K}^{\infty}_{com,\boldsymbol{\varphi}}(\mathbf{H}, \mathbf{d}; \mathbf{L})$ can be employed to estimate the relative forward errors $\boldsymbol{r}_k$, $\boldsymbol{r}_m$ and $\boldsymbol{r}_c$, respectively. We select the matrix $\mathbf{L}$ as

$$\mathbf{L}_0 = I_l,\ \mathbf{L}_n = \begin{bmatrix} I_n & \mathbf{0}_{n\times(m+p)} \end{bmatrix},\ \mathbf{L}_m = \begin{bmatrix} \mathbf{0}_{m\times n} & I_m & \mathbf{0}_{m\times p} \end{bmatrix} \text{ and } \mathbf{L}_p = \begin{bmatrix} \mathbf{0}_{p\times(n+m)} & I_p \end{bmatrix}$$

to obtain the CNs for $\boldsymbol{w}$, $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, respectively.

Example 5.2.1. We consider the DSPP (5.2.1) taken from [75] with

$$A = \begin{bmatrix} I_q \otimes J + J \otimes I_q & \mathbf{0} \\ \mathbf{0} & I_q \otimes J + J \otimes I_q \end{bmatrix} \in \mathbb{R}^{2q^2\times 2q^2},$$
$$B = \begin{bmatrix} I_q \otimes Z & Z \otimes I_q \end{bmatrix} \in \mathbb{R}^{q^2\times 2q^2}, \text{ and } C = Y \otimes Z \in \mathbb{R}^{q^2\times q^2},$$

where $J = \frac{1}{(q+1)^2}\,\mathtt{tridiag}(-1, 2, -1) \in \mathbb{R}^{q\times q}$, $Z = \frac{1}{q+1}\,\mathtt{tridiag}(0, 1, -1) \in \mathbb{R}^{q\times q}$ and $Y = \mathrm{diag}(1, q+1, \ldots, q^2 - q + 1) \in \mathbb{R}^{q\times q}$. Further, we take $D = I_m$ and $E = I_p$. Here, the dimension of the coefficient matrix $\mathfrak{B}$ of the DSPP is $l = 4q^2$. We take $\mathbf{d} = \mathtt{randn}(l, 1) \in \mathbb{R}^l$. Further, we consider $\Psi = \|\mathfrak{B}\|_F$ and $\boldsymbol{\chi} = \|\mathbf{d}\|_2$. We use Theorems 5.2.5, 5.2.8 to compute partial CNs and Corollaries 5.2.1 and 5.2.2 for their upper bounds. The numerical results for different choices for $\mathbf{L}$ and $q = 4 : 2 : 16$ are presented in Tables 5.2.1-5.2.4.

The results in Tables 5.2.1-5.2.4 show that the upper bounds of the NCN, MCN and CCN provide very sharp estimates of the exact NCN, MCN and CCN, respectively. Additionally, it is observed that both the MCN and CCN, along with their upper bounds, are at most two orders of magnitude larger than the actual relative forward errors, offering more accurate estimates compared to the NCN and its upper bounds. These numerical

Table 5.2.1: Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_0$ for Example 5.2.1.

q	$\boldsymbol{r}_k$	$\epsilon_1\mathfrak{R}_{\boldsymbol{\varphi}}^{(2)}(\mathbf{H},\mathbf{b};\mathbf{L}_0)$	$\epsilon_1\mathfrak{R}_{\boldsymbol{\varphi}}^{(2),u}(\mathbf{H},\mathbf{b};\mathbf{L}_0)$	$\boldsymbol{r}_m$	$\epsilon_2\mathfrak{R}_{mix,\boldsymbol{\varphi}}^{\infty}(\mathbf{H},\mathbf{b};\mathbf{L}_0)$	$\epsilon_2\mathfrak{R}_{mix,\boldsymbol{\varphi}}^{\infty,u}(\mathbf{H},\mathbf{b};\mathbf{L}_0)$	$\boldsymbol{r}_c$	$\epsilon_2\mathfrak{R}_{com,\boldsymbol{\varphi}}^{\infty}(\mathbf{H},\mathbf{b};\mathbf{L}_0)$	$\epsilon_2\mathfrak{R}_{com,\boldsymbol{\varphi}}^{\infty,u}(\mathbf{H},\mathbf{b};\mathbf{L}_0)$
4	$1.1882e-08$	$5.1234e-06$	$6.3888e-06$	$1.9907e-08$	$3.5552e-07$	$3.8637e-07$	$2.1532e-07$	$6.9184e-06$	$8.6300e-06$
6	$2.0426e-08$	$8.2886e-06$	$1.6984e-05$	$2.3603e-08$	$3.9780e-07$	$4.0885e-07$	$3.7248e-07$	$1.2135e-05$	$1.2490e-05$
8	$3.1951e-08$	$1.9154e-05$	$3.6586e-05$	$4.1045e-08$	$4.9025e-07$	$4.9808e-07$	$8.3715e-06$	$1.3746e-04$	$1.3964e-04$
10	$2.9187e-08$	$5.5170e-05$	$1.1519e-04$	$4.2771e-08$	$1.2319e-06$	$1.2706e-06$	$1.8721e-07$	$1.9213e-05$	$2.1515e-05$
12	$2.3289e-08$	$9.4616e-05$	$1.7034e-04$	$2.9289e-08$	$1.0722e-06$	$1.0798e-06$	$4.5669e-07$	$3.5850e-05$	$3.8136e-05$
14	$3.9431e-08$	$1.6571e-04$	$3.3889e-04$	$3.9739e-08$	$1.2469e-06$	$1.2747e-06$	$6.0883e-07$	$3.1225e-05$	$3.5558e-05$
16	$3.9038e-08$	$2.7496e-04$	$5.6909e-04$	$3.9932e-08$	$1.8001e-06$	$1.8356e-06$	$4.2789e-06$	$1.7386e-04$	$1.8883e-04$

Table 5.2.2: Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_n$ for Example 5.2.1.

q	$\boldsymbol{r}_k$	$\epsilon_1\mathfrak{R}_{\boldsymbol{\varphi}}^{(2)}(\mathbf{H},\mathbf{b};\mathbf{L}_n)$	$\epsilon_1\mathfrak{R}_{\boldsymbol{\varphi}}^{(2),u}(\mathbf{H},\mathbf{b};\mathbf{L}_n)$	$\boldsymbol{r}_m$	$\epsilon_2\mathfrak{R}_{mix,\boldsymbol{\varphi}}^{\infty}(\mathbf{H},\mathbf{b};\mathbf{L}_n)$	$\epsilon_2\mathfrak{R}_{mix,\boldsymbol{\varphi}}^{\infty,u}(\mathbf{H},\mathbf{b};\mathbf{L}_n)$	$\boldsymbol{r}_c$	$\epsilon_2\mathfrak{R}_{com,\boldsymbol{\varphi}}^{\infty}(\mathbf{H},\mathbf{b};\mathbf{L}_n)$	$\epsilon_2\mathfrak{R}_{com,\boldsymbol{\varphi}}^{\infty,u}(\mathbf{H},\mathbf{b};\mathbf{L}_n)$
4	$2.3210e-08$	$1.7899e-06$	$2.2325e-06$	$4.4418e-08$	$3.9328e-07$	$3.9952e-07$	$6.1112e-08$	$7.7010e-07$	$7.9196e-07$
6	$2.2133e-08$	$2.6980e-06$	$7.6732e-06$	$4.2425e-08$	$8.8875e-07$	$9.5134e-07$	$5.3918e-08$	$5.8823e-06$	$7.3855e-06$
8	$3.8470e-08$	$4.8135e-06$	$1.5701e-05$	$8.7676e-08$	$1.4710e-06$	$1.5322e-06$	$5.7409e-07$	$5.8895e-05$	$7.0437e-05$
10	$2.9894e-08$	$9.6923e-06$	$3.1204e-05$	$5.9374e-08$	$2.0857e-06$	$2.1065e-06$	$2.7350e-07$	$9.5571e-05$	$1.0598e-04$
12	$3.3113e-08$	$3.6464e-05$	$4.9737e-05$	$5.6822e-08$	$3.0914e-06$	$3.0928e-06$	$4.5080e-07$	$9.3747e-06$	$9.4743e-06$
14	$8.0906e-08$	$2.4824e-05$	$7.6132e-05$	$1.5910e-07$	$3.3988e-06$	$3.4220e-06$	$8.4585e-07$	$2.5939e-05$	$2.9725e-05$
16	$4.4663e-08$	$3.5262e-05$	$8.8133e-05$	$8.0314e-08$	$4.7189e-06$	$4.7301e-06$	$4.5566e-07$	$3.7542e-05$	$3.9779e-05$

Table 5.2.3: Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_m$ for Example 5.2.1.

q	$\boldsymbol{r}_k$	$\epsilon_1\mathfrak{R}_{\boldsymbol{\varphi}}^{(2)}(\mathbf{H},\mathbf{b};\mathbf{L}_m)$	$\epsilon_1\mathfrak{R}_{\boldsymbol{\varphi}}^{(2),u}(\mathbf{H},\mathbf{b};\mathbf{L}_m)$	$\boldsymbol{r}_m$	$\epsilon_2\mathfrak{R}_{mix,\boldsymbol{\varphi}}^{\infty}(\mathbf{H},\mathbf{b};\mathbf{L}_m)$	$\epsilon_2\mathfrak{R}_{mix,\boldsymbol{\varphi}}^{\infty,u}(\mathbf{H},\mathbf{b};\mathbf{L}_m)$	$\boldsymbol{r}_c$	$\epsilon_2\mathfrak{R}_{com,\boldsymbol{\varphi}}^{\infty}(\mathbf{H},\mathbf{b};\mathbf{L}_m)$	$\epsilon_2\mathfrak{R}_{com,\boldsymbol{\varphi}}^{\infty,u}(\mathbf{H},\mathbf{b};\mathbf{L}_m)$
4	$1.4640e-08$	$4.1828e-06$	$8.1301e-06$	$1.7030e-08$	$3.2527e-07$	$3.6699e-07$	$6.3575e-08$	$1.6374e-06$	$1.7734e-06$
6	$3.8057e-08$	$3.4032e-05$	$5.4748e-05$	$3.9292e-08$	$8.5355e-07$	$1.0120e-06$	$6.8320e-08$	$2.2698e-06$	$2.3143e-06$
8	$6.2425e-08$	$1.1653e-04$	$1.7568e-04$	$5.3397e-08$	$1.9931e-06$	$2.1774e-06$	$3.0610e-07$	$1.3012e-05$	$1.3193e-05$
10	$4.7836e-08$	$1.5526e-04$	$3.3655e-04$	$5.1330e-08$	$1.8322e-06$	$2.2332e-06$	$5.0917e-07$	$2.8394e-05$	$2.8957e-05$
12	$3.7645e-08$	$9.3405e-05$	$4.7395e-04$	$3.6702e-08$	$1.1129e-06$	$1.4824e-06$	$7.7646e-08$	$7.9413e-06$	$8.8422e-06$
14	$7.7970e-08$	$1.6058e-04$	$9.2172e-04$	$5.7666e-08$	$1.7296e-06$	$2.1704e-06$	$2.9851e-06$	$7.6483e-05$	$7.9562e-05$
16	$4.7259e-08$	$4.3435e-04$	$1.5265e-03$	$6.6721e-08$	$2.2484e-06$	$2.8853e-06$	$2.9895e-07$	$5.8651e-05$	$6.0528e-05$

results highlight the effectiveness of the proposed CNs and their corresponding upper bounds.

Example 5.2.2. We consider the DSPP (5.2.1) with the block matrices given by

$$A = \operatorname{diag}((2ZZ^T + \Sigma_1),\, \Sigma_2,\, \Sigma_3) \in \mathbb{R}^{n\times n},\ B = \begin{bmatrix} N & -I_{2\tilde{q}} & I_{2\tilde{q}} \end{bmatrix} \in \mathbb{R}^{m\times n},$$

$$D = \texttt{toeplitz}(\boldsymbol{d}) \in \mathbb{R}^{m\times m},\ C = M,\ \text{and}\ E = \texttt{toeplitz}(\boldsymbol{e}) \in \mathbb{R}^{p\times p},$$

Table 5.2.4: Comparison of the NCN, MCN, and CCN, and their upper bounds, with the corresponding relative errors for $\mathbf{L} = \mathbf{L}_p$ for Example 5.2.1.

q	$\boldsymbol{r}_k$	$\boldsymbol{\epsilon}_1\mathfrak{K}^{(2)}_{\boldsymbol{\varphi}}(\mathbf{H},\mathbf{b};\mathbf{L}_p)$	$\boldsymbol{\epsilon}_1\mathfrak{K}^{(2),u}_{\boldsymbol{\varphi}}(\mathbf{H},\mathbf{b};\mathbf{L}_p)$	$\boldsymbol{r}_m$	$\boldsymbol{\epsilon}_2\mathfrak{K}^{\infty}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{b};\mathbf{L}_p)$	$\boldsymbol{\epsilon}_2\mathfrak{K}^{\infty,u}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{b};\mathbf{L}_p)$	$\boldsymbol{r}_c$	$\boldsymbol{\epsilon}_2\mathfrak{K}^{\infty}_{com,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{b};\mathbf{L}_p)$	$\boldsymbol{\epsilon}_2\mathfrak{K}^{\infty,u}_{com,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{b};\mathbf{L}_p)$
4	$7.9587e-09$	$3.2626e-06$	$5.0845e-06$	$9.2868e-09$	$2.7450e-07$	$2.8678e-07$	$6.4428e-08$	$5.9782e-07$	$6.2833e-07$
6	$1.3913e-08$	$9.6919e-06$	$1.1609e-05$	$1.8364e-08$	$4.4130e-07$	$4.4401e-07$	$4.1485e-08$	$1.1621e-06$	$1.1751e-06$
8	$2.1543e-08$	$2.5173e-05$	$4.6780e-05$	$2.7719e-08$	$7.9552e-07$	$8.4231e-07$	$4.9306e-08$	$4.2242e-06$	$4.4848e-06$
10	$3.2132e-08$	$5.7696e-05$	$1.1250e-04$	$4.3437e-08$	$1.0618e-06$	$1.1161e-06$	$8.3724e-08$	$4.1932e-06$	$4.3258e-06$
12	$1.6735e-08$	$8.7652e-05$	$1.4248e-04$	$3.1936e-08$	$1.0011e-06$	$1.0138e-06$	$6.2623e-08$	$2.5824e-06$	$2.5860e-06$
14	$2.0065e-08$	$1.5497e-04$	$1.9181e-04$	$2.2200e-08$	$1.0421e-06$	$1.0457e-06$	$7.7378e-08$	$4.1703e-06$	$4.1741e-06$
16	$1.0902e-08$	$2.4487e-04$	$3.5342e-04$	$1.5817e-08$	$1.6369e-06$	$1.6456e-06$	$1.0229e-07$	$3.4527e-06$	$3.4614e-06$

where $Z = [z_{ij}] \in \mathbb{R}^{\hat{q}\times\hat{q}}$ with $z_{ij} = e^{-2((i/3)^2+(j/3)^2)}$, $\Sigma_1 = I_{\hat{q}}$, and $\Sigma_k = \operatorname{diag}(d_j^{(k)}) \in \mathbb{R}^{2\tilde{q}\times 2\tilde{q}}$, $k = 2, 3$, are diagonal matrices with

$$d_j^{(2)} = \begin{cases} 1, & \text{for } 1 \le j \le \tilde{q}, \\ 10^{-5}(j-\tilde{q})^2, & \text{for } \tilde{q}+1 \le j \le 2\tilde{q}, \end{cases}$$

$d_j^{(3)} = 10^{-5}(j+\tilde{q})^2$ for $1 \le j \le 2\tilde{q}$, where $\tilde{q} = q^2$ and $\hat{q} = q(q+1)$. Further, $N = \begin{bmatrix} \widehat{N} \otimes I_q \\ I_q \otimes \widehat{N} \end{bmatrix} \in \mathbb{R}^{2\tilde{q}\times\hat{q}}$, $\widehat{N} = \texttt{tridiag}(0, 2, -1) \in \mathbb{R}^{q\times(q+1)}$, $M = \begin{bmatrix} \widehat{M} \otimes I_q & I_q \otimes \widehat{M} \end{bmatrix} \in \mathbb{R}^{\hat{q}\times 2\tilde{q}}$,

$$\widehat{M} = \begin{bmatrix} q+1 & -\frac{q-1}{q} & \frac{q-2}{q} & \dots & \frac{(-1)^{q-1}}{q} \\ -\frac{q-1}{q} & 2q+1 & \ddots & \ddots & \vdots \\ \frac{q-2}{q} & \ddots & \ddots & \ddots & \frac{q-2}{q} \\ \vdots & \ddots & \ddots & \ddots & -\frac{q-1}{q} \\ \frac{(-1)^{q-1}}{q} & \vdots & \frac{q-2}{q} & -\frac{q-1}{q} & q^2+1 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{(q+1)\times q},$$

$\boldsymbol{d} = \texttt{randn}(m, 1) \in \mathbb{R}^m$, and $\boldsymbol{e} = \texttt{randn}(p, 1) \in \mathbb{R}^p$. Moreover, we get $l = 8q^2 + 2q$. The vector $\mathbf{d} \in \mathbb{R}^l$ is chosen as in Example 5.2.1.

The structured partial NCN is calculated using Theorem 5.2.11, while the structured partial MCN and CCN are computed from Theorem 5.2.12. Numerical results for various choices of $\mathbf{L}$ and $q = 2, 3, 4$ are summarized in Table 5.2.5. We observed for all choices of $\mathbf{L}$, $\mathfrak{K}^{\infty,\mathbb{S}}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L})$ and $\mathfrak{K}^{\infty,\mathbb{S}}_{comp,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L})$ are almost one order smaller than the $\mathfrak{K}^{\infty}_{mix,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L})$ and $\mathfrak{K}^{\infty}_{comp,\boldsymbol{\varphi}}(\mathbf{H},\mathbf{d};\mathbf{L})$, respectively.

Table 5.2.5: Comparison of the partial NCN, MCN, and CCN with their structured counterparts for Example 5.2.2.

$\mathbf{L}$	q	$\mathfrak{K}_{\varphi}^{(2)}(\mathbf{H},\mathbf{d};\mathbf{L})$	$\mathfrak{K}_{\varphi}^{(2),\mathbb{S}}(\mathbf{H},\mathbf{d};\mathbf{L})$	$\mathfrak{K}_{mix,\varphi}^{\infty}(\mathbf{H},\mathbf{d};\mathbf{L})$	$\mathfrak{K}_{mix,\varphi}^{\infty,\mathbb{S}}(\mathbf{H},\mathbf{d};\mathbf{L})$	$\mathfrak{K}_{com,\varphi}^{\infty}(\mathbf{H},\mathbf{d};\mathbf{L})$	$\mathfrak{K}_{com,\varphi}^{\infty,\mathbb{S}}(\mathbf{H},\mathbf{d};\mathbf{L})$
$\mathbf{L}_0$	2	$9.0961e+03$	$8.9842e+03$	$1.1174e+02$	$9.2003e+01$	$1.0211e+03$	$7.8013e+02$
	3	$1.0996e+04$	$8.6963e+03$	$1.1735e+02$	$9.1218e+01$	$1.2372e+04$	$9.5905e+03$
	4	$2.6066e+04$	$1.9983e+04$	$3.5456e+02$	$2.6052e+02$	$2.5284e+05$	$6.7124e+04$
$\mathbf{L}_n$	2	$9.1011e+03$	$8.9891e+03$	$1.1174e+02$	$9.2003e+01$	$2.9170e+02$	$2.1175e+02$
	3	$1.0972e+04$	$8.6781e+03$	$1.1735e+02$	$9.1218e+01$	$1.2372e+04$	$9.5905e+03$
	4	$2.6371e+04$	$2.0138e+04$	$3.5456e+02$	$2.6052e+02$	$2.7573e+04$	$1.8454e+04$
$\mathbf{L}_m$	2	$8.4606e+03$	$8.3300e+03$	$1.5279e+02$	$1.1486e+02$	$2.5276e+02$	$1.5433e+02$
	3	$1.8511e+04$	$1.4614e+04$	$2.2517e+02$	$1.6131e+02$	$9.9621e+02$	$5.3614e+02$
	4	$3.2837e+04$	$2.7473e+04$	$8.9167e+02$	5.8061e+02	$2.5284e+05$	$6.7124e+04$
$\mathbf{L}_p$	2	$1.1952e+04$	$1.1786e+04$	$2.1568e+02$	$1.6042e+02$	$1.0211e+03$	$7.8013e+02$
	3	$9.4332e+03$	$7.4505e+03$	$1.2141e+02$	$8.3080e+01$	$1.1802e+04$	$8.2273e+03$
	4	$1.8803e+04$	$1.5754e+04$	$7.0206e+02$	$4.8744e+02$	$7.4731e+03$	$4.8946e+03$

5.2.7. Summary

This section introduced a unified framework for investigating the partial CN for the solution of the DSPP. We derived compact formulas for the partial unified CN, and by considering specific norms, we obtained the partial NCN, MCN, and CCN for the solution of the DSSP. Additionally, sharp upper bounds for the partial CNs that are free from Kronecker products are provided. Moreover, we compute structured partial NCN, MCN, and CCN by introducing perturbations that maintain the structure of the block matrices of the coefficient matrix. Using our theoretical findings and by leveraging the relationship between the EILS problems and the DSPP, we recovered previously established results for the EILS problems. Experimental results demonstrated that the derived upper bounds for the partial CNs provide tight estimates of the actual partial CNs. Furthermore, the proposed partial CNs and their upper bounds provide sharp error estimation for the solution, highlighting their effectiveness and reliability.

CHAPTER 6

Condition Numbers for Moore-Penrose Inverse and Least Square Problem*

This chapter addresses and analyzes structured MCN and CCN for the Moore-Penrose (M-P) inverse and the minimum norm least squares (MNLS) solution of least squares (LS) problem involving rank-structured matrices, which include the Cauchy-Vandermonde (CV) matrices and $\{1,1\}$-Quasiseparable (QS) matrices. A general framework has been developed to compute the upper bounds for MCN and CCN of rank deficient parameterized matrices. This framework leads to faster computation of upper bounds for structured MCN and CCN for CV and $\{1,1\}$-QS matrices. Furthermore, comparisons of obtained upper bounds are investigated theoretically and experimentally. In addition, the structured effective CNs for the M-P inverse and the MNLS solution of $\{1,1\}$-QS matrices are presented. Numerical tests reveal the reliability of the proposed upper bounds as well as demonstrate that the structured effective CNs can be substantially smaller compared to the unstructured CNs.

6.1. Background

The M-P inverse holds a pivotal position in matrix computation, offering a generalization of the standard inverse for rectangular or rank deficient matrices. The M-P inverse finds its practical significance in solving the linear LS problem. The M-P inverse and the LS problem have various applications in digital image restoration and reconstruction [48, 47], Gauss–Markov model [112], and so on. The literature on CNs for the M-P inverse [53] and LS problems [145] is quite rich. The normwise CN for the M-P inverse and the LS problem is investigated in [67, 99, 53], while MCN and CCN are considered in [51, 50]. For structured matrices, structured CNs for the M-P inverse and the LS problem have been investigated in [152, 50], which involves the preservation of the inherent matrix structure within the perturbation matrices.

* S. S. Ahmad and **P. Khatun**, " Condition numbers for the Moore-Penrose inverse and the least squares problem involving rank-structured matrices." *Linear and Multilinear Algebra,* 4:1-37, 2024.

In the past few years, many fast algorithms have been developed for various problems involving rank-structured matrices, such as computing eigenvalues and singular values [132, 155], solving linear systems [131] and LS problems [78], and computing the M-P inverse [42]. The QS [58], Cauchy [78], and CV [78] matrices are popular examples of rank-structured matrices that arise in many applications, such as in boundary value problem [68, 85], acoustic and electromagnetic scattering theory [49], interpolation problems [105], rational models of regression and E-optimal design [81], and so on.

One of the striking properties of the rank-structured matrices is that they can be parameterized by $\mathcal{O}(m+n)$ parameters rather than mn entries. Based on this property, many fast algorithms with lower computational costs have been developed [131, 130]. Plenty of works involving rank-structured matrices have been done in recent years to investigate the structured CNs for eigenvalue problems [56, 52], the solution of a linear system having a single as well as multiple right-hand sides [57, 103], the Sylvester matrix equation [54], and so on, by considering perturbations on the parameters. Based on the above discussions, it is more sensible to investigate structured CNs by addressing perturbations on the parameters rather than directly on the matrix entries and to identify which set of parameters will be more suited for the development of fast algorithms. Thus, the forgoing discussion motivates us to consider perturbations on parameters instead of directly on entries in this chapter.

In [145], authors have presented general parameterized QS representation and Givens-vector (GV) representation for the rectangular $m\times n$ $(m \geq n)$ $\{1,1\}$-QS matrices (a special case of QS matrices), which are natural extensions of the square matrix case discussed in [56, 57, 130]. Then, the authors studied the structured MCN for the LS problems when the coefficient matrix is a full column rank $m \times n$ $\{1,1\}$-QS matrices. However, the above investigations do not address the rank deficient case. Furthermore, the MCN and CCN for the M-P inverse of rank deficient rank-structured matrices still need to be explored in the literature. Nevertheless, when dealing with rank deficient matrices, a prominent challenge in analyzing the CNs arises from the fact that even slight changes to the matrix can yield enormous variations in the computed M-P inverse. In light of this, normwise CNs for rank deficient unstructured matrices have been considered in [140, 142], and for structured matrices in [152] under the assumptions: $\mathcal{R}(\Delta M) \subset \mathcal{R}(M)$ and $\mathcal{R}(\Delta M^T) \subset \mathcal{R}(M^T)$, on the perturbation matrix ΔM in M, where $\mathcal{R}(M)$ denotes the range of M. Whereas, in [141], upper bounds are investigated for CCN for unstructured matrices under the above assumptions.

This chapter's central aim is to study the structured MCN and CCN for the M-P inverse and the LS problem when dealing with rank deficient rank-structured matrices. This investigation adheres to the rank-preserving constraint, denoted as $\text{rank}(M+\Delta M) = \text{rank}(M)$, which encompasses a broader class of perturbation matrices than those constrained by $\mathcal{R}(\Delta M) \subset \mathcal{R}(M)$ and $\mathcal{R}(\Delta M^T) \subset \mathcal{R}(M^T)$. This perspective expands the horizons of our study and offers valuable insights into structured CNs for this class of matrices.

The following highlights the main contributions of this chapter:

- The MCN and CCN for two problems, the M-P inverse and the MNLS solution of the LS problem, involving rank deficient CV and $\{1,1\}$-QS matrices are considered under the broader rank condition, i.e., $\text{rank}(M + \Delta M) = \text{rank}(M)$.
- By considering matrix entries to be differentiable functions of a set of real parameters, we develop a general framework to compute the upper bounds of the MCN and CCN of the M-P inverse and LS problem for rank deficient parameterized matrices. In addition, exact expressions in the full column rank case of the MCN and CCN are also obtained.
- For the CV and $\{1,1\}$-QS matrices, compact upper bounds are obtained for structured MCN and CCN. Two important parameter representations for $\{1,1\}$-QS matrices are considered: the QS representation and GV representation.
- For $\{1,1\}$-QS matrices, structured effective CNs are proposed and shown that they can reliably estimate the actual conditioning of these matrices.
- Numerical experiments are reported to demonstrate that structured CNs are significantly smaller compared to unstructured CNs and align consistently with the theoretical results.

The remaining part of this chapter is structured as follows. Section 6.2 provides a few notations and preliminary results. In Section 6.3, for the M-P inverse and the MNLS solution, we develop expressions of upper bounds for MCN and CCN for a general class of parameterized matrices. These frameworks are utilized in Sections 6.4 and 6.5 to derive the bounds for structured MCN and CCN for CV and $\{1,1\}$-QS matrices. Further, Section 6.5 studies comparison results between different structured and unstructured CNs. In Section 6.6, numerical experiments are performed to illustrate our findings. Section 6.7 ends with conclusions and a line of future research.

6.2. Preliminaries

For $M \in \mathbb{R}^{m\times n}$, set $\mathbf{E}_M := I_m - MM^\dagger$ and $\mathbf{F}_M := I_n - M^\dagger M$. We denote $E_{ij}^{mn} = e_i^m (e_j^n)^T$ as the matrix with ij-element is 1 and zero elsewhere. For matrices $M, N \in \mathbb{R}^{m\times n}$, we define M/N as $(M/N)_{ij} = n_{ij}^\ddagger m_{ij}$, where for any $a \in \mathbb{R}$, $a^\ddagger = \frac{1}{a}$ when $a \neq 0$, otherwise $a^\ddagger = 1$. The notation $i = 1 : n$ indicates that i takes the values $1, 2, \ldots, n$. For any $a \in \mathbb{R}$, $\text{sign}(a) := \frac{a}{|a|}$ for $a \neq 0$ and $\text{sign}(a) := 0$ for $a = 0$, and $\text{sign}(M) := [\text{sign}(m_{ij})]$.

Next, we discuss some important properties of the M-P inverse, which will be crucial for our main finding results. The following lemma states that for a full column rank matrices M, its M-P inverse is a continuous function of its data entries.

Lemma 6.2.1. *[133] Let $M \in \mathbb{R}^{m\times n}$ with full column rank and $\{E_j\}$ be a collection of real $m \times n$ matrices satisfying $\lim_{j\to 0} E_j = \mathbf{0}$. Then, $(M + E_j)$ has full column rank when j is small enough and $\lim_{j\to 0}(M + E_j)^\dagger = M^\dagger$.*

However, M does not share the above property when it is singular or rank deficient. Small perturbation ΔM on M can produce the computed M-P inverse far from the actual one. To tackle this situation, perturbation theory for the M-P inverse has been studied in certain specific constraints. Next, we recall the definition of 'acute' perturbation [128].

Definition 6.2.1. *An acute perturbation $\widetilde{M} = M + \Delta M \in \mathbb{R}^{m\times n}$ of $M \in \mathbb{R}^{m\times n}$ is a perturbed matrix for which $\|MM^\dagger - \widetilde{M}\widetilde{M}^\dagger\|_2 < 1$ and $\|M^\dagger M - \widetilde{M}^\dagger\widetilde{M}\|_2 < 1$.*

Proposition 6.2.2 provides an if and only if condition for the continuity of $M^\dagger$ of any matrix $M \in \mathbb{R}^{m\times n}$.

Proposition 6.2.2. *Let $M \in \mathbb{R}^{m\times n}$. Consider the set*

$$\mathcal{S}^1(M) = \left\{\Delta M \in \mathbb{R}^{m\times n} : \|M^\dagger\|_2\|\Delta M\|_2 < 1\right\}.$$

Then, $\lim_{\Delta M\to \mathbf{0}} (M + \Delta M)^\dagger = M^\dagger$ if and only if $\text{rank}(M + \Delta M) = \text{rank}(M)$, where $\Delta M \in \mathcal{S}^1(M)$.

Proof. For $\Delta M \in \mathcal{S}^1(M)$, we have $\|M^\dagger\|_2\|\Delta M\|_2 < 1$. Then $M + \Delta M$ is an acute perturbation of M if and only if $\text{rank}(M + \Delta M) = \text{rank}(M)$ [91, Lemma 1]. Since on the set of acute perturbations of M, its M-P inverse $M^\dagger$ is a continuous function about M [128, Page 140]. Therefore, it follows that $M^\dagger$ is continuous on the set $S^1(M)$ if and only if $\text{rank}(M + \Delta M) = \text{rank}(M)$. Hence, the proof is completed. ■

Remark 6.2.3. *When M has full column rank (or row rank), from Lemma 6.2.1, the rank condition in Proposition 6.2.2 holds trivially.*

6.3. MCN and CCN for General Parameterized Matrices

In this part, initially, we define structured MCN and CCN for the M-P inverse and the unique MNLS solution for a general class of parameterized matrices. Suppose that each entry of $M \in \mathbb{R}^{m\times n}$ is a differentiable function of a set of real parameters $\Psi = [\psi_1, \psi_2, \ldots, \psi_p]^T \in \mathbb{R}^p$ and write the matrix as $M(\Psi)$. We employ this notation for the rest of the chapter. Due to the fact that a number of important classes of matrices can be parameterized by a collection of parameters, it is reasonable to consider perturbations on the parameters rather than directly on their entries. Let $\Delta\Psi \in \mathbb{R}^p$ be the perturbation on the parameters set $\Psi \in \mathbb{R}^p$, we consider the admissible perturbation in the matrix $M(\Psi)$ as $M(\Psi + \Delta\Psi) - M(\Psi) = \Delta M(\Psi)$. For maintaining the continuity property for $M^\dagger(\Psi)$, according to the Proposition 6.2.2, we restrict the perturbation on Ψ to the following set

$$\mathcal{S}(\Psi) := \Big\{\Delta\Psi \in \mathbb{R}^p : \operatorname{rank}(M(\Psi)) = \operatorname{rank}(M(\Psi + \Delta\Psi)) = r,\ \|M^\dagger(\Psi)\|_2\|\Delta M(\Psi)\|_2 < 1\Big\}.$$

Next, we provide an example to show that $\mathcal{S}(\Psi)$ is nonempty.

Example 6.3.1. Consider the parameter set

$$\Psi = [\{2,4\},\{1\},\{-3,1\},\{5,2,6\},\{1,3\},\{2\},\{3,1\}]^T \in \mathbb{R}^{13}$$

of a $\{1,1\}$-QS matrix given as in (6.5.1) and using the formula provided in Definition 6.5.1, we have

$$M(\Psi) = \begin{bmatrix} 5 & 3 & 2 \\ -6 & 2 & 3 \\ -12 & 4 & 6 \end{bmatrix}. \tag{6.3.1}$$

Taking $\Delta\Psi = [\{0,0\},\{0\},\{0,0\},\{\mu,0,0\},\{\mu,0\},\{0\},\{0,0\}]^T \in \mathbb{R}^{13}$, we get

$$M(\Psi + \Delta\Psi) = \begin{bmatrix} 5+\mu & 3+3\mu & 2+2\mu \\ -6 & 2 & 3 \\ -12 & 4 & 6 \end{bmatrix}.$$

Here, $\|M^\dagger(\Psi)\|_2 = 0.1812$ and $\|\Delta M^\dagger(\Psi)\|_2 = \sqrt{14}|\mu|$. Clearly, $\operatorname{rank}(M(\Psi)) = \operatorname{rank}(M(\Psi+\Delta\Psi))$ and $\|M^\dagger(\Psi)\|_2\|\Delta M(\Psi)\|_2 < 1$, whenever $|\mu| < 1.4747$. Thus, $\mathcal{S}(\Psi)$ contains all perturbations $\Delta\Psi$ such that $|\mu| < 1.4747$.

6.3.1. M-P Inverse of General Parameterized Matrices

We introduce structured MCN and CCN in Definition 6.3.1 for a general class of parameterized matrices for its M-P inverse. We provide general expressions for the upper bounds of these CNs in Theorem 6.3.2. Also, we present exact formulae for these CNs in Theorem 6.3.5 for full column rank matrices.

Definition 6.3.1. *Let $M(\Psi) \in \mathbb{R}^{m\times n}$, $\text{rank}(M(\Psi)) = r \leq \min\{m,n\}$. Then, we define structured MCN and CCN for $M^\dagger(\Psi)$ as follows:*

$$\mathscr{M}^\dagger\big(M(\Psi)\big) := \limsup_{\boldsymbol{\epsilon}\to 0} \left\{ \frac{\|M^\dagger(\Psi+\Delta\Psi) - M^\dagger(\Psi)\|_{\max}}{\boldsymbol{\epsilon}\|M^\dagger\|_{\max}} : \|\Delta\Psi/\Psi\|_\infty \leq \boldsymbol{\epsilon}, \Delta\Psi \in \mathcal{S}(\Psi) \right\},$$

$$\mathscr{C}^\dagger\big(M(\Psi)\big) := \limsup_{\boldsymbol{\epsilon}\to 0} \left\{ \frac{1}{\boldsymbol{\epsilon}} \left\| \frac{M^\dagger(\Psi+\Delta\Psi) - M^\dagger(\Psi)}{M^\dagger} \right\|_{\max} : \|\Delta\Psi/\Psi\|_\infty \leq \boldsymbol{\epsilon}, \Delta\Psi \in \mathcal{S}(\Psi) \right\}.$$

To formulate the general expressions for the upper bounds of the CNs outlined in Definition 6.3.1, we present the following perturbation expression for $M^\dagger(\Psi)$.

Lemma 6.3.1. *Let $M(\Psi) \in \mathbb{R}^{m\times n}$ and $\text{rank}(M(\Psi)) = r$. Suppose $\Delta\Psi \in \mathcal{S}(\Psi)$ is the perturbation on the parameter set Ψ. Then*

$$M^\dagger(\Psi+\Delta\Psi) - M^\dagger(\Psi) = \sum_{k=1}^{p} \Big(-M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} M^\dagger + M^\dagger M^{\dagger T} \Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T \mathbf{E}_M + \mathbf{F}_M \Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T M^{\dagger T} M^\dagger \Big) \Delta\psi_k + \mathcal{O}(\|\Delta\Psi\|_\infty^2).$$

Proof. Given that the elements of $M(\Psi)$ are differentiable functions of $\Psi = [\psi_1, \psi_2, ..., \psi_p]^T$, using matrix differentiation, for an infinitesimal change in $M(\Psi)$, we get

$$\Delta M(\Psi) = M(\Psi+\Delta\Psi) - M(\Psi) = \sum_{k=1}^{p} \frac{\partial M(\Psi)}{\partial \psi_k} \Delta\psi_k + \mathcal{O}(\|\Delta\Psi\|_\infty^2), \tag{6.3.2}$$

where $\Delta\Psi = [\Delta\psi_1, \ldots, \Delta\psi_p]^T$. Since $\Delta\Psi \in \mathcal{S}(\Psi)$, using the perturbation expression for the M-P inverse [128], we obtain

$$\Delta M^\dagger(\Psi) = M^\dagger(\Psi+\Delta\Psi) - M^\dagger(\Psi) = -M^\dagger \Delta M(\Psi) M^\dagger + M^\dagger M^{\dagger T} (\Delta M(\Psi))^T \mathbf{E}_M + \mathbf{F}_M (\Delta M(\Psi))^T M^{\dagger T} M^\dagger + \mathcal{O}(\|\Delta\Psi\|_\infty^2). \tag{6.3.3}$$

Putting (6.3.2) in (6.3.3), we get

$$\Delta M^\dagger(\Psi) = -M^\dagger \Big(\sum_{k=1}^{p} \frac{\partial M(\Psi)}{\partial \psi_k} \Delta\psi_k \Big) M^\dagger + M^\dagger M^{\dagger T} \Big(\sum_{k=1}^{p} \frac{\partial M(\Psi)}{\partial \psi_k} \Delta\psi_k \Big)^T \mathbf{E}_M + \mathbf{F}_M \Big(\sum_{k=1}^{p} \frac{\partial M(\Psi)}{\partial \psi_k} \Delta\psi_k \Big)^T M^{\dagger T} M^\dagger + \mathcal{O}(\|\Delta\Psi\|_\infty^2).$$

Hence, the desired expression is obtained. ■

In Theorem 6.3.2, for $M^\dagger(\Psi)$, we derive general expressions for upper bounds of the proposed CNs when $\text{rank}(M(\Psi)) = r$.

Theorem 6.3.2. *For $M(\Psi) \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(M(\Psi)) = r$, we have*

$$\mathscr{M}^{\dagger}\big(M(\Psi)\big) \leq \frac{\|\mathcal{X}_{\Psi}^{\dagger}\|_{\max}}{\|M^{\dagger}\|_{\max}} =: \widetilde{\mathscr{M}^{\dagger}}\big(M(\Psi)\big) \quad \textit{and} \quad \mathscr{C}^{\dagger}\big(M(\Psi)\big) \leq \Big\|\frac{\mathcal{X}_{\Psi}^{\dagger}}{M^{\dagger}}\Big\|_{\max} =: \widetilde{\mathscr{C}^{\dagger}}\big(M(\Psi)\big),$$

where

$$\mathcal{X}_{\Psi}^{\dagger} = \sum_{k=1}^{p}\Big(\Big|M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}M^{\dagger}\Big| + \Big|M^{\dagger}M^{\dagger^T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T\mathbf{E}_M\Big| + \Big|\mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T M^{\dagger^T}M^{\dagger}\Big|\Big)|\psi_k|.$$

Proof. From Lemma 6.3.1 and using the properties of absolute values, we have

$$\begin{aligned}\Big|M^{\dagger}(\Psi+\Delta\Psi) - M^{\dagger}(\Psi)\Big| \leq \sum_{k=1}^{p}\Big(&\Big|M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}M^{\dagger}\Big| + \Big|M^{\dagger}M^{\dagger^T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T\mathbf{E}_M\Big| \\ &+ \Big|\mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T M^{\dagger^T}M^{\dagger}\Big|\Big)|\Delta\psi_k| + \mathcal{O}(\|\Delta\Psi\|_{\infty}^2).\end{aligned}$$

Now, by Definition 6.3.1, $\|\Delta\Psi/\Psi\|_{\infty} \leq \boldsymbol{\epsilon}$ implies that $|\Delta\psi_k| \leq \boldsymbol{\epsilon}\,|\psi_k|$ for all $k = 1, 2, \ldots, p$, and using the properties of the max norm, we find that

$$\begin{aligned}\|M^{\dagger}(\Psi+\Delta\Psi) - M^{\dagger}(\Psi)\|_{\max} \leq \boldsymbol{\epsilon}\Big\|\sum_{k=1}^{p}\Big(&\Big|M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}M^{\dagger}\Big| + \Big|M^{\dagger}M^{\dagger^T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T\mathbf{E}_M\Big| \\ &+ \Big|\mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^T M^{\dagger^T}M^{\dagger}\Big|\Big)|\psi_k|\Big\|_{\max} + \mathcal{O}(\boldsymbol{\epsilon}^2).\end{aligned}$$

Then, if we take $\boldsymbol{\epsilon} \to 0$, and from Definition 6.3.1, we get the desired result of the first claim.

In a similar manner, we obtain the second part of the claim. ■

In the next corollary, we obtain bounds for the CNs for $M^{\dagger}$ in the unstructured case.

Corollary 6.3.1. *Suppose $M \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(M) = r$. Then*

$$\begin{aligned}\widetilde{\mathscr{M}^{\dagger}}(M) &= \frac{1}{\|M^{\dagger}\|_{\max}}\Big\||M^{\dagger}||M||M^{\dagger}| + |M^{\dagger}M^{\dagger^T}||M^T||\mathbf{E}_M| + |\mathbf{F}_M||M^T||M^{\dagger^T}M^{\dagger}|\Big\|_{\max}, \\ \widetilde{\mathscr{C}^{\dagger}}(M) &= \Big\|\frac{1}{M^{\dagger}}\Big(|M^{\dagger}||M||M^{\dagger}| + |M^{\dagger}M^{\dagger^T}||M^T||\mathbf{E}_M| + |\mathbf{F}_M||M^T||M^{\dagger^T}M^{\dagger}|\Big)\Big\|_{\max}.\end{aligned}$$

Proof. For any $M = [m_{ij}] \in \mathbb{R}^{m\times n}$ and any two column vectors a and b, we have

$$\frac{\partial M}{\partial m_{ij}} = e_i^m(e_j^n)^T \quad \text{and} \tag{6.3.4}$$

$$|ab^T| = |a||b^T|. \tag{6.3.5}$$

By considering the parameters as the entries of M itself, i.e., $\Psi = [\{m_{ij}\}_{i,j=1}^{m,n}]^T \in \mathbb{R}^{mn}$, and using (6.3.4), the sum expression in Theorem 6.3.2 can be written as:

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n}\Big(\Big|M^\dagger\frac{\partial M}{\partial m_{ij}}M^\dagger\Big| + \Big|M^\dagger {M^\dagger}^T\Big(\frac{\partial M}{\partial m_{ij}}\Big)^T\mathbf{E}_M\Big| + \Big|\mathbf{F}_M\Big(\frac{\partial M}{\partial m_{ij}}\Big)^T {M^\dagger}^T M^\dagger\Big|\Big)|m_{ij}| \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\Big(\Big|M^\dagger e_i^m (e_j^n)^T M^\dagger\Big| + \Big|M^\dagger {M^\dagger}^T e_j^n (e_i^m)^T\mathbf{E}_M| + \Big|\mathbf{F}_M e_j^n (e_i^m)^T {M^\dagger}^T M^\dagger\Big|\Big)|m_{ij}|.
\end{aligned} \tag{6.3.6}$$

Again, using (6.3.5), we can write (6.3.6) as

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n}\Big(|M^\dagger(:,i)||m_{ij}||M^\dagger(j,:)| + |M^\dagger {M^\dagger}^T(:,j)||m_{ij}||\mathbf{E}_M(i,:)| \\
&\qquad + |\mathbf{F}_M(:,j)||m_{ij}||{M^\dagger}^T M^\dagger(i,:)|\Big) \\
&= |M^\dagger||M||M^\dagger| + |M^\dagger {M^\dagger}^T||M^T||\mathbf{E}_M| + |\mathbf{F}_M||M^T||{M^\dagger}^T M^\dagger|.
\end{aligned} \tag{6.3.7}$$

The desired upper bounds will be obtained by substituting (6.3.7) in Theorem 6.3.2. ■

Next, we estimate the bounds for CNs under the constraints $\mathcal{R}(\Delta M(\Psi)) \subseteq \mathcal{R}(M(\Psi))$ and $\mathcal{R}(\Delta M^T(\Psi)) \subseteq \mathcal{R}(M^T(\Psi))$.

Proposition 6.3.3. *Let $M(\Psi) \in \mathbb{R}^{m\times n}$ be such that $\operatorname{rank}(M(\Psi)) = r$. Suppose that $\Delta\Psi \in \mathbb{R}^p$ is the perturbation on the parameter set Ψ satisfying the conditions, $\|M^\dagger\|_2\|\Delta M(\Psi)\|_2 < 1$, $\mathcal{R}(\Delta M(\Psi)) \subseteq \mathcal{R}(M(\Psi))$ and $\mathcal{R}(\Delta M^T(\Psi)) \subseteq \mathcal{R}(M^T(\Psi))$. Then*

$$M^\dagger(\Psi+\Delta\Psi) - M^\dagger(\Psi) = -\sum_{k=1}^{p} M^\dagger\frac{\partial M(\Psi)}{\partial\psi_k}M^\dagger\Delta\psi_k + \mathcal{O}(\|\Delta\Psi\|_\infty^2).$$

Furthermore,

$$\begin{aligned}
\widetilde{\mathscr{M}^\dagger}\big(M(\Psi)\big) &= \frac{\|\sum_{k=1}^{p}|M^\dagger\frac{\partial M(\Psi)}{\partial\psi_k}M^\dagger||\Delta\psi_k|\|_{\max}}{\|M^\dagger\|_{\max}}, \\
\widetilde{\mathscr{C}^\dagger}\big(M(\Psi)\big) &= \Big\|\frac{1}{M^\dagger}\sum_{k=1}^{p}|M^\dagger\frac{\partial M(\Psi)}{\partial\psi_k}M^\dagger||\Delta\psi_k|\Big\|_{\max}.
\end{aligned}$$

Proof. If $M(\Psi), \Delta M(\Psi) \in \mathbb{R}^{m\times n}$ satisfies the assumptions $\mathcal{R}(\Delta M(\Psi)) \subseteq \mathcal{R}(M(\Psi))$ and $\mathcal{R}(\Delta M^T(\Psi)) \subseteq \mathcal{R}(M^T(\Psi))$. Then

$$M(\Psi)M^\dagger(\Psi)\Delta M(\Psi) = \Delta M(\Psi) \quad\text{and}\quad M^T(\Psi){M^\dagger(\Psi)}^T\Delta M^T(\Psi) = \Delta M^T(\Psi). \tag{6.3.8}$$

In addition, if $\|M^\dagger(\Psi)\|_2\|\Delta M(\Psi)\|_2 < 1$ holds, it is shown in [22] that

$$M^\dagger(\Psi+\Delta\Psi) = (I_n + M^\dagger(\Psi)\Delta M(\Psi))^{-1}M^\dagger(\Psi). \tag{6.3.9}$$

Now, (6.3.8) implies $\Delta M^T(\Psi)\mathbf{E}_M = \mathbf{0}$ and $\mathbf{F}_M \Delta M^T(\Psi) = \mathbf{0}$. Again, (6.3.9) implies that $\text{rank}(M(\Psi + \Delta\Psi)) = \text{rank}(M(\Psi))$. Therefore, $\Delta\Psi \in \mathcal{S}(\Psi)$. Hence, from Lemma 6.3.1, we get the desired expression.

The proof of the second part is a direct consequence of the derived perturbation expansion and the proof technique employed in Theorem 6.3.2. ■

Remark 6.3.4. *Using Proposition 6.3.3, and in an analogous approach to Corollary 6.3.1, we can recover the bounds for unstructured CNs obtained in [141].*

For the matrices with full column rank, the next theorem provides exact expressions of CNs for $M^\dagger(\Psi)$, introduced in Definition 6.3.1.

Theorem 6.3.5. *For $M(\Psi) \in \mathbb{R}^{m\times n}$ with full column rank, we have*

$$\mathscr{M}^\dagger(M(\Psi)) = \frac{\|\hat{\mathcal{X}}_\Psi^\dagger\|_{\max}}{\|M^\dagger\|_{\max}} \quad \textit{and} \quad \mathscr{C}^\dagger(M(\Psi)) = \left\| \frac{\hat{\mathcal{X}}_\Psi^\dagger}{M^\dagger} \right\|_{\max},$$

where

$$\hat{\mathcal{X}}_\Psi^\dagger = \sum_{k=1}^{p} \left| M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} M^\dagger - (M^T M)^{-1} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{E}_M \right| |\psi_k|.$$

Proof. Since $M(\Psi)$ is of full column rank matrix, we have $\mathbf{F}_M = \mathbf{0}$. Now, applying Remark 6.2.3 on Lemma 6.3.1, we get following perturbation expression for $M^\dagger(\Psi)$

$$\Delta M^\dagger(\Psi) = M^\dagger(\Psi + \Delta\Psi) - M^\dagger(\Psi) = \sum_{k=1}^{p} \Big(- M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} M^\dagger + (M^T M)^{-1} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{E}_M \Big) \Delta\psi_k + \mathcal{O}(\|\Delta\Psi\|_\infty^2). \tag{6.3.10}$$

By employing a similar proof technique as in Theorem 6.3.2, and considering the given condition $\|\Delta\Psi/\Psi\|_\infty \leq \boldsymbol{\epsilon}$, we can establish the following bound:

$$\mathscr{M}^\dagger(M(\Psi)) \leq \frac{\left\| \sum_{k=1}^{p} \left| M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} M^\dagger - (M^T M)^{-1} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{E}_M \right| |\psi_k| \right\|_{\max}}{\|M^\dagger\|_{\max}}. \tag{6.3.11}$$

On the other hand, from Lemma 6.2.1 and Remark 6.2.3, it follows that we can consider arbitrary perturbation $\Delta\Psi \in \mathbb{R}^p$ on the parameter set Ψ. Choose

$$\Delta\psi_k = -\boldsymbol{\epsilon}\, \text{sign}(M_k)_{lq}\, \text{sign}(\psi_k)\, \psi_k, \tag{6.3.12}$$

where $M_k = M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} M^\dagger - (M^T M)^{-1} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{E}_M$, for $k = 1 : p$, $(M_k)_{lq}$ denotes the lq-th entry of the matrix M_k, and the indices l and q are such that

$$\left\| \sum_{k=1}^{p} |M_k||\psi_k| \right\|_{\max} = \left(\sum_{k=1}^{p} |M_k||\psi_k| \right)_{lq}.$$

The upper bound in (6.3.11) is obtained by inserting the values of (6.3.12) in the perturbation expression (6.3.10) and from the Definition 6.3.1. Therefore, the proof of the first claim is concluded.
The second part of the claim can be obtained in a similar approach. ■

6.3.2. MNLS Solution for General Parameterized Coefficient Matrices

Let us consider the LS problem for the parameterized matrix $M(\Psi) \in \mathbb{R}^{m\times n}$

$$\min_{z\in\mathbb{R}^n} \|M(\Psi)z - b\|_2 \tag{6.3.13}$$

with $\text{rank}(M(\Psi)) = r$ and $b \in \mathbb{R}^m$. When $M(\Psi)$ is rank deficient, the unique MNLS solution is provided by $\boldsymbol{x} = M(\Psi)^\dagger b$. Moreover, in this situation, $\boldsymbol{x}$ is not even a continuous function of the data, and small changes in $M(\Psi)$ can produce large changes to $\boldsymbol{x}$. This happens as a consequence of the behavior of the M-P inverse for any rank deficient matrix. Thus, according to Proposition 6.2.2, we consider the perturbation $\Delta\Psi \in \mathcal{S}(\Psi)$ for the parameters, and then the perturbed problem

$$\min_{z\in\mathbb{R}^n} \|M(\Psi + \Delta\Psi)z - (b + \Delta b)\|_2$$

has the MNLS solution $\tilde{\boldsymbol{x}} = M(\Psi + \Delta\Psi)^\dagger(b + \Delta b)$. Consider $\Delta\boldsymbol{x} = \tilde{\boldsymbol{x}} - \boldsymbol{x}$.

In Definition 6.3.2, for the MNLS solution $\boldsymbol{x}$, we introduce its structured MCN and CCN.

Definition 6.3.2. *Let $M(\Psi) \in \mathbb{R}^{m\times n}$ with $\text{rank}(M(\Psi)) = r$ and $b \in \mathbb{R}^m$. Then, we define structured MCN and CCN of $\boldsymbol{x}$ as follows:*

$$\mathscr{M}^\dagger(M(\Psi), b) := \limsup_{\boldsymbol{\epsilon}\to 0} \left\{ \frac{\|\Delta\boldsymbol{x}\|_\infty}{\boldsymbol{\epsilon}\|\boldsymbol{x}\|_\infty} : \|\Delta\Psi/\Psi\|_\infty \le \boldsymbol{\epsilon}, \|\Delta b/b\|_\infty \le \boldsymbol{\epsilon},\ \Delta\Psi \in \mathcal{S}(\Psi), \Delta b \in \mathbb{R}^m \right\},$$

$$\mathscr{C}^\dagger(M(\Psi), b) := \limsup_{\boldsymbol{\epsilon}\to 0} \left\{ \frac{1}{\boldsymbol{\epsilon}} \left\| \frac{\Delta\boldsymbol{x}}{\boldsymbol{x}} \right\|_\infty : \|\Delta\Psi/\Psi\|_\infty \le \boldsymbol{\epsilon}, \|\Delta b/b\|_\infty \le \boldsymbol{\epsilon},\ \Delta\Psi \in \mathcal{S}(\Psi), \Delta b \in \mathbb{R}^m \right\}.$$

Our main objective of this section is to find general expressions of bounds for the CNs introduced in Definition 6.3.2, and the following lemma provides the perturbation expansion for the MNLS solution.

Lemma 6.3.6. *Let $M(\Psi) \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(M(\Psi)) = r$ and $b \in \mathbb{R}^m$. Suppose $\Delta\Psi \in \mathcal{S}(\Psi)$ and $\Delta b \in \mathbb{R}^m$, and set $\mathbf{r} := b - M(\Psi)\boldsymbol{x}$. Then*

$$\Delta\boldsymbol{x} = \sum_{k=1}^{p}\Big(-M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}\boldsymbol{x} + M^{\dagger}M^{\dagger T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}\mathbf{r} + \mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}M^{\dagger T}\boldsymbol{x}\Big)\Delta\psi_k + \sum_{i=1}^{m} M^{\dagger}e_i^m\Delta b_i + \mathcal{O}(\|[\Delta\Psi,\Delta b]\|_{\infty}^2).$$

Proof. Since $\Delta\boldsymbol{x} = M^{\dagger}(\Psi+\Delta\Psi)(b+\Delta b) - M^{\dagger}(\Psi)b$ and $\Delta\Psi \in \mathcal{S}(\Psi)$, using Lemma 6.3.1, we get

$$\begin{aligned}\Delta\boldsymbol{x} &= \Bigg(\sum_{k=1}^{p}\Big(-M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}M^{\dagger} + M^{\dagger}M^{\dagger T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}\mathbf{E}_M + \mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}M^{\dagger T}M^{\dagger}\Big)\Delta\psi_k \\ &\quad + M^{\dagger}(\Psi)\Bigg)(b+\Delta b) - M^{\dagger}(\Psi)b + \mathcal{O}(\|\Delta\Psi\|_{\infty}^2) \qquad (6.3.14)\\ &= \sum_{k=1}^{p}\Big(-M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}\boldsymbol{x} + M^{\dagger}M^{\dagger T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}\mathbf{r} + \mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}M^{\dagger T}\boldsymbol{x}\Big)\Delta\psi_k \\ &\quad + M^{\dagger}(\Psi)\Delta b + \mathcal{O}(\|[\Delta\Psi,\Delta b]\|_{\infty}^2). \qquad (6.3.15)\end{aligned}$$

On the other hand, for the perturbation Δb in b, we can write $\Delta b = \sum_{i=1}^{m} e_i^m\Delta b_i$. Therefore, from (6.3.15), we get

$$\begin{aligned}\Delta\boldsymbol{x} &= \sum_{k=1}^{p}\Big(-M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}\boldsymbol{x} + M^{\dagger}M^{\dagger T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}\mathbf{r} + \mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}M^{\dagger T}\boldsymbol{x}\Big)\Delta\psi_k \\ &\quad + M^{\dagger}\sum_{i=1}^{m} e_i^m\Delta b_i + \mathcal{O}(\|[\Delta\Psi,\Delta b]\|_{\infty}^2),\end{aligned}$$

and hence, the desired result is obtained. ■

In Theorem 6.3.7, we provide general expressions for the upper bounds of $\mathscr{M}^{\dagger}(M(\Psi),b)$ and $\mathscr{C}^{\dagger}(M(\Psi),b)$.

Theorem 6.3.7. *Let $M(\Psi) \in \mathbb{R}^{m\times n}$ be a matrix having $\operatorname{rank}(M(\Psi)) = r$ and $b \in \mathbb{R}^m$. Then,*

$$\begin{aligned}\mathscr{M}^{\dagger}(M(\Psi),b) &\leq \frac{\|\mathcal{X}_{\Psi}^{ls}\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}} =: \widetilde{\mathscr{M}^{\dagger}}(M(\Psi),b),\\ \mathscr{C}^{\dagger}(M(\Psi),b) &\leq \left\|\mathfrak{D}_{\boldsymbol{x}^{\ddagger}}\mathcal{X}_{\Psi}^{ls}\right\|_{\infty} =: \widetilde{\mathscr{C}^{\dagger}}(M(\Psi),b),\end{aligned}$$

where

$$\mathcal{X}_{\Psi}^{ls} = \sum_{k=1}^{p}\Big(\Big|M^{\dagger}\frac{\partial M(\Psi)}{\partial \psi_k}\boldsymbol{x}\Big| + \Big|M^{\dagger}M^{\dagger T}\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}\mathbf{r}\Big| + \Big|\mathbf{F}_M\Big(\frac{\partial M(\Psi)}{\partial \psi_k}\Big)^{T}M^{\dagger T}\boldsymbol{x}\Big|\Big)|\psi_k| + |M^{\dagger}||b|$$

and $\mathfrak{D}_{\boldsymbol{x}^\ddagger} = \text{diag}(\boldsymbol{x}^\ddagger)$.

Proof. From Lemma 6.3.6 and utilizing the properties of absolute values, we obtain

$$|\Delta \boldsymbol{x}| \leq \sum_{k=1}^{p} \left(\left| M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} \boldsymbol{x} \right| + \left| M^\dagger M^{\dagger T} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{r} \right| + \left| \mathbf{F}_M \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T M^{\dagger T} \boldsymbol{x} \right| \right) |\Delta \psi_k| + \sum_{i=1}^{m} |M^\dagger| |\Delta b_i| + \mathcal{O}(\|[\Delta \Psi, \Delta b]_\infty\|^2). \tag{6.3.16}$$

Now, by Definition 6.3.2, $\|\Delta \Psi / \Psi\|_\infty \leq \boldsymbol{\epsilon}$ and $\|\Delta b / b\|_\infty \leq \boldsymbol{\epsilon}$ implies that for $k = 1 : p$, $|\Delta \psi_k| \leq \boldsymbol{\epsilon} \, |\psi_k|$, and for $i = 1 : m$, $|\Delta b_i| \leq \boldsymbol{\epsilon} |b_i|$, respectively. Taking infinity norm in (6.3.16), we deduced that

$$\|\Delta \boldsymbol{x}\|_\infty \leq \boldsymbol{\epsilon} \Big\| \sum_{k=1}^{p} \left(\left| M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} \boldsymbol{x} \right| + \left| M^\dagger M^{\dagger T} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{r} + \left| \mathbf{F}_M \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T M^{\dagger T} \boldsymbol{x} \right| \right) |\psi_k| + |M^\dagger| |b| \Big\|_\infty + \mathcal{O}(\boldsymbol{\epsilon}^2). \tag{6.3.17}$$

Then, if we take $\boldsymbol{\epsilon} \to 0$ in (6.3.17) and from Definition 6.3.2, we attain the desired result of the first assertion. The second assertion follows in a similar manner, as we can express $\left\| \frac{\Delta \boldsymbol{x}}{\boldsymbol{x}} \right\|_\infty = \|\mathfrak{D}_{\boldsymbol{x}^\ddagger} \Delta \boldsymbol{x}\|_\infty$. ■

Next, we discuss the bounds of the CNs for the MNLS solution of the LS problem (6.3.13) corresponding to unstructured matrices.

Corollary 6.3.2. *For* $M \in \mathbb{R}^{m \times n}$ *having* $\text{rank}(M) = r$ *and* $b \in \mathbb{R}^m$, *we have*

$$\widetilde{\mathscr{M}^\dagger}(M, b) := \frac{\Big\| |M^\dagger||M||\boldsymbol{x}| + |M^\dagger M^{\dagger T}||M^T||\mathbf{r}| + |\mathbf{F}_M||M^T||M^{\dagger T}\boldsymbol{x}| + |M^\dagger||b| \Big\|_\infty}{\|\boldsymbol{x}\|_\infty},$$

$$\widetilde{\mathscr{C}^\dagger}(M, b) := \Big\| \mathfrak{D}_{\boldsymbol{x}^\ddagger} \Big(|M^\dagger||M||\boldsymbol{x}| + |M^\dagger M^{\dagger T}||M^T||\mathbf{r}| + |\mathbf{F}_M||M^T||M^{\dagger T}\boldsymbol{x}| + |M^\dagger||b| \Big) \Big\|_\infty.$$

Proof. The proof follows in an analogous way to the Corollary 6.3.1 by considering $\Psi = [\{m_{ij}\}_{i,j=1}^{m,n}]^T \in \mathbb{R}^{mn}$ in Theorem 6.3.7 and using (6.3.4) and (6.3.5). ■

The next theorem offers explicit formulae of structured CNs for the unique LS solution $\boldsymbol{x} = M^\dagger(\Psi) b$ for full column rank matrices.

Theorem 6.3.8. *For* $M(\Psi) \in \mathbb{R}^{m \times n}$ *having full column rank and* $b \in \mathbb{R}^m$, *we get*

$$\mathscr{M}^\dagger\big(M(\Psi), b\big) = \frac{\|\hat{\mathcal{X}}_\Psi^{ls}\|_\infty}{\|\boldsymbol{x}\|_\infty} \quad \text{and} \quad \mathscr{C}^\dagger\big(M(\Psi), b\big) = \Big\| \mathfrak{D}_{\boldsymbol{x}^\ddagger} \hat{\mathcal{X}}_\Psi^{ls} \Big\|_\infty,$$

where

$$\hat{\mathcal{X}}_\Psi^{ls} = \sum_{k=1}^{p} \Big| M^\dagger \frac{\partial M(\Psi)}{\partial \psi_k} \boldsymbol{x} - (M^T M)^{-1} \Big(\frac{\partial M(\Psi)}{\partial \psi_k} \Big)^T \mathbf{r} \Big| |\psi_k| + |M^\dagger||b|. \tag{6.3.18}$$

Proof. In Lemma 6.3.6, using the fact that for any full column rank matrix $\mathbf{F}_M = \mathbf{0}$, we have

$$\Delta \boldsymbol{x} = \sum_{k=1}^{p} \left(-M^{\dagger} \frac{\partial M(\Psi)}{\partial \psi_k} \boldsymbol{x} + M^{\dagger} M^{\dagger T} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{r} \right) \Delta \psi_k + \sum_{i=1}^{m} M^{\dagger} e_i^m \Delta b_i + \mathcal{O}(\|[\Delta \psi, \Delta b]\|_{\infty}^2). \tag{6.3.19}$$

Now, by applying the proof method used in Theorem 6.3.5 and considering the given conditions $\|\Delta\Psi/\Psi\|_{\infty} \leq \boldsymbol{\epsilon}$ and $\|\Delta b/b\|_{\infty} \leq \boldsymbol{\epsilon}$, we obtain

$$\mathscr{M}^{\dagger}(M(\Psi), b) \leq \frac{\left\| \sum_{k=1}^{p} \left| M^{\dagger} \frac{\partial M(\Psi)}{\partial \psi_k} \boldsymbol{x} - (M^T M)^{-1} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{r} \right| |\psi_k| + |M^{\dagger}||b| \right\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}}. \tag{6.3.20}$$

From Lemma 6.2.1 and Remark 6.2.3, we can choose the perturbation $\Delta\Psi$ on the parameters set Ψ arbitrarily from $\mathbb{R}^p$. Consider the following perturbations

$$\Delta b = \boldsymbol{\epsilon}\, \boldsymbol{\Theta}_{M^{\dagger}} \boldsymbol{\Theta}_b\, b,$$

where $\boldsymbol{\Theta}_{M^{\dagger}}$ and $\boldsymbol{\Theta}_b$ are the diagonal matrices having diagonal entries $\boldsymbol{\Theta}_{M^{\dagger}}(i,i) = \operatorname{sign}(M^{\dagger}(l,i))$, and $\boldsymbol{\Theta}_b(i,i) = \operatorname{sign}(b_i)$ for $i = 1:m$, respectively, and

$$\Delta \psi_k = -\boldsymbol{\epsilon} \operatorname{sign}(M_{\boldsymbol{x},k})_l \operatorname{sign}(\psi_k)\, \psi_k,$$

where $M_{\boldsymbol{x},k} := M^{\dagger} \frac{\partial M(\Psi)}{\partial \psi_k} \boldsymbol{x} - (M^T M)^{-1} \left(\frac{\partial M(\Psi)}{\partial \psi_k} \right)^T \mathbf{r}$ and l is the index so that

$$\left\| \sum_{k=1}^{p} |M_{\boldsymbol{x},k}||\psi_k| + |M^{\dagger}||b| \right\|_{\infty} = \left(\sum_{k=1}^{p} |M_{\boldsymbol{x},k}||\psi_k| + |M^{\dagger}||b| \right)_l.$$

The upper bound in (6.3.20) will be attained by substituting these perturbations in (6.3.19) and from Definition 6.3.2, and hence the desired expression is attained. Analogously, we can obtain the expression for the CCN. ■

Remark 6.3.9. *The formula for the MCN $\mathscr{M}^{\dagger}(M(\Psi), b)$ in Theorem 6.3.8 is also presented in [145]. However, our approach to proving the theorem differs slightly. Interested readers may also refer to the proof method in [145].*

6.4. CNs for Cauchy-Vandermonde (CV) Matrices

In this section, we start by reviewing the definition of CV matrices. Subsequently, we provide the derivative expressions for the matrix with respect to its parameter set. These expressions play a pivotal role in the derivation of computationally feasible upper bounds for the structured CNs of the M-P inverse and the solution of the LS problem for a rank deficient CV matrix given in Theorems 6.4.2 and 6.4.5, respectively. Explicit formulations

for these CNs are also provided in the Theorems 6.4.3 and 6.4.6, respectively, when the matrix has full column rank.

Definition 6.4.1. *[78] A matrix $M \in \mathbb{R}^{m\times n}$ is classified as a CV matrix if it satisfies the following conditions: for $c = [c_1, c_2, \ldots, c_m]^T \in \mathbb{R}^m$ and $d = [d_1, d_2, \ldots, d_l]^T \in \mathbb{R}^l$, where $c_i \neq d_j$ for $i = 1 : m$ and $j = 1 : l$, with $0 \leq l \leq n$, the matrix M can be represented as follows:*

$$M = \begin{bmatrix} \frac{1}{c_1-d_1} & \frac{1}{c_1-d_2} & \cdots & \frac{1}{c_1-d_l} & 1 & c_1 & c_1^2 & \cdots & c_1^{n-l-1} \\ \frac{1}{c_2-d_1} & \frac{1}{c_2-d_2} & \cdots & \frac{1}{c_2-d_l} & 1 & c_2 & c_2^2 & \cdots & c_2^{n-l-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{c_m-d_1} & \frac{1}{c_m-d_2} & \cdots & \frac{1}{c_m-d_l} & 1 & c_m & c_m^2 & \cdots & c_m^{n-l-1} \end{bmatrix}. \tag{6.4.1}$$

M becomes the Vandermonde matrix when $l = 0$, and the Cauchy matrix when $l = n$.

For a CV matrix of the form (6.4.1), $\Psi_{\mathbb{CV}} := [\{c_i\}_{i=1}^m, \{d_i\}_{i=1}^l]^T \in \mathbb{R}^{m+l}$ represents its parameter set. We use the notation $M(\Psi_{\mathbb{CV}})$ to refer a CV matrix parameterized by $\Psi_{\mathbb{CV}}$.

For the M-P inverse and the MNLS solution involving the CV matrix, our objective is to estimate the structured CNs. Lemma 6.4.1 accomplishes our claim. Before that, we will construct the following matrices. For any positive integers p, q and any vector $y = [y_1, y_2, \ldots, y_p]^T \in \mathbb{R}^p$, define the matrices

$$\mathcal{Q}_{y,i}^{pq} := \Big[\mathbf{1}, \ldots, \mathbf{1}, y, \mathbf{1}, \ldots, \mathbf{1}\Big] \in \mathbb{R}^{p\times q},$$

for $i = 1 : q$, with the i-th column is y and $\mathbf{1} \in \mathbb{R}^p$ have all entries equal to 1. Also, set

$$\mathcal{M}_1 := \Big[-M(\Psi_{\mathbb{CV}})(:, 1 : l) \quad \mathbf{0} \quad M(\Psi_{\mathbb{CV}})(:, l + 2 : n)\Big] \in \mathbb{R}^{m\times n} \tag{6.4.2}$$

$$\text{and} \quad \mathcal{M}_2 := \Big[M(\Psi_{\mathbb{CV}})(:, 1 : l) \quad \mathbf{0}\Big] \in \mathbb{R}^{m\times n}. \tag{6.4.3}$$

The following lemma presents the derivative expressions of a CV matrix for the parameters in $\Psi_{\mathbb{CV}}$.

Lemma 6.4.1. *Suppose $M(\Psi_{\mathbb{CV}}) \in \mathbb{R}^{m\times n}$ having rank r, represented by a set of real parameter $\Psi_{\mathbb{CV}} = \Big[\{c_i\}_{i=1}^m, \{d_i\}_{i=1}^l\Big]^T \in \mathbb{R}^{m+l}$, where $c_i \neq d_j$, $i = 1 : m$ and $j = 1 : l$. Then, each entry of $M(\Psi_{\mathbb{CV}})$ is a differentiable function of $\Psi_{\mathbb{CV}}$, and*

1. $\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} = e_i^m(\mathcal{M}_1 \odot (\mathcal{Q}_{\mathbf{c}_i',i}^{nm})^T)(i, :)$ *for* $i = 1 : m$,
2. $\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial d_j} = (\mathcal{M}_2 \odot \mathcal{Q}_{\mathbf{d}_j',j}^{mn})(:, j)(e_j^n)^T$ *for* $j = 1 : l$,

where

$$\mathbf{c}'_i := \Big[\frac{1}{c_i-d_1}, \frac{1}{c_i-d_2}, \ldots, \frac{1}{c_i-d_l}, 1, \frac{1}{c_i}, \frac{2}{c_i}, \ldots, \frac{(n-l-1)}{c_i}\Big]^T \in \mathbb{R}^n,$$
$$\mathbf{d}'_j := \Big[\frac{1}{c_1-d_j}, \frac{1}{c_2-d_j}, \ldots, \frac{1}{c_m-d_j}\Big]^T \in \mathbb{R}^m,$$

for $i = 1:m$ *and* $j = 1:l$.

Proof. By observing that, when $c_i \neq d_j$, where $i = 1:m$ and $j = 1:l$, partial derivatives corresponding to the parameters $\{c_i\}_{i=1}^m$ will be

$$\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{-1}{(c_i-d_1)^2} & \cdots & \frac{-1}{(c_i-d_l)^2} & 0 & 1 & 2c_i & \cdots & (n-l-1)c_i^{n-l-2} \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Now, on the right-hand side of the above, using Hadamard product with the matrix $(\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T$, we get

$$\begin{aligned}\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} &= \Big(e_i^m \begin{bmatrix} -M(\Psi_{\mathbb{CV}})(i,1:l) & 0 & M(\Psi_{\mathbb{CV}})(i,l+2:n)\end{bmatrix}\Big) \odot (\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T \\ &= \Big(e_i^m \mathcal{M}_1(i,:)\Big) \odot (\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T = e_i^m(\mathcal{M}_1 \odot (\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T)(i,:).\end{aligned}$$

Hence, proof of the first statement follows.

In a similar argument, we can prove the second part of the statement. ■

For the structured CNs, computationally feasible upper bounds are provided in the following theorem for $M^\dagger(\Psi_{\mathbb{CV}})$ addressed in Definition 6.3.1.

Theorem 6.4.2. *Suppose* $M(\Psi_{\mathbb{CV}}) \in \mathbb{R}^{m\times n}$ *with* $\operatorname{rank}(M(\Psi_{\mathbb{CV}})) = r$. *Then*

$$\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{CV}})) = \frac{\|\mathcal{X}^\dagger_{\mathbb{CV}}\|_{\max}}{\|M^\dagger\|_{\max}} \quad \textit{and} \quad \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{CV}})) = \left\|\frac{\mathcal{X}^\dagger_{\mathbb{CV}}}{M^\dagger}\right\|_{\max},$$

where

$$\begin{aligned}\mathcal{X}^\dagger_{\mathbb{CV}} =& |M^\dagger||\mathfrak{D}_c|(\mathcal{M}_1 \odot \mathcal{Q})M^\dagger| + |M^\dagger M^{\dagger T}(\mathcal{M}_1 \odot \mathcal{Q})^T||\mathfrak{D}_c||\mathbf{E}_M| \\ &+ |\mathbf{F}_M(\mathcal{M}_1 \odot \mathcal{Q})^T||\mathfrak{D}_c||M^{\dagger T}M^\dagger| + |M^\dagger(\mathcal{M}_2 \odot \mathcal{M}_2)||\mathfrak{D}_{d'}||M^\dagger| \\ &+ |M^\dagger M^{\dagger T}||\mathfrak{D}_{d'}||(\mathcal{M}_2 \odot \mathcal{M}_2)^T\mathbf{E}_M| + |\mathbf{F}_M||\mathfrak{D}_{d'}||(\mathcal{M}_2 \odot \mathcal{M}_2)^T M^{\dagger T}M^\dagger|,\end{aligned}$$

$d' = [d_1, \ldots, d_l, 0, \ldots, 0]^T \in \mathbb{R}^n$, $\mathcal{Q} = [\mathbf{c}'_1, \mathbf{c}'_2, \ldots, \mathbf{c}'_m]^T \in \mathbb{R}^{m\times n}$ *and* $\mathcal{M}_1, \mathcal{M}_2$ *as defined in* (6.4.2) *and* (6.4.3), *respectively.*

Proof. For deriving the desired expressions for $\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathbb{CV}}))$ and $\widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathbb{CV}}))$, we calculate the contribution of each parameter subset to the expressions outlined in Theorem 6.3.2. For the parameters $\{c_i\}_{i=1}^{l}$, using (1) of Lemma 6.4.1, we get:

$$\begin{aligned}
\mathcal{E}_c :&= \sum_{i=1}^{m} \Big(\Big| M^{\dagger} \frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} M^{\dagger} \Big| + \Big| M^{\dagger T} M^{\dagger} \Big(\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} \Big)^T \mathbf{E}_M \Big| \\
&\quad + \Big| \mathbf{F}_M \Big(\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} \Big)^T M^{\dagger T} M^{\dagger} \Big| \Big) |c_i| \\
&= \sum_{i=1}^{m} \Big(\Big| M^{\dagger} e_i^m ((\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T \odot \mathcal{M}_1)(i,:) M^{\dagger} \Big| + \Big| M^{\dagger} M^{\dagger T} \Big(e_i^m ((\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T \odot \mathcal{M}_1)(i,:) \Big)^T \mathbf{E}_M \Big| \\
&\quad + \Big| \mathbf{F}_M \Big(e_i^m ((\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T \odot \mathcal{M}_1)(i,:) \Big)^T M^{\dagger T} M^{\dagger} \Big| \Big) |c_i|. \qquad (6.4.4)
\end{aligned}$$

Using (6.3.5) in (6.4.4), we get

$$\begin{aligned}
\mathcal{E}_c &= \sum_{i=1}^{m} \Big(|M^{\dagger}(:,i)||c_i||((\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T \odot \mathcal{M}_1)(i,:) M^{\dagger}| + |M^{\dagger} M^{\dagger T} (\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T \odot \mathcal{M}_1)^T (i,:)||c_i||\mathbf{E}_M(i,:)| \\
&\quad + |\mathbf{F}_M (\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T \odot \mathcal{M}_1)^T (i,:)||c_i||M^{\dagger T} M^{\dagger}(i,:)| \Big) \\
&= |M^{\dagger}||\mathfrak{D}_c||(\mathcal{M}_1 \odot \mathcal{Q}) M^{\dagger}| + |M^{\dagger} M^{\dagger T} (\mathcal{M}_1 \odot \mathcal{Q})^T||\mathfrak{D}_c||\mathbf{E}_M| + |\mathbf{F}_M (\mathcal{M}_1 \odot \mathcal{Q})^T||\mathfrak{D}_c||M^{\dagger T} M^{\dagger}|.
\end{aligned}$$

Analogously, for the parameters $\{d_i\}_{i=1}^{l}$, using (2) of Lemma 6.4.1, we have

$$\begin{aligned}
\mathcal{E}_d :&= \sum_{i=1}^{l} \Big(\Big| M^{\dagger} \frac{\partial M(\Psi_{\mathbb{CV}})}{\partial d_i} M^{\dagger} \Big| + \Big| M^{\dagger T} M^{\dagger} \Big(\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial d_i} \Big)^T \mathbf{E}_M \Big| \\
&\quad + \Big| \mathbf{F}_M \Big(\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial d_i} \Big)^T M^{\dagger T} M^{\dagger} \Big| \Big) |d_i| \\
&= |M^{\dagger}(\mathcal{M}_2 \odot \mathcal{M}_2)||\mathfrak{D}_{d'}||M^{\dagger}| + |M^{\dagger} M^{\dagger T}||\mathfrak{D}_{d'}||(\mathcal{M}_2 \odot \mathcal{M}_2)^T \mathbf{E}_M| \\
&\quad + |\mathbf{F}_M||\mathfrak{D}_{d'}||(\mathcal{M}_2 \odot \mathcal{M}_2)^T M^{\dagger T} M^{\dagger}|.
\end{aligned}$$

Applying Theorem 6.3.2 yields

$$\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathbb{CV}})) = \frac{\|\mathcal{E}_d + \mathcal{E}_c\|_{\max}}{\|M^{\dagger}\|_{\max}} \quad \text{and} \quad \widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathbb{CV}})) = \Big\| \frac{\mathcal{E}_d + \mathcal{E}_c}{M^{\dagger}} \Big\|_{\max}.$$

Hence, the proof is completed. ■

We now employ Theorem 6.3.5 to deduce the explicit formulations for structured CNs of $M^{\dagger}(\Psi_{\mathbb{CV}})$ by considering $M(\Psi)$ has full column rank, which is presented next.

Theorem 6.4.3. *Suppose $M(\Psi_{\mathbb{CV}}) \in \mathbb{R}^{m\times n}$ has full column rank. Then,*

$$\mathscr{M}^{\dagger}(M(\Psi_{\mathbb{CV}})) = \frac{\|\hat{\mathcal{X}}^{\dagger}_{\mathbb{CV}}\|_{\max}}{\|M^{\dagger}\|_{\max}} \quad \textit{and} \quad \mathscr{C}^{\dagger}(M(\Psi_{\mathbb{CV}})) = \left\| \frac{\hat{\mathcal{X}}^{\dagger}_{\mathbb{CV}}}{M^{\dagger}} \right\|_{\max},$$

where

$$\begin{aligned}\hat{\mathcal{X}}^{\dagger}_{\mathbb{CV}} =& \sum_{i=1}^{m} \left| M^{\dagger} E^{mm}_{ii} (\mathcal{M}_1 \odot \mathcal{Q}) M^{\dagger} - (M^T M)^{-1} (E^{mm}_{ii} (\mathcal{M}_1 \odot \mathcal{Q}))^T \mathbf{E}_M \right| |c_i| \\ &+ \sum_{i=1}^{l} \left| M^{\dagger} (\mathcal{M}_2 \odot \mathcal{M}_2) E^{nn}_{jj} M^{\dagger} - (M^T M)^{-1} ((\mathcal{M}_2 \odot \mathcal{M}_2) E^{nn}_{jj})^T \mathbf{E}_M \right| |d_i|,\end{aligned}$$

$E^{mn}_{ij} = e^m_i (e^n_j)^T$ and $\mathcal{Q}$ is as defined in Theorem 6.4.2.

Proof. To employ the expressions given in Theorem 6.3.5, we need to compute the contribution for each subset of parameters in an analogous method to the proof of Theorem 6.4.2. For the parameters $\{c_i\}_{i=1}^m$, using (1) of Lemma 6.4.1, we have

$$\begin{aligned}\mathcal{E}'_c :=& \sum_{i=1}^{m} \left| M^{\dagger} \frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} M^{\dagger} - (M^T M)^{-1} \Big(\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial c_i} \Big)^T \mathbf{E}_M \right| |c_i| \\ =& \sum_{i=1}^{m} \left| M^{\dagger} \Big(e^m_i (\mathcal{M}_1 \odot (\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T)(i,:) \Big) M^{\dagger} - (M^T M)^{-1} \Big(e^m_i (\mathcal{M}_1 \odot (\mathcal{Q}^{nm}_{\mathbf{c}'_i,i})^T)(i,:) \Big)^T \mathbf{E}_M \right| |c_i| \\ =& \sum_{i=1}^{m} \left| M^{\dagger} E^{mm}_{ii} (\mathcal{M}_1 \odot \mathcal{Q}) M^{\dagger} - (M^T M)^{-1} (E^{mm}_{ii} (\mathcal{M}_1 \odot \mathcal{Q}))^T \mathbf{E}_M \right| |c_i|.\end{aligned}$$

Similarly, for the parameters $\{d_i\}_{i=1}^l$, using (2) of Lemma 6.4.1, we get

$$\begin{aligned}\mathcal{E}'_d :=& \sum_{i=1}^{l} \left| M^{\dagger} \frac{\partial M(\Psi_{\mathbb{CV}})}{\partial d_i} M^{\dagger} - (M^T M)^{-1} \Big(\frac{\partial M(\Psi_{\mathbb{CV}})}{\partial d_i} \Big)^T \mathbf{E}_M \right| |d_i| \\ =& \sum_{i=1}^{l} \left| M^{\dagger} (\mathcal{M}_2 \odot \mathcal{M}_2) E^{nn}_{jj} M^{\dagger} - (M^T M)^{-1} ((\mathcal{M}_2 \odot \mathcal{M}_2) E^{nn}_{jj})^T \mathbf{E}_M \right| |d_i|.\end{aligned}$$

From Theorem 6.3.5, we have

$$\mathscr{M}^{\dagger}(M(\Psi_{\mathbb{CV}})) = \frac{\|\mathcal{E}'_c + \mathcal{E}'_d\|_{\max}}{\|M^{\dagger}\|_{\max}} \quad \text{and} \quad \mathscr{C}^{\dagger}(M(\Psi_{\mathbb{CV}})) = \left\| \frac{\mathcal{E}'_c + \mathcal{E}'_d}{M^{\dagger}} \right\|_{\max},$$

and hence, our proof is completed. ■

Remark 6.4.4. *If we consider $l = 0$ or $l = n$ in the preceding results, we can calculate the structured CNs for the M-P inverse Vandermonde and Cauchy matrices, respectively.*

Next, we consider the LS problem (6.3.13) corresponding to a rank deficient CV matrix. Using the expressions given in Theorem 6.3.7, we deduce upper bounds for structured CNs of $\boldsymbol{x}$, presented next.

Theorem 6.4.5. *Suppose* $M(\Psi_{\mathbb{CV}}) \in \mathbb{R}^{m\times n}$ *with rank* r *and* $b \in \mathbb{R}^m$. *Set* $\mathbf{r} := b - M(\Psi_{\mathbb{CV}})\boldsymbol{x}$. *Then,*

$$\widetilde{\mathscr{M}}^{\dagger}\big(M(\Psi_{\mathbb{CV}}), b\big) = \frac{\|\mathcal{X}^{ls}_{\mathbb{CV}}\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}} \quad \textit{and} \quad \widetilde{\mathscr{C}}^{\dagger}\big(M(\Psi_{\mathbb{CV}}, b)\big) = \Big\|\mathfrak{D}_{\boldsymbol{x}^{\ddagger}}\mathcal{X}^{ls}_{\mathbb{CV}}\Big\|_{\infty},$$

where

$$\begin{aligned}\mathcal{X}^{ls}_{\mathbb{CV}} =& |M^{\dagger}||b| + |M^{\dagger}||\mathfrak{D}_c||(\mathcal{M}_1 \odot \mathcal{Q})\boldsymbol{x}| + |M^{\dagger}M^{\dagger T}(\mathcal{M}_1 \odot \mathcal{Q})^T||\mathfrak{D}_c||\mathbf{r}| \\ &+ |\mathbf{F}_M(\mathcal{M}_1 \odot \mathcal{Q})^T||\mathfrak{D}_c||M^{\dagger T}\boldsymbol{x}| + |M^{\dagger}(\mathcal{M}_2 \odot \mathcal{M}_2)||\mathfrak{D}_{d'}||\boldsymbol{x}| \\ &+ |M^{\dagger}M^{\dagger T}||\mathfrak{D}_{d'}||\mathcal{M}_2 \odot \mathcal{M}_2)^T\mathbf{r}| + |\mathbf{F}_M||\mathfrak{D}_{d'}||(\mathcal{M}_2 \odot \mathcal{M}_2)^T M^{\dagger T}\boldsymbol{x}|.\end{aligned}$$

Proof. By evaluating in an analogous method to the proof of Theorem 6.4.2, for each subset of parameters using the expressions given in Theorem 6.3.7, the proof is followed. Hence, we omit it here. ■

The structured CNs to the LS problem (6.3.13) corresponding to a full column rank CV matrix are stated next.

Theorem 6.4.6. *Let* $M(\Psi_{\mathbb{CV}}) \in \mathbb{R}^{m\times n}$ *and* $b \in \mathbb{R}^m$. *Set* $\mathbf{r} := b - M\boldsymbol{x}$. *Then,*

$$\mathscr{M}^{\dagger}(M(\Psi_{\mathbb{CV}}), b) = \frac{\|\hat{\mathcal{X}}^{ls}_{\mathbb{CV}}\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}} \quad \textit{and} \quad \mathscr{C}^{\dagger}(M(\Psi_{\mathbb{CV}}), b) = \Big\|\mathfrak{D}\boldsymbol{x}^{\ddagger}\hat{\mathcal{X}}^{ls}_{\mathbb{CV}}\Big\|_{\infty},$$

where

$$\begin{aligned}\hat{\mathcal{X}}^{ls}_{\mathbb{CV}} =& \sum_{i=1}^{m}\Big|M^{\dagger}E^{mm}_{ii}(\mathcal{M}_1 \odot \mathcal{Q})\boldsymbol{x} - (M^TM)^{-1}(E^{mm}_{ii}(\mathcal{M}_1 \odot \mathcal{Q}))^T\mathbf{r}\Big||c_i| \\ &+ \sum_{i=1}^{l}\Big|M^{\dagger}(\mathcal{M}_2 \odot \mathcal{M}_2)E^{nn}_{jj}\boldsymbol{x} - (M^TM)^{-1}((\mathcal{M}_2 \odot \mathcal{M}_2)E^{nn}_{jj})^T\mathbf{r} + |M^{\dagger}||b|.\end{aligned}$$

Proof. Since the proof follows in an analogous method to the proof of Theorem 6.4.3, by finding the contribution of each parameter set in the expressions of Theorem 6.3.8. ■

6.5. Quasiseparable (QS) Matrices

The outset of this section begins with a quick introduction to QS matrices, which is a specific type of rank-structured matrices. Specifically, CNs are investigated for two important representations known as QS representation [58] and GV representation [58]. Upper bounds for the CNs of the M-P inverse and the MNLS solution are obtained corresponding to QS representation in Subsection 6.5.1 and for the GV representation in Subsection 6.5.2. The relationship between different CNs is also investigated in Subsection 6.5.3.

For the first time, QS matrices were investigated in [58]. In this work, we focus solely on considering $\{1,1\}$-QS matrices, which is a special case of QS matrices. Let M be in $\mathbb{R}^{n\times n}$. If every submatrix of M completely contained in the strictly lower triangular (resp., upper triangular) part is of rank at most 1 (resp., 1), and there is at least one of these submatrices has rank equal to 1 (resp., 1), then M is called a $\{1,1\}$-QS matrix. Equivalently, we can write:

$$\max_i \operatorname{rank}(M(i+1:n,1:i)) = 1 \quad \text{and} \quad \max_i \operatorname{rank}(M(1:i,i+1:n)) = 1.$$

6.5.1. CNs Corresponding to QS Representation

In [58], the notion of QS representation was proposed for $\{1,1\}$-QS matrices. In this subsection, we first recall this representation and then discuss the structured MCN and CCN.

Definition 6.5.1. *A matrix $M \in \mathbb{R}^{n\times n}$ is classified to be a $\{1,1\}$-QS matrix if it can be parameterized by the following set of $7n-8$ real parameters,*

$$\Psi_{\mathbb{QS}} = \left[\{\mathbf{a_i}\}_{i=2}^{n}, \{\mathbf{e_i}\}_{i=2}^{n-1}, \{\mathbf{b_i}\}_{i=1}^{n-1}, \{\mathbf{d_i}\}_{i=1}^{n}, \{\mathbf{f}_i\}_{i=1}^{n-1}, \{\mathbf{g}_i\}_{i=2}^{n-1}, \{\mathbf{h_i}\}_{i=2}^{n}\right]^T \in \mathbb{R}^{7n-8}, \quad (6.5.1)$$

as follows,

$$M = \begin{bmatrix} \mathbf{d}_1 & \mathbf{f}_1\mathbf{h}_2 & \mathbf{f}_1\mathbf{g}_2\mathbf{h}_3 & \cdots & \mathbf{f}_1\mathbf{g}_2\cdots\mathbf{g}_{n-1}\mathbf{h}_n \\ \mathbf{a}_2\mathbf{b}_1 & \mathbf{d}_2 & \mathbf{f}_2\mathbf{h}_3 & \cdots & \mathbf{f}_2\mathbf{g}_3\cdots\mathbf{g}_{n-1}\mathbf{h}_n \\ \mathbf{a}_3\mathbf{e}_2\mathbf{b}_1 & \mathbf{a}_3\mathbf{b}_2 & \mathbf{d}_3 & \cdots & \mathbf{f}_3\mathbf{g}_4\cdots\mathbf{g}_{n-1}\mathbf{h}_n \\ \mathbf{a}_4\mathbf{e}_3\mathbf{e}_2\mathbf{b}_1 & \mathbf{a}_4\mathbf{e}_3\mathbf{b}_2 & \mathbf{a}_4\mathbf{b}_3 & \cdots & \mathbf{f}_4\mathbf{g}_5\cdots\mathbf{g}_{n-1}\mathbf{h}_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_n\mathbf{e}_{n-1}\ldots\mathbf{e}_2\mathbf{b}_1 & \mathbf{a}_n\mathbf{e}_{n-1}\ldots\mathbf{e}_3\mathbf{b}_2 & \mathbf{a}_n\mathbf{e}_{n-1}\ldots\mathbf{e}_4\mathbf{b}_3 & \cdots & \mathbf{d}_n \end{bmatrix}.$$

The set of real parameters $\Psi_{\mathbb{QS}}$ as in (6.5.1) is called QS representation of M. We use the notation $M(\Psi_{\mathbb{QS}})$ to refer a $\{1,1\}$-QS matrix parameterized by the set $\Psi_{\mathbb{QS}}$. For the rest part, we set $M(\Psi_{\mathbb{QS}}) := \mathrm{L}_M + \mathrm{D}_M + \mathrm{U}_M$, where L_M and U_M denote the strictly lower and upper triangular part of $M(\Psi_{\mathbb{QS}})$, respectively, and D_M denotes the diagonal part of $M(\Psi_{\mathbb{QS}})$.

In Lemma 6.5.1, we recall the derivative expressions of a $\{1,1\}$-QS matrix $M(\Psi_{\mathbb{QS}})$ for the parameters in $\Psi_{\mathbb{QS}}$ provided in Definition 6.5.1, which are useful to obtain the desired upper bounds for CNs. These results are discussed in [56].

Lemma 6.5.1. *Let $M(\Psi_{\mathbb{QS}}) \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-QS matrix. Then $M(\Psi_{\mathbb{QS}})$ has entries that are differentiable functions of $\Psi_{\mathbb{QS}}$ defined as in (6.5.1), and*

1. $\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{d}_i} = e_i^n (e_i^n)^T$, *for* $i = 1:n$.
2. $\mathbf{a}_i \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{a}_i} = e_i^n \mathrm{L}_M(i,:)$, *for* $i = 2:n$.
3. $\mathbf{e}_i \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{e}_i} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ M(\Psi_{\mathbb{QS}})(i+1:n, 1:i-1) & \mathbf{0} \end{bmatrix} := \mathcal{F}_i$, *for* $i = 2:n-1$.
4. $\mathbf{b}_i \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{b}_i} = \mathrm{L}_M(:,i)(e_i^n)^T$, *for* $i = 1:n-1$.
5. $\mathbf{g}_i \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{g}_i} = \begin{bmatrix} \mathbf{0} & M(\Psi_{\mathbb{QS}})(1:i-1, i+1:n) \\ \mathbf{0} & \mathbf{0} \end{bmatrix} := \mathcal{G}_i$, *for* $i = 2:n-1$.
6. $\mathbf{f}_i \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{f}_i} = e_i^n \mathrm{U}_M(i,:)$, *for* $i = 1:n-1$.
7. $\mathbf{h}_i \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{h}_i} = \mathrm{U}_M(i,:)(e_i^n)^T$, *for* $i = 2:n$.

Next, we use the derivative expressions given in Lemma 6.5.1 and Theorem 6.3.2 to compute the bounds of the structured CNs for $M^\dagger(\Psi_{\mathbb{QS}})$.

Theorem 6.5.2. *For* $M(\Psi_{\mathbb{QS}}) \in \mathbb{R}^{n\times n}$ *with* $\operatorname{rank}(M(\Psi_{\mathbb{QS}})) = r$, *we have*

$$\widetilde{\mathscr{M}}^\dagger(M(\Psi_{\mathbb{QS}})) = \frac{\|\mathcal{X}^\dagger_{\mathbb{QS}}\|_{\max}}{\|M^\dagger\|_{\max}} \quad \textit{and} \quad \widetilde{\mathscr{C}}^\dagger(M(\Psi_{\mathbb{QS}})) = \left\| \frac{\mathcal{X}^\dagger_{\mathbb{QS}}}{M^\dagger} \right\|_{\max},$$

where

$$\begin{aligned}
\mathcal{X}^\dagger_{\mathbb{QS}} :=\ & |M^\dagger||\mathrm{D}_M||M^\dagger| + |M^\dagger M^{\dagger T}||\mathrm{D}_M||\mathbf{E}_M| + |\mathbf{F}_M||\mathrm{D}_M||M^{\dagger T} M^\dagger| + |M^\dagger||\mathrm{L}_M M^\dagger| \\
& + |M^\dagger M^{\dagger T} \mathrm{L}_M^T||\mathbf{E}_M| + |\mathbf{F}_M \mathrm{L}_M^T||M^{\dagger T} M^\dagger| + |M^\dagger \mathrm{L}_M||M^\dagger| + |M^\dagger M^{\dagger T}||\mathrm{L}_M^T \mathbf{E}_M| \\
& + |\mathbf{F}_M||\mathrm{L}_M^T M^{\dagger T} M^\dagger| + |M^\dagger||\mathrm{U}_M M^\dagger| + |M^\dagger M^{\dagger T} \mathrm{U}_M^T||\mathbf{E}_M| + |\mathbf{F}_M \mathrm{U}_M^T||M^{\dagger T} M^\dagger| \\
& + |M^\dagger \mathrm{U}_M||M^\dagger| + |M^\dagger M^{\dagger T}||\mathrm{U}_M^T \mathbf{E}_M| + |\mathbf{F}_M||\mathrm{U}_M^T M^{\dagger T} M^\dagger| \\
& + \sum_{i=2}^{n-1} \Big(|M^\dagger \mathcal{F}_i M^\dagger| + |M^\dagger M^{\dagger T} \mathcal{F}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{F}_i^T M^{\dagger T} M^\dagger| \Big) \\
& + \sum_{j=2}^{n-1} \Big(|M^\dagger \mathcal{G}_j M^\dagger| + |M^\dagger M^{\dagger T} \mathcal{G}_j^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{G}_j^T M^{\dagger T} M^\dagger| \Big),
\end{aligned}$$

$\mathcal{F}_i$, *and* $\mathcal{G}_i$ *defined as in Lemma 6.5.1.*

Proof. The proof of the above assertions involves determining the contribution of each subset of parameters to the expressions provided in Theorem 6.3.2. Using (1) of Lemma 6.5.1 for the parameters $\{\mathbf{d}_i\}_{i=1}^n$, we have:

$$\begin{aligned}
\mathcal{E}_{\mathbf{d}} := \sum_{i=1}^{n} \Big(& |M^\dagger \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{d}_i} M^\dagger||\mathbf{d}_i| + |M^\dagger M^{\dagger T} \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{d}_i} \Big)^T \mathbf{E}_M||\mathbf{d}_i| \\
& + |\mathbf{F}_M \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{d}_i} \Big)^T M^{\dagger T} M^\dagger||\mathbf{d}_i| \Big)
\end{aligned}$$

$$= \sum_{i=1}^{n} \Big(|M^{\dagger} e_i^n (e_i^n)^T M^{\dagger}||\mathbf{d}_i| + |M^{\dagger} {M^{\dagger}}^T e_i^n (e_i^n)^T \mathbf{E}_M||\mathbf{d}_i| + |\mathbf{F}_M e_i^n (e_i^n)^T {M^{\dagger}}^T M^{\dagger}||\mathbf{d}_i| \Big). \tag{6.5.2}$$

Using (6.3.5) in (6.5.2), we deduce

$$\begin{aligned}\mathcal{E}_{\mathbf{d}} &= \sum_{i=1}^{n} \Big(|M^{\dagger}(:,i)||\mathbf{d}_i||M^{\dagger}(i,:)| + |M^{\dagger} {M^{\dagger}}^T(:,i)||\mathbf{d}_i||\mathbf{E}_M(i,:)| + |\mathbf{F}_M(:,i)||\mathbf{d}_i||{M^{\dagger}}^T M^{\dagger}(i,:)| \Big) \\ &= |M^{\dagger}||\mathrm{D}_M||M^{\dagger}| + |M^{\dagger} {M^{\dagger}}^T||\mathrm{D}_M||\mathbf{E}_M| + |\mathbf{F}_M||\mathrm{D}_M||{M^{\dagger}}^T M^{\dagger}|.\end{aligned}$$

Similarly, for $\{\mathbf{a}_i\}_{i=2}^{n}$ and using (2) of Lemma 6.5.1, we get:

$$\begin{aligned}\mathcal{E}_{\mathbf{a}} := \sum_{i=2}^{n} \Big(&|M^{\dagger} \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{a}_i} M^{\dagger}||\mathbf{a}_i| + |M^{\dagger} {M^{\dagger}}^T \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{a}_i} \Big)^T \mathbf{E}_M||\mathbf{a}_i| \\ &+ |\mathbf{F}_M \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{a}_i} \Big)^T {M^{\dagger}}^T M^{\dagger}||\mathbf{a}_i| \Big) \\ = |M^{\dagger}||\mathrm{L}_M M^{\dagger}| &+ |M^{\dagger} {M^{\dagger}}^T \mathrm{L}_M^T||\mathbf{E}_M| + |\mathbf{F}_M \mathrm{L}_M^T||{M^{\dagger}}^T M^{\dagger}|.\end{aligned}$$

For the parameters $\{\mathbf{b}_i\}_{i=1}^{n-1}$ and using (4) of Lemma 6.5.1, we deduce:

$$\begin{aligned}\mathcal{E}_{\mathbf{b}} : = \sum_{i=2}^{n} \Big(&|M^{\dagger} \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{b}_i} M^{\dagger}||\mathbf{b}_i| + |M^{\dagger} {M^{\dagger}}^T \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{b}_i} \Big)^T \mathbf{E}_M||\mathbf{b}_i| \\ &+ |\mathbf{F}_M \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{b}_i} \Big)^T {M^{\dagger}}^T M^{\dagger}||\mathbf{b}_i| \Big) \\ = |M^{\dagger} \mathrm{L}_M||M^{\dagger}| &+ |M^{\dagger} {M^{\dagger}}^T||\mathrm{L}_M^T \mathbf{E}_M| + |\mathbf{F}_M||\mathrm{L}_M^T {M^{\dagger}}^T M^{\dagger}|.\end{aligned}$$

For the parameters $\{\mathbf{e}_i\}_{i=2}^{n-1}$, using (3) of Lemma 6.5.1, we have:

$$\begin{aligned}\mathcal{E}_{\mathbf{e}} := \sum_{i=2}^{n-1} \Big(&|M^{\dagger} \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{e}_i} M^{\dagger}||\mathbf{e}_i| + |M^{\dagger} {M^{\dagger}}^T \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{e}_i} \Big)^T \mathbf{E}_M||\mathbf{e}_i| \\ &+ |\mathbf{F} \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{e}_i} \Big)^T {M^{\dagger}}^T M^{\dagger}||\mathbf{e}_i| \Big) \\ = \sum_{i=2}^{n-1} \Big(&|M^{\dagger} \mathcal{F}_i M^{\dagger}| + |M^{\dagger} {M^{\dagger}}^T \mathcal{F}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{F}_i^T {M^{\dagger}}^T M^{\dagger}| \Big).\end{aligned}$$

In a similar approach, for the parameters $\{\mathbf{f}_i\}_{i=1}^{n-1}$, $\{\mathbf{g}_i\}_{i=2}^{n-1}$ and $\{\mathbf{h}_i\}_{i=2}^{n-1}$, we get:

$$\begin{aligned}\mathcal{E}_{\mathbf{f}} := \sum_{i=1}^{n-1} \Big(&|M^{\dagger} \frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{f}_i} M^{\dagger}||\mathbf{f}_i| + |M^{\dagger} {M^{\dagger}}^T \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{f}_i} \Big)^T \mathbf{E}_M||\mathbf{f}_i| \\ &+ |\mathbf{F}_M \Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{f}_i} \Big)^T {M^{\dagger}}^T M^{\dagger}||\mathbf{f}_i| \Big) \\ = |M^{\dagger}||\mathrm{U}_M M^{\dagger}| &+ |M^{\dagger} {M^{\dagger}}^T \mathrm{U}_M^T||\mathbf{E}_M| + |\mathbf{F}_M \mathrm{U}_M^T||{M^{\dagger}}^T M^{\dagger}|.\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{\mathbf{h}} &:= \sum_{i=2}^{n}\Big(|M^{\dagger}\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{h}_i}M^{\dagger}||\mathbf{h}_i| + |M^{\dagger}M^{\dagger^T}\Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{h}_i}\Big)^T\mathbf{E}_M||\mathbf{h}_i| \\ &\quad + |\mathbf{F}_M\Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{h}_i}\Big)^T M^{\dagger^T}M^{\dagger}||\mathbf{h}_i|\Big) \\ &= |M^{\dagger}\mathrm{U}_M||M^{\dagger}| + |M^{\dagger}||M^{\dagger^T}\mathrm{U}_M^T\mathbf{E}_M| + |\mathbf{F}_M||\mathrm{U}_M^T M^{\dagger^T}M^{\dagger}|.\end{aligned}$$

$$\begin{aligned}\mathcal{E}_{\mathbf{g}} &:= \sum_{i=2}^{n-1}\Big(|M^{\dagger}\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{g}_i}M^{\dagger}||\mathbf{g}_i| + |M^{\dagger}M^{\dagger^T}\Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{g}_i}\Big)^T\mathbf{E}_M||\mathbf{g}_i| \\ &\quad + |\mathbf{F}_M\Big(\frac{\partial M(\Psi_{\mathbb{QS}})}{\partial \mathbf{g}_i}\Big)^T M^{\dagger^T}M^{\dagger}||\mathbf{g}_i|\Big) \\ &= \sum_{i=2}^{n-1}\Big(|M^{\dagger}\mathcal{G}_i M^{\dagger}| + |M^{\dagger}M^{\dagger^T}\mathcal{G}_i^T\mathbf{E}_M| + |\mathbf{F}_M\mathcal{G}_i^T M^{\dagger^T}M^{\dagger}|\Big).\end{aligned}$$

Now, it is straightforward from Theorem 6.3.2 that

$$\mathcal{X}_{\mathbb{QS}}^{\dagger} = \mathcal{E}_{\mathbf{d}} + \mathcal{E}_{\mathbf{a}} + \mathcal{E}_{\mathbf{b}} + \mathcal{E}_{\mathbf{e}} + \mathcal{E}_{\mathbf{f}} + \mathcal{E}_{\mathbf{g}} + \mathcal{E}_{\mathbf{h}},$$

and hence, the proof is completed. ■

Remark 6.5.3. *When $M(\Psi_{\mathbb{QS}})$ is a $n\times n$ nonsingular $\{1,1\}$-QS matrix, we have $\mathbf{E}_M = \mathbf{0}$. Then, using the expressions in Theorem 6.3.5, and an analogous way to Theorem 6.5.2, exact formulae of the structured CNs can be deduced for the inversion of $M(\Psi_{\mathbb{QS}})$.*

For any $\{1,1\}$-QS matrix $M(\Psi_{\mathbb{QS}})$, it is worth noting that there exist infinitely many QS representations, as indicated by Dopico and Pomés [57]. The next result demonstrates that $\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathbb{QS}}))$ and $\widetilde{\mathscr{C}^{\dagger}}((\Psi_{\mathbb{QS}}))$ are independent of the QS representation used.

Proposition 6.5.4. *For any two representations $\Psi_{\mathbb{QS}}$ and $\Psi'_{\mathbb{QS}}$ of a $\{1,1\}$-QS matrix $M\in\mathbb{R}^{m\times n}$, we get*

$$\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathbb{QS}})) = \widetilde{\mathscr{M}^{\dagger}}(M(\Psi'_{\mathbb{QS}})) \quad \textit{and} \quad \widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathbb{QS}})) = \widetilde{\mathscr{C}^{\dagger}}(M(\Psi'_{\mathbb{QS}})).$$

Proof. Observe that the formulae given by the Theorem 6.5.2 only depend on the entries of M, $M^{\dagger}$, $\mathbf{E}_M$ and $\mathbf{F}_M$, but not on the specific selection of the parameter set; hence, the proof follows. ■

We provide upper bounds on the structured CNs of the MNLS solution of the LS problem (6.3.13) corresponding to $\{1,1\}$-QS matrices in the next theorem.

Theorem 6.5.5. *Let* $M(\Psi_{\mathbb{QS}}) \in \mathbb{R}^{n\times n}$ *be such that* $\operatorname{rank}(M(\Psi_{\mathbb{QS}})) = r$ *and* $b \in \mathbb{R}^n$. *Set* $\mathbf{r} := b - M(\Psi_{\mathbb{QS}})\boldsymbol{x}$. *Then*

$$\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathbb{QS}}), b) = \frac{\|\mathcal{X}^{ls}_{\mathbb{QS}}\|_{\infty}}{\|\boldsymbol{x}\|_{\infty}} \quad \textit{and} \quad \widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathbb{QS}}), b) = \left\| \frac{\mathcal{X}^{ls}_{\mathbb{QS}}}{\boldsymbol{x}} \right\|_{\infty},$$

where

$$\begin{aligned}
\mathcal{X}^{ls}_{\mathbb{QS}} :=\ & |M^{\dagger}||b| + |M^{\dagger}||\mathrm{D}_M||\boldsymbol{x}| + |M^{\dagger}M^{\dagger^T}||\mathrm{D}_M||\mathbf{r}| + |\mathbf{F}_M||\mathrm{D}_M||M^{\dagger^T}\boldsymbol{x}| + |M^{\dagger}||\mathrm{L}_M\boldsymbol{x}| \\
& + |M^{\dagger}M^{\dagger^T}\mathrm{L}_M^T||\mathbf{r}| + |\mathbf{F}_M\mathrm{L}_M^T||M^{\dagger^T}\boldsymbol{x}| + |M^{\dagger}\mathrm{L}_M||\boldsymbol{x}| + |M^{\dagger}M^{\dagger^T}||\mathrm{L}_M^T\mathbf{r}| \\
& + |\mathbf{F}_M||\mathrm{L}_M^T|M^{\dagger^T}\boldsymbol{x}| + |M^{\dagger}||\mathrm{U}_M\boldsymbol{x}| + |M^{\dagger}M^{\dagger^T}\mathrm{U}_M^T||\mathbf{r}| + |\mathbf{F}_M\mathrm{U}_M^T||M^{\dagger^T}\boldsymbol{x}| \\
& + |M^{\dagger}\mathrm{U}_M||\boldsymbol{x}| + |M^{\dagger}M^{\dagger^T}||\mathrm{U}_M^T\mathbf{r}| + |\mathbf{F}_M||\mathrm{U}_M^TM^{\dagger^T}\boldsymbol{x}| \\
& + \sum_{i=2}^{n-1}\Big(|M^{\dagger}\mathcal{F}_i\boldsymbol{x}| + |M^{\dagger}M^{\dagger^T}\mathcal{F}_i^T\mathbf{r}| + |\mathbf{F}_M\mathcal{F}_i^TM^{\dagger^T}\boldsymbol{x}|\Big) \\
& + \sum_{j=2}^{n-1}\Big(|M^{\dagger}\mathcal{G}_j\boldsymbol{x}| + |M^{\dagger}M^{\dagger^T}\mathcal{G}_j^T\mathbf{r}| + |\mathbf{F}_M\mathcal{G}_j^TM^{\dagger^T}\boldsymbol{x}|\Big).
\end{aligned}$$

Proof. The statement can be easily verified using a similar justification to that of the proof of the Theorem 6.5.2 and using the Theorem 6.3.7. Hence, we omit the proof. ■

Remark 6.5.6. *By considering* $M(\Psi_{\mathbb{QS}})$ *nonsingular, for the linear system* $M(\Psi_{\mathbb{QS}})\boldsymbol{x} = b$, *using the expression in Theorem 6.3.8 and* $\mathbf{r} = 0$, *we can obtain following expression for the structured MCN of* $\boldsymbol{x}$:

$$\begin{aligned}
\mathscr{M}^{\dagger}\big(M(\Psi_{\mathbb{QS}}), b\big) =& \frac{1}{\|\boldsymbol{x}\|_{\infty}}\Big\| |M^{-1}||b| + |M^{-1}||\mathrm{D}_M||\boldsymbol{x}| + |M^{-1}||\mathrm{L}_M\boldsymbol{x}| + |M^{-1}\mathrm{L}_M||\boldsymbol{x}| \\
& + |M^{-1}||\mathrm{U}_M\boldsymbol{x}| + |M^{-1}\mathrm{U}_M||\boldsymbol{x}| + \sum_{i=2}^{n-1}|M^{-1}\mathcal{F}_i\boldsymbol{x}| + \sum_{j=2}^{n-1}|M^{-1}\mathcal{G}_j\boldsymbol{x}|\Big\|_{\max}.
\end{aligned}$$

This result is the same as obtained by Dopico and Pomés [57].

6.5.2. CNs Corresponding to GV Representation

The GV representation, proposed initially in [130], is another essential representation for $\{1, 1\}$-QS matrices. This representation is used to enhance the stability of fast algorithms. In this subsection, we first review the GV representation together with its minor variant called GV representation through tangent.

Definition 6.5.2. *[130] Any $M \in \mathbb{R}^{n\times n}$ is classified to be a $\{1,1\}$-QS matrix if it can be represented by the parameter set*

$$\Psi_{\mathbb{QS}}^{\mathcal{GV}} = \left[\{p_i, q_i\}_{i=2}^{n-1}, \{u_i\}_{i=1}^{n-1}, \{\mathbf{d}_i\}_{i=1}^{n}, \{v_i\}_{i=1}^{n-1}, \{r_i, s_i\}_{i=2}^{n-1}\right]^T \in \mathbb{R}^{7n-10}, \tag{6.5.3}$$

satisfying the following properties,

1. *$\{p_i, q_i\}$ is a cosine-sine pair with $p_i^2 + q_i^2 = 1$, for every $i \in \{2 : n-1\}$,*
2. *$\{u_i\}_{i=1}^{n-1}$, $\{\mathbf{d}_i\}_{i=1}^{n}$, and $\{v_i\}_{i=1}^{n-1}$ are independent parameters,*
3. *$\{r_i, s_i\}$ is a cosine-sine pair with $r_i^2 + s_i^2 = 1$, for every $i \in \{2 : n-1\}$,*

as follows:

$$M = \begin{bmatrix} \mathbf{d}_1 & v_1r_2 & v_1s_2r_3 & \dots & v_1s_2\dots s_{n-2}r_{n-1} & v_1s_2\dots s_{n-1} \\ p_2u_1 & \mathbf{d}_2 & v_2r_3 & \dots & v_2s_3\dots s_{n-2}r_{n-1} & v_2s_3\dots s_{n-1} \\ p_3q_2u_1 & p_3u_2 & \mathbf{d}_3 & \dots & v_3s_4\dots s_{n-2}r_{n-1} & v_3s_4\dots s_{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{n-1}q_{n-2}\dots q_2u_1 & p_{n-1}q_{n-2}\dots q_3u_2 & p_{n-1}q_{n-2}\dots q_4u_3 & \cdots & \mathbf{d}_{n-1} & v_{n-1} \\ q_{n-1}\dots q_2u_1 & q_{n-1}\dots q_3u_2 & q_{n-1}\dots q_4u_3 & \cdots & u_{n-1} & \mathbf{d}_n \end{bmatrix}.$$

Note that the $\Psi_{\mathbb{QS}}^{\mathcal{GV}}$ is a special case of $\Psi_{\mathbb{QS}}$ by considering $\{\mathbf{a}_i, \mathbf{e}_i\}_{i=2}^{n-1} = \{p_i, q_i\}_{i=2}^{n-1}$, $\{\mathbf{b}_i\}_{i=1}^{n-1} = \{u_i\}_{i=1}^{n-1}$, $\{\mathbf{d}_i\}_{i=1}^{n} = \{\mathbf{d}_i\}_{i=1}^{n}$, $\{\mathbf{f}_i\}_{i=1}^{n-1} = \{v_i\}_{i=1}^{n-1}$, $\{\mathbf{g}_i, \mathbf{h}_i\}_{i=2}^{n-1} = \{s_i, r_i\}_{i=2}^{n-1}$, and $\mathbf{a}_n = \mathbf{h}_n = 1$ with additional conditions on the parameters. Since the parameters p_i and q_i are dependent, arbitrary perturbation to $\Psi_{\mathbb{QS}}^{\mathcal{GV}}$ will destroy the cosine-sine pairs and the same is true for r_i and s_i. Thus, it will be more sensible to restrict the perturbation that preserves the cosine-sine pair. Consequently, Dopico and Pomés [57] introduced a new representation called GV representation through tangent using their tangents.

Definition 6.5.3. *For the GV representation $\Psi_{\mathbb{QS}}^{\mathcal{GV}}$ as in (6.5.3), the GV representation through tangent is defined as*

$$\Psi_{\mathcal{GV}} = \left[\{t_i\}_{i=2}^{n-1}, \{u_i\}_{i=1}^{n-1}, \{\mathbf{d_i}\}_{i=1}^{n}, \{v_i\}_{i=1}^{n-1}, \{w_i\}_{i=2}^{n-1}\right]^T \in \mathbb{R}^{5n-6}, \tag{6.5.4}$$

where $p_i = \frac{1}{\sqrt{1+t_i^2}}$, $q_i = \frac{t_i}{\sqrt{1+t_i^2}}$ and $r_i = \frac{1}{\sqrt{1+w_i^2}}$, $s_i = \frac{w_i}{\sqrt{1+w_i^2}}$, for $i = 2 : n-1$.

We employ the notation $M(\Psi_{\mathcal{GV}})$ to refer a $\{1,1\}$-QS matrix parameterized by the set $\Psi_{\mathcal{GV}}$. The derivative expressions corresponding to the parameters in the representation $\Psi_{\mathcal{GV}}$ are revisited in the next lemma:

Lemma 6.5.7. *[56] Let $M(\Psi_{\mathcal{GV}}) \in \mathbb{R}^{n\times n}$ having $\operatorname{rank}(M(\Psi_{\mathcal{GV}})) = r$. Then each entry of $M(\Psi_{\mathcal{GV}})$ is differentiable functions of the elements in $\Psi_{\mathcal{GV}}$, and*

1. $t_i \frac{\partial M(\Psi_{\mathcal{GV}})}{\partial t_i} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ -q_i^2 M(\Psi_{\mathcal{GV}})(i, 1:i-1) & \mathbf{0} \\ p_i^2 M(\Psi_{\mathcal{GV}})(i+1:n, 1:i-1) & \mathbf{0} \end{bmatrix} := \mathcal{K}_i$, *for* $i = 2 : n-1$.

2. $w_i \frac{\partial M(\Psi_{\mathcal{GV}})}{\partial w_i} = \begin{bmatrix} \mathbf{0} & -s_i^2 M(\Psi_{\mathcal{GV}})(1:i-1, i) & r_i^2 M(\Psi_{\mathcal{GV}})(1:i-1, i+1:n) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} := \mathcal{L}_i$, *for* $i = 2 : n-1$.

Note: Partial derivative expressions corresponding to the parameters $\{u_i\}_{i=1}^{n-1}, \{\mathbf{d}_i\}_{i=1}^{n}$ and $\{v_i\}_{i=1}^{n-1}$ are same as the expression for the parameters $\{\mathbf{b}_i\}_{i=1}^{n-1}, \{\mathbf{d}_i\}_{i=1}^{n}$ and $\{\mathbf{f}_i\}_{i=1}^{n-1}$, respectively, given in the Lemma 6.5.1.

In Theorem 6.5.8, we discuss computationally feasible upper bounds for the structured CNs introduced in Definition 6.3.1 corresponding to the representation $\Psi_{\mathcal{GV}}$.

Theorem 6.5.8. *Let* $M(\Psi_{\mathcal{GV}}) \in \mathbb{R}^{n\times n}$ *with* $\operatorname{rank}(M(\Psi_{\mathcal{GV}})) = r$, *then*

$$\widetilde{\mathscr{M}^\dagger}\big(M(\Psi_{\mathcal{GV}})\big) = \frac{\|\mathcal{X}_{\mathcal{GV}}^\dagger\|_{\max}}{\|M^\dagger\|_{\max}} \quad \textit{and} \quad \widetilde{\mathscr{C}^\dagger}\big(M(\Psi_{\mathcal{GV}})\big) = \left\|\frac{\mathcal{X}_{\mathcal{GV}}^\dagger}{M^\dagger}\right\|_{\max},$$

where

$$\begin{aligned}
\mathcal{X}_{\mathcal{GV}}^\dagger :=\ & |M^\dagger||\mathrm{D}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{D}_M||\mathbf{E}_M| + |\mathbf{F}_M||\mathrm{D}_M||{M^\dagger}^T M^\dagger| + |M^\dagger \mathrm{L}_M||M^\dagger| \\
& + |M^\dagger {M^\dagger}^T||\mathrm{L}_M^T \mathbf{E}_M| + |\mathbf{F}_M||\mathrm{L}_M^T {M^\dagger}^T M^\dagger| + |M^\dagger||\mathrm{U}_M M^\dagger| + |M^\dagger {M^\dagger}^T \mathrm{U}_M^T||\mathbf{E}_M| \\
& + |\mathbf{F}_M \mathrm{U}_M^T||{M^\dagger}^T M^\dagger| + \sum_{i=2}^{n-1} \Big(|M^\dagger \mathcal{K}_i M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{K}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{K}_i^T {M^\dagger}^T M^\dagger|\Big) \\
& + \sum_{j=2}^{n-1} \Big(|M^\dagger \mathcal{L}_j M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{L}_j^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{L}_j^T {M^\dagger}^T M^\dagger|\Big),
\end{aligned}$$

$\mathcal{K}_i$ *and* $\mathcal{L}_i$ *are defined as in Lemma 6.5.7.*

The bounds of the structured CNs for $\boldsymbol{x}$ with respect to the representation $\Psi_{\mathcal{GV}}$ are given in the next theorem.

Theorem 6.5.9. *Let* $M(\Psi_{\mathcal{GV}}) \in \mathbb{R}^{n\times n}$ *with* $\operatorname{rank}(M(\Psi_{\mathcal{GV}})) = r$. *Set* $\mathbf{r} := b - M(\Psi_{\mathcal{GV}})\boldsymbol{x}$, *then*

$$\widetilde{\mathscr{M}^\dagger}\big(M(\Psi_{\mathcal{GV}}), b\big) = \frac{\|\mathcal{X}_{\mathcal{GV}}^{ls}\|_\infty}{\|\boldsymbol{x}\|_\infty} \quad \textit{and} \quad \widetilde{\mathscr{C}}\big(M(\Psi_{\mathcal{GV}}), b\big) = \left\|\frac{\mathcal{X}_{\mathcal{GV}}^{ls}}{\boldsymbol{x}}\right\|_\infty,$$

where

$$\begin{aligned}\mathcal{X}^{ls}_{\mathcal{GV}} :=& |M^\dagger||\mathrm{D}_M||\boldsymbol{x}| + |M^\dagger {M^\dagger}^T||\mathrm{D}_M||\mathbf{r}| + |\mathbf{F}_M||\mathrm{D}_M||{M^\dagger}^T\boldsymbol{x}| + |M^\dagger \mathrm{L}_M||\boldsymbol{x}| \\ &+ |M^\dagger {M^\dagger}^T||\mathrm{L}_M^T\mathbf{r}| + |\mathbf{F}_M||\mathrm{L}_M^T{M^\dagger}^T\boldsymbol{x}| + |M^\dagger||\mathrm{U}_M\boldsymbol{x}| + |M^\dagger {M^\dagger}^T\mathrm{U}_M^T||\mathbf{r}| \\ &+ |\mathbf{F}_M\mathrm{U}_M^T||{M^\dagger}^T\boldsymbol{x}| + \sum_{i=2}^{n}\Big(|M^\dagger\mathcal{K}_i\boldsymbol{x}| + |M^\dagger {M^\dagger}^T\mathcal{K}_i^T\mathbf{r}| + |\mathbf{F}_M\mathcal{K}_i^T{M^\dagger}^T\boldsymbol{x}|\Big) \\ &+ \sum_{i=2}^{n}\Big(|M^\dagger\mathcal{L}_i\boldsymbol{x}| + |M^\dagger {M^\dagger}^T\mathcal{L}_i^T\mathbf{r}| + |\mathbf{F}_M\mathcal{L}_i^T{M^\dagger}^T\boldsymbol{x}|\Big) + |M^\dagger||b|,\end{aligned}$$

$\mathcal{K}_i$ and $\mathcal{L}_i$ are defined as in Lemma 6.5.7.

The proof of Theorems 6.5.8 and 6.5.9 follows by using the similar proof technique of Theorem 6.5.2.

6.5.3. Comparisons Between Different CNs for $\{1,1\}$-QS Matrices

We compare structured and unstructured CNs for the M-P inverse in Proposition 6.5.10 and the MNLS solution in Proposition 6.5.11 for $\{1,1\}$-QS matrix. For unstructured CNs, we use the expressions given in Corollary 6.3.1 and Corollary 6.3.2. The next result describes that structured CNs for the parameter set $\Psi_{\mathbb{QS}}$ are smaller than unstructured ones for the M-P inverse up to an order of n.

Proposition 6.5.10. *Let $M(\Psi_{\mathbb{QS}}) \in \mathbb{R}^{n\times n}$ be such that* $\mathrm{rank}(M(\Psi_{\mathbb{QS}})) = r$*, then we get the following relations*

$$\widetilde{\mathscr{M}^\dagger}\left(M(\Psi_{\mathbb{QS}})\right) \leq n\ \widetilde{\mathscr{M}^\dagger}(M) \text{ and } \widetilde{\mathscr{C}^\dagger}\left(M(\Psi_{\mathbb{QS}})\right) \leq n\ \widetilde{\mathscr{C}^\dagger}(M).$$

Proof. Using the properties of absolute values and Theorem 6.5.2, we have:

$$\begin{aligned}\mathcal{X}^{\dagger}_{\mathbb{QS}} \leq& |M^\dagger||\mathrm{D}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{D}_M||\mathbf{E}_M| + |\mathbf{F}_M||\mathrm{D}_M||{M^\dagger}^T M^\dagger| \\ &+ 2|M^\dagger||\mathrm{L}_M||M^\dagger| + 2|M^\dagger {M^\dagger}^T||\mathrm{L}_M^T||\mathbf{E}_M| + 2|\mathbf{F}_M||\mathrm{L}_M^T||{M^\dagger}^T M^\dagger| \\ &+ 2|M^\dagger||\mathrm{U}_M||M^\dagger| + 2|M^\dagger {M^\dagger}^T||\mathrm{U}_M^T||\mathbf{E}_M| + 2|\mathbf{F}_M||\mathrm{U}_M^T||{M^\dagger}^T M^\dagger| \\ &+ \sum_{i=2}^{n-1}\Big(|M^\dagger||\mathcal{F}_i||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathcal{F}_i^T||\mathbf{E}_M| + |\mathbf{F}_M||\mathcal{F}_i^T||{M^\dagger}^T M^\dagger|\Big) \\ &+ \sum_{i=2}^{n-1}\Big(|M^\dagger||\mathcal{G}_i||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathcal{G}_i^T||\mathbf{E}_M| + |\mathbf{F}_M||\mathcal{G}_i^T||{M^\dagger}^T M^\dagger|\Big).\end{aligned}$$

Using $|\mathcal{F}_i| \leq |\mathrm{L}_M|$ and $|\mathcal{G}_i| \leq |\mathrm{U}_M|$, we get

$$\mathcal{X}^{\dagger}_{\mathbb{QS}} \leq n\Big(|M^\dagger||M||M^\dagger| + |M^\dagger {M^\dagger}^T||M^T||\mathbf{E}_M| + |\mathbf{F}_M||M^T||{M^\dagger}^T M^\dagger|\Big).$$

Therefore, the desired relations are obtained from Theorem 6.5.2 and Corollary 6.3.1. ■

A similar type of result also holds for the LS problem, which is given next. We remove the proof since it is analogous to Proposition 6.5.10.

Proposition 6.5.11. *Let $M(\Psi_{\mathbb{QS}}) \in \mathbb{R}^{n\times n}$ be as in Proposition 6.5.10 and $b \in \mathbb{R}^m$. Then, we get the following relations*

$$\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{QS}}), b) \leq n\ \widetilde{\mathscr{M}^\dagger}(M, b) \quad \textit{and} \quad \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}}), b) \leq\ n\ \widetilde{\mathscr{C}^\dagger}(M, b).$$

Next result discuss about the relationship between the $\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{QS}}))$ with $\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathcal{GV}}))$ and $\widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}}))$ with $\widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathcal{GV}}))$.

Proposition 6.5.12. *For the representations $\Psi_{\mathbb{QS}}$ and $\Psi_{\mathcal{GV}}$ of a $\{1,1\}$-QS matrix $M \in \mathbb{R}^{n\times n}$ with $\operatorname{rank}(M) = r$, following holds:*

$$\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathcal{GV}})) \leq \widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{QS}})) \quad \textit{and} \quad \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathcal{GV}})) \leq \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}})).$$

Proof. The proof will be followed by observing that

$$\begin{aligned}
\mathcal{K}_i &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ -q_i^2 M(i, 1:i-1) & \mathbf{0} \\ p_i^2 M(i+1:n, 1:i-1) & \mathbf{0} \end{bmatrix} \\
&= -q_i^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ Ms(i, 1:i-1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + p_i^2 \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ M(i+1:n, 1:i-1) & \mathbf{0} \end{bmatrix} \\
&= -q_i^2\, e_i^m \mathrm{L}_M(i,:) + p_i^2\, \mathcal{F}_i.
\end{aligned}$$

Now, using the properties $|p_i|^2 \leq 1$ and $|q_i|^2 \leq 1$, and (6.3.5), we obtain

$$\begin{aligned}
\sum_{i=2}^{n-1} \Big(|M^\dagger \mathcal{K}_i M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{K}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{K}_i^T {M^\dagger}^T M^\dagger| \Big) &\leq |M^\dagger||\mathrm{L}_M M^\dagger| + |M^\dagger {M^\dagger}^T \mathrm{L}_M^T||\mathbf{E}_M| \\
&+ |\mathbf{F}_M \mathrm{L}_M^T||{M^\dagger}^T M^\dagger| + \sum_{i=2}^{n-1} \Big(|M^\dagger \mathcal{F}_i M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{F}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{F}_i^T {M^\dagger}^T M^\dagger| \Big).
\end{aligned}$$

Similarly, we can write $\mathcal{L}_i = -s_i^2\, \mathrm{U}_M(:,i)(e_i^n)^T + r_i^2\, \mathcal{G}_i$. Therefore, using $|s_i|^2 \leq 1$ and $|r_i|^2 \leq 1$, and (6.3.5), we get

$$\begin{aligned}
\sum_{i=2}^{n-1} \Big(|M^\dagger \mathcal{L}_i M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{L}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{L}_i^T {M^\dagger}^T M^\dagger| \Big) &\leq |M^\dagger \mathrm{U}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{U}_M^T \mathbf{E}_M| \\
&+ |\mathbf{F}_M||\mathrm{U}_M^T {M^\dagger}^T M^\dagger| + \sum_{i=2}^{n-1} \Big(|M^\dagger \mathcal{G}_i M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{G}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{G}_i^T {M^\dagger}^T M^\dagger| \Big).
\end{aligned}$$

Hence, we get the desired relations from the above two inequalities, expressions from the Theorems 6.5.2 and 6.5.8. ■

Proposition 6.5.13 provides the relationship between CNs for LS problem (6.3.13) for any $\{1,1\}$-QS matrix corresponding to the parameter sets $\Psi_{\mathbb{QS}}$ and $\Psi_{\mathcal{GV}}$.

Proposition 6.5.13. *For the representations $\Psi_{\mathbb{QS}}$ and $\Psi_{\mathcal{GV}}$ of a $\{1,1\}$-QS matrix $M \in \mathbb{R}^{n\times n}$ having rank r and $b \in \mathbb{R}^n$, following holds:*

$$\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathcal{GV}}),b) \leq \widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{QS}}),b) \quad \text{and} \quad \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathcal{GV}}),b) \leq \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}}),b).$$

6.5.4. The Structured Effective CNs

The expressions in Theorems 6.5.2 and 6.5.5 can be computationally very expensive for large matrices due to the involvement of two sums. The effective CN for $\{1,1\}$-QS matrices was initially considered in [56, 57] for eigenvalue problem and linear system to reduce the computation complexity. In a similar fashion to avoid such problems, we propose in Definition 6.5.4, structured effective CNs $\widetilde{\mathscr{M}_f^\dagger}(M(\Psi_{\mathbb{QS}}))$ and $\widetilde{\mathscr{C}_f^\dagger}(M(\Psi_{\mathbb{QS}}))$, which have similar contribution as $\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{QS}}))$ and $\widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}}))$, respectively.

Definition 6.5.4. *Let $M(\Psi_{\mathbb{QS}}) \in \mathbb{R}^{n\times n}$ having $\operatorname{rank}(M(\Psi_{\mathbb{QS}})) = r$. Then for $M^\dagger(\Psi_{\mathbb{QS}})$, we define structured effective MCN and CCN as*

$$\widetilde{\mathscr{M}_f^\dagger}(M(\Psi_{\mathbb{QS}})) := \frac{\|\mathcal{X}_{f,\mathbb{QS}}^\dagger\|_{\max}}{\|M^\dagger\|_{\max}} \quad \text{and} \quad \widetilde{\mathscr{C}_f^\dagger}(M(\Psi_{\mathbb{QS}})) := \left\| \frac{\mathcal{X}_{f,\mathbb{QS}}^\dagger}{M^\dagger} \right\|_{\max},$$

where

$$\begin{aligned}
\mathcal{X}_{f,\mathbb{QS}}^\dagger :=\ & |M^\dagger||\mathrm{D}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{D}_M||\mathbf{E}_M| + |\mathbf{F}_M||\mathrm{D}_M||{M^\dagger}^T M^\dagger| + |M^\dagger||\mathrm{L}_M M^\dagger| \\
& + |M^\dagger {M^\dagger}^T \mathrm{L}_M^T||\mathbf{E}_M| + |\mathbf{F}_M \mathrm{L}_M^T||{M^\dagger}^T M^\dagger| + |M^\dagger \mathrm{L}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{L}_M^T \mathbf{E}_M| \\
& + |\mathbf{F}_M||\mathrm{L}_M^T {M^\dagger}^T M^\dagger| + |M^\dagger||\mathrm{U}_M M^\dagger| + |M^\dagger {M^\dagger}^T \mathrm{U}_M^T||\mathbf{E}_M| + |\mathbf{F}_M \mathrm{U}_M^T||{M^\dagger}^T M^\dagger| \\
& + |M^\dagger \mathrm{U}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{U}_M^T \mathbf{E}_M| + |\mathbf{F}_M||\mathrm{U}_M^T {M^\dagger}^T M^\dagger|.
\end{aligned}$$

The following theorem demonstrates that the contribution of the sum terms in the expression of $\widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{QS}}))$ and $\widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}}))$ are negligible and reliably estimated up to a multiple n.

Theorem 6.5.14. *Under the same hypothesis as in Definition 6.5.4, following relations holds*

$$\begin{aligned}
\widetilde{\mathscr{M}_f^\dagger}(M(\Psi_{\mathbb{QS}})) \leq \widetilde{\mathscr{M}^\dagger}(M(\Psi_{\mathbb{QS}})) \leq (n-1)\,\widetilde{\mathscr{M}_f^\dagger}(M(\Psi_{\mathbb{QS}})), \\
\widetilde{\mathscr{C}_f^\dagger}(M(\Psi_{\mathbb{QS}})) \leq \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}})) \leq (n-1)\,\widetilde{\mathscr{C}_f^\dagger}(M(\Psi_{\mathbb{QS}})).
\end{aligned}$$

Proof. The inequalities on the left side are trivial. For inequalities on the right side, we set

$$\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ M(\Psi_{\mathbb{QS}})(i+1:n,1:i-1) & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathrm{L}_M(i+1:n,:) \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathrm{L}_M(i+1:n,i:n) \end{bmatrix}.$$

By using the above, we get

$$\begin{aligned}\Big(|M^\dagger \mathcal{F}_i M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{F}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{F}_i^T {M^\dagger}^T M^\dagger|\Big) &\leq |M^\dagger||\mathrm{L}_M M^\dagger| + |M^\dagger {M^\dagger}^T \mathrm{L}_M^T||\mathbf{E}_M| \\ &+ |\mathbf{F}_M \mathrm{L}_M^T||{M^\dagger}^T M^\dagger| + |M^\dagger \mathrm{L}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{L}_M^T \mathbf{E}_M| + |\mathbf{F}_M||\mathrm{L}_M^T|{M^\dagger}^T M^\dagger|.\end{aligned}$$

Again,

$$\begin{bmatrix} \mathbf{0} & M(\Psi_{\mathbb{QS}})(1:i-1,i+1:n) \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathrm{U}_M(:,i+1:n) \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathrm{U}_M(i:n,i+1:n) \end{bmatrix}.$$

By using the above, we get

$$\begin{aligned}\Big(|M^\dagger \mathcal{G}_i M^\dagger| + |M^\dagger {M^\dagger}^T \mathcal{G}_i^T \mathbf{E}_M| + |\mathbf{F}_M \mathcal{G}_i^T {M^\dagger}^T M^\dagger|\Big) &\leq |M^\dagger||\mathrm{U}_M M^\dagger| + |M^\dagger {M^\dagger}^T \mathrm{U}_M^T||\mathbf{E}_M| \\ &+ |\mathbf{F}_M \mathrm{U}_M^T||{M^\dagger}^T M^\dagger| + |M^\dagger \mathrm{U}_M||M^\dagger| + |M^\dagger {M^\dagger}^T||\mathrm{U}_M^T \mathbf{E}_M| + |\mathbf{F}_M||\mathrm{U}_M^T {M^\dagger}^T M^\dagger|.\end{aligned}$$

Hence, the desired result is straightforward from Definition 6.5.4. ■

Next, similarly to the Definition 6.5.4, we define structured effective CNs for the MNLS solution.

Definition 6.5.5. *Let $M(\Psi_{\mathbb{QS}}) \in \mathbb{R}^{n\times n}$ having $\mathrm{rank}(M(\Psi_{\mathbb{QS}})) = r$ and $b \in \mathbb{R}^m$. Then, for the MNLS solution $\boldsymbol{x}$, we define structured effective MCN and CCN as follows:*

$$\widetilde{\mathscr{M}_f^\dagger}(M(\Psi_{\mathbb{QS}}),b) := \frac{\|\mathcal{X}_{f,\mathbb{QS}}^{ls}\|_\infty}{\|\boldsymbol{x}\|_\infty}, \quad \widetilde{\mathscr{C}^\dagger}(M(\Psi_{\mathbb{QS}}),b) := \left\| \frac{\mathcal{X}_{f,\mathbb{QS}}^{ls}}{\boldsymbol{x}} \right\|_\infty,$$

where

$$\begin{aligned}\mathcal{X}_{f,\mathbb{QS}}^{ls} :&= |M^\dagger||\mathrm{D}_M||\boldsymbol{x}| + |M^\dagger {M^\dagger}^T||\mathrm{D}_M||\mathbf{r}| + |\mathbf{F}_M||\mathrm{D}_M||{M^\dagger}^T \boldsymbol{x}| + |M^\dagger||\mathrm{L}_M \boldsymbol{x}| \\ &+ |M^\dagger {M^\dagger}^T \mathrm{L}_M^T||\mathbf{r}| + |\mathbf{F}_M \mathrm{L}_M^T||{M^\dagger}^T \boldsymbol{x}| + |M^\dagger \mathrm{L}_M||\boldsymbol{x}| + |M^\dagger {M^\dagger}^T||\mathrm{L}_M^T \mathbf{r}| \\ &+ |\mathbf{F}_M||\mathrm{L}_M^T {M^\dagger}^T \boldsymbol{x}| + |M^\dagger||\mathrm{U}_M \boldsymbol{x}| + |M^\dagger {M^\dagger}^T \mathrm{U}_M^T||\mathbf{r}| + |\mathbf{F}_M \mathrm{U}_M^T||{M^\dagger}^T \boldsymbol{x}| \\ &+ |M^\dagger \mathrm{U}_M||\boldsymbol{x}| + |M^\dagger {M^\dagger}^T||\mathrm{U}_M^T \mathbf{r}| + |\mathbf{F}_M||\mathrm{U}_M^T {M^\dagger}^T \boldsymbol{x}| + |M^\dagger||b|.\end{aligned}$$

Similar results also hold for the MNLS solution as discussed in Theorem 6.5.14.

6.6. Numerical Experiments

We reported a few numerical experiments in this section to illustrate the theoretical findings covered in Sections 6.4 and 6.5, respectively. For all numerical computations, we have used MATLAB R2022b.

Example 6.6.1. [155] Let $M \in \mathbb{R}^{12\times 20}$ be a CV matrix as in Definition 6.5.1 with the parameter set $\Psi_{\mathbb{CV}} = [\{c_i\}_{i=1}^{12}, \{d_i\}_{i=1}^{8}]^T \in \mathbb{R}^{20}$, where

$$\begin{cases} c_i = \frac{i}{20}, & i = 1:12, \\ d_j = \frac{j+4}{50}, & j = 1:8. \end{cases} \tag{6.6.1}$$

For the MNLS solution, we generate a random vector $b \in \mathbb{R}^{12}$ in MATLAB by the command `randn`. The computed results for the bounds of structured and unstructured CNs are listed in Table 6.6.1. We observed that the bounds for the structured CNs are less than the order of 3 or 4 compared to the unstructured case.

Table 6.6.1: Comparison between upper bounds of structured and unstructured CNs for $M^{\dagger}(\Psi_{\mathbb{CV}})$ and $M^{\dagger}(\Psi_{\mathbb{CV}}, b)$ for Example 6.6.1.

$\widetilde{\mathscr{M}^{\dagger}}(M)$	$\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathbb{CV}}))$	$\widetilde{\mathscr{C}^{\dagger}}(M)$	$\widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathbb{CV}}))$	$\widetilde{\mathscr{M}^{\dagger}}(M,b)$	$\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathbb{CV}}),b)$	$\widetilde{\mathscr{C}^{\dagger}}(M,b)$	$\widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathbb{CV}}),b)$
$1.1568e+05$	$8.5419e+01$	$5.3826e+07$	$3.3542e+04$	$1.8008e+05$	$8.9389e+01$	$3.6180e+05$	$1.7959e+02$

Example 6.6.2. We consider several $\{1,1\}$-QS matrices of different orders. In fact, we choose $n = 5$, $n = 7$ and $n = 10$. We generate the random vectors $\mathbf{a}, \mathbf{b}, \mathbf{e}, \mathbf{d}, \mathbf{f}, \mathbf{g}$, and $\mathbf{h}$ by the command `randn` in MATLAB. After generating these vectors, we take following scaling

$$\mathbf{a} = a * 10^k, \ \mathbf{e} = e * 10^k, \ \mathbf{h} = h * 10^k,$$

where $k \in \{-1, -2, 0, 1, 2, 3\}$ to get unbalanced lower and upper right corner.

In Table 6.6.2, we compare $\widetilde{\mathscr{M}_f^{\dagger}}(M(\Psi_{\mathbb{QS}}))$ and $\widetilde{\mathscr{C}_f^{\dagger}}(M(\Psi_{\mathbb{QS}}))$ with $\widetilde{\mathscr{M}^{\dagger}}(M)$ and $\widetilde{\mathscr{C}^{\dagger}}(M)$, respectively, for the M-P inverse with different values of n. For the MNLS solution, we generate a random vector $b \in \mathbb{R}^n$ for each choice of n. The computed bounds for the CNs are listed in Table 6.6.2. These results demonstrate the reliability of proposed CNs.

Example 6.6.3. In this example, we compare structured MCN and CCN with respect to QS representation and GV representation through tangent, structured effective CNs with their unstructured ones for M-P inverse and MNLS solution. On account of these, we generate the random vectors $t \in \mathbb{R}^{n-2}, u \in \mathbb{R}^{n-1}, v \in \mathbb{R}^{n-1}, w \in \mathbb{R}^{n-1}$ in MATLAB by the

Table 6.6.2: Comparison between the upper bounds of unstructured, structured CNs and structured effective CNs for the M-P inverse and the MNLS solution of $\{1,1\}$-QS matrices for Example 6.6.2.

n	$\widetilde{\mathscr{M}^{\dagger}}(M)$	$\widetilde{\mathscr{M}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{C}^{\dagger}}(M)$	$\widetilde{\mathscr{C}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{M}^{\dagger}}(M,b)$	$\widetilde{\mathscr{M}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}),b)$	$\widetilde{\mathscr{C}^{\dagger}}(M,b)$	$\widetilde{\mathscr{C}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}),b)$
5	1.4330e+04	2	2.1690e+12	910	9.333e+04	3.0030	1.1910e+05	4.5451
7	5.9139e+03	3.0246	9.2340e+08	74.2974	6.5311e+04	4.6655	2.0255e+ 04	6.7386
10	7.3293e+04	2.0101	1.5858e+10	94.0499	6.3988e+04	1.0127	2.5381e+05	11.1271

comand `randn` for the parameter set $\Psi_{\mathcal{GV}}$. We set $\mathbf{d} = \mathbf{0} \in \mathbb{R}^n$ and rescale the vector v as $v(1) = 0$ and $v(n-1) = 10^2$. Then, we compute the $\{1,1\}$-QS matrix M as in Definition 6.5.2. For different values of n, we generate 100 rank deficient $\{1,1\}$-QS matrices. We use the formulae provided in Theorem 6.5.8 to compute the upper bounds $\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathcal{GV}}))$ and $\widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathcal{GV}}))$. Again, we use the formulae for $\widetilde{\mathscr{M}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}))$ and $\widetilde{\mathscr{C}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}))$ presented as in Definition 6.5.4, and for $\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathbb{QS}}))$ and $\widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathbb{QS}}))$ presented as in Theorem 6.5.2. For the upper bounds of unstructured CNs, we consider the formulae given in Corollary 6.3.1. We computed the above values for 100 randomly generated $\{1,1\}$-QS matrices for $n = 30, 40, 50$ and 60. In Table 6.6.3, average values of each upper bound of the CNs for the M-P inverse of these $\{1,1\}$-QS matrices are listed.

Table 6.6.3: Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the M-P inverse of $\{1,1\}$-QS matrices for Example 6.6.3.

mean	n	$\widetilde{\mathscr{M}^{\dagger}}(M)$	$\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathcal{GV}}))$	$\widetilde{\mathscr{M}^{\dagger}}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{M}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{C}^{\dagger}}(M)$	$\widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathcal{GV}}))$	$\widetilde{\mathscr{C}^{\dagger}}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{C}^{\dagger}_{f}}(M(\Psi_{\mathbb{QS}}))$
	30	$1.0667e+02$	$7.4873e+01$	$1.2388e+02$	$8.9215e+01$	$2.3780e+05$	$1.2730e+04$	$2.2102e+04$	$1.5815e+04$
	40	$1.1429e+02$	$7.4757e+01$	$1.2356e+02$	$8.9427e+01$	$6.1267e+08$	$1.4452e+05$	$2.3753e+05$	$1.7285e+05$
	50	$1.6711e+02$	$1.0323e+02$	$1.7182e+02$	$1.2296e+02$	$6.9215e+06$	$1.6133e+05$	$3.2430e+05$	$2.1980e+05$
	60	$3.2135e+02$	$1.8140e+02$	$2.9452e+02$	$2.1359e+02$	$2.7856e+07$	$3.0500e+05$	$5.2242e+05$	$3.9791e+05$

Next, we generate a random vector $b \in \mathbb{R}^n$, for each value of n, and upper bounds for the structured CNs with respect to QS representation and GV representation through tangent, structured effective CNs and unstructured CNs for the MNLS solution are listed in Table 6.6.4.

The computed results in Tables 6.6.3 and 6.6.4 show the consistency of Propositions 6.5.10–6.5.13 and Theorem 6.5.14. We can observe that the upper bounds of CNs for GV representation through tangent give more reliable bounds compared to the other CNs.

Table 6.6.4: Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the MNLS solution for $\{1,1\}$-QS matrices for Example 6.6.3.

mean	n	$\widetilde{\mathscr{M}}^{\dagger}(M,b)$	$\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathcal{GV}}),b)$	$\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$	$\widetilde{\mathscr{M}}_f^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$	$\widetilde{\mathscr{C}}^{\dagger}(M,b)$	$\widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathcal{GV}}),b)$	$\widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$	$\widetilde{\mathscr{C}}_f^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$
	30	$1.1278e+02$	$7.8851e+01$	$2.6113e+02$	$9.4414e+01$	$4.6667e+03$	$1.7445e+03$	$2.2102e+04$	$2.1646e+03$
	40	$1.1990e+02$	$7.7823e+01$	$2.4226e+02$	$9.3005e+01$	$7.3841e+03$	$4.1774e+03$	$2.3753e+05$	$5.2069e+03$
	50	$1.6512e+02$	$1.0080e+02$	$3.2184e+02$	$1.2072e+02$	$9.2143e+03$	$4.2351e+03$	$3.2430e+05$	$5.3584e+03$
	60	$3.3149e+02$	$1.8715e+02$	$7.6179e+02$	$2.2120e+02$	$7.0839e+04$	$5.9482e+04$	$5.2242e+05$	$7.2054e+04$

Table 6.6.5: Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the M-P inverse of $\{1,1\}$-QS matrices for Example 6.6.4.

mean	n	$\widetilde{\mathscr{M}}^{\dagger}(M)$	$\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathcal{GV}}))$	$\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{M}}_f^{\dagger}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{C}}^{\dagger}(M)$	$\widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathcal{GV}}))$	$\widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathbb{QS}}))$	$\widetilde{\mathscr{C}}_f^{\dagger}(M(\Psi_{\mathbb{QS}}))$
	100	$4.3964e+02$	$3.6260e+02$	$5.3677e+02$	$3.9861e+02$	$4.5243e+07$	$2.4837e+04$	$3.6640e+04$	$2.7217e+04$
	150	$4.1692e+01$	$3.6846e+01$	$5.2059e+01$	$4.3613e+01$	$1.0494e+09$	$7.5769e+06$	$1.1606e+07$	$9.2937e+06$
	200	$3.3461e+02$	$1.9504e+02$	$2.9550e+02$	$2.4237e+02$	$1.5109e+07$	$1.0192e+05$	$1.9086e+06$	$1.4250e+05$
	250	$2.1300e+02$	$1.3904e+02$	$2.1856e+02$	$1.5019e+02$	$4.8391e+08$	$2.4495e+06$	$4.1373e+06$	$3.1253e+06$
	300	$2.0939e+02$	$9.8454e+01$	$1.6871e+02$	$1.1458e+02$	$1.1827e+09$	$1.6759e+06$	$1.7875e+07$	$1.7776e+06$

Table 6.6.6: Comparison between upper bounds of unstructured, structured CNs and structured effective CNs for the MNLS solution for $\{1,1\}$-QS matrices for Example 6.6.4.

mean	n	$\widetilde{\mathscr{M}}^{\dagger}(M,b)$	$\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathcal{GV}}),b)$	$\widetilde{\mathscr{M}}^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$	$\widetilde{\mathscr{M}}_f^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$	$\widetilde{\mathscr{C}}^{\dagger}(M,b)$	$\widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathcal{GV}}),b)$	$\widetilde{\mathscr{C}}^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$	$\widetilde{\mathscr{C}}_f^{\dagger}(M(\Psi_{\mathbb{QS}}),b)$
	100	$4.3739e+02$	$3.6376e+02$	$2.0884e+02$	$4.0061e+02$	$6.1525e+03$	$4.8648e+02$	$2.5191e+03$	$5.4354e+03$
	150	$4.7556e+01$	$4.3981e+01$	$7.0553e+01$	$5.1641e+01$	$5.3904e+03$	$3.3664e+02$	$1.2982e+03$	$3.8831e+02$
	200	$3.4402e+02$	$2.0275e+02$	$1.0998e+02$	$2.5110e+02$	$1.4342e+04$	$5.3603e+03$	$1.1052e+04$	$6.5164e+03$
	250	$1.0986e+02$	$1.0489e+02$	$1.8630e+02$	$1.1328e+02$	$8.4676e+04$	$5.1797e+03$	$1.0935e+04$	$6.0154e+03$
	300	$2.1839e+02$	$1.0345e+02$	$2.6652e+02$	$1.2097e+02$	$3.4116e+03$	$1.4340e+03$	$6.5971e+03$	$1.7712e+03$

Example 6.6.4. In this example, we consider $\{1,1\}$-QS matrices of different orders. To construct the $\{1,1\}$-QS matrices using the formula given in Definition 6.5.2, we generate the random vectors $t \in \mathbb{R}^{n-2}, u \in \mathbb{R}^{n-1}, v \in \mathbb{R}^{n-1}, w \in \mathbb{R}^{n-1}$ as in Example 6.6.3. Moreover, we generate $\mathbf{d} = \texttt{randn}(n) \in \mathbb{R}^n$ and set $\mathbf{d}(1) = 0$. Further, we rescale v by setting $v(1) = 0$ and $v(n-1) = 1$. For MNLS solution, we choose $b = \texttt{randn}(n) \in \mathbb{R}^N$. For different values of n, the computed upper bounds structured CNs with respect to

QS representation and GV representation through tangent, structured effective CNs, and unstructured CNs are reported in Tables 6.6.5 and 6.6.6. These results confirm that our proposed bounds are reliable for large matrices, and structured ones are much sharper and smaller than unstructured ones.

6.7. Summary

For the M-P inverse and the MNLS solution, we investigated structured MCN and CCN corresponding to a class of parameterized matrices, with each entry as a differentiable function of some real parameters. This framework has been used to derive the upper bounds of structured CNs for CV and $\{1,1\}$-QS matrices. QS representation and the GV representation through tangent are considered for $\{1,1\}$-QS matrices to investigate their structured CNs. It is proved that upper bounds for the structured CNs for GV representation through tangent are always smaller than the QS representation. Numerical examples demonstrate that the proposed structured effective CNs are significantly smaller in most cases.

CHAPTER 7

Conclusions and Scope for Future Work

Conclusion

This thesis focuses on developing efficient iterative methods, preconditioners, structured BEs, and structured CNs for GSPP, DSPP, and LS problems. It addresses key challenges such as slow convergence, scalability, and robustness while incorporating sparsity and structure-preserving perturbations. The following provides an overview of the major contributions discussed in each chapter:

Chapter 1 highlighted key applications of SPPs and provided a review of iterative methods, including Krylov subspace methods, GMRES, and preconditioners. It also presents essential preliminary results to support the subsequent chapters.

In Chapter 2, we proposed the PESS iterative method, corresponding PESS preconditioner, and its relaxed variant LPESS preconditioner to solve the DSPP (2.1.1). For the convergence of the proposed PESS iterative method, necessary and sufficient criteria are derived. Moreover, we estimated the spectral bounds of the proposed PESS and LPESS preconditioned matrices. This empowers us to derive spectral bounds for existing SS and EGSS preconditioned matrices. Extensive experiments validate the effectiveness of the proposed PESS and LPESS preconditioners. Key findings include superior performance over existing preconditioners in terms of IT and CPU time, a significant reduction in the CN of $\mathcal{A}$, and improved spectral clustering compared to baseline preconditioners.

In Chapter 3, we developed the GSS preconditioner and its relaxed variants for solving the DSPP (1.0.6), addressing cases where the diagonal block matrices are both symmetric and nonsymmetric. We analyzed the spectral properties of each preconditioned matrix and demonstrated empirically that our proposed preconditioner outperforms existing state-of-the-art preconditioners, resulting in a well-conditioned system for computing the robust solution of the DSPP.

In Chapter 4, we investigated the structured BEs for circulant, Toeplitz, symmetric-Toeplitz, and Hermitian structured GSPPs with and without preserving the sparsity pattern of block matrices. Moreover, we studied structured BEs for DSPP in three cases when the diagonal block matrices preserve symmetric structure. Additionally, we derived

minimal perturbation matrices for which an approximate solution becomes the exact solution of a nearly perturbed GSPP or DSPP, which preserves their inherent block structure and sparsity pattern. Our obtained results are used to derive structured BE for WRLS problems with Toeplitz or symmetric-Toeplitz coefficient matrices. We have used the obtained structured BEs formulae to show that a backward stable algorithm may not always exhibit strong backward stability for solving the SPPs.

Chapter 5 investigated both unstructured and structured partial NCN, MCN, and CCN for the solution of GSPPs and DSPPs by analyzing structure-preserving perturbations on block matrices. Furthermore, we introduced the concept of partial unified CN for DSPPs, providing a general framework that encompasses traditional NCN, MCN, and CCN. Additionally, we derived compact formulas and specific upper bounds free of Kronecker products. Finally, leveraging our theoretical findings and the connections between the WRLS problem and GSPP, as well as the EILS problem and DSPP, we established their corresponding CNs.

In Chapter 6, we explore structured MCN and CCN for the M-P inverse and MNLS solution of LS problems involving rank-structured matrices, including CV and $\{1,1\}$-QS matrices. A comprehensive framework is developed to establish the upper bounds for the structured CNs of CV and $\{1,1\}$-QS matrices. We consider both QS and GV representations through tangent for $\{1,1\}$-QS matrices to examine their structured CNs. Both the theoretical and numerical results show that the upper bounds for structured CNs in the GV representation are smaller than those in the QS representation.

Future Scope

Building on the findings of this thesis, several promising research directions emerge, raising important questions for further exploration. For instance, can we develop a strongly backward stable algorithm for solving SPP, i.e., the computed solution satisfies a slightly perturbed SPP? Can SS-type preconditioners improve the efficiency of solving DSPPs arising from liquid crystal director modeling? Additionally, how can effective preconditioners for GSPPs with Toeplitz structures enhance numerical performance? Another key area of study involves developing iterative methods and preconditioners for SPPs with multiple right-hand sides, prompting the question: what are the structured BEs and CNs for such problems? Addressing these questions will deepen our understanding and advance computational techniques in this field.

REFERENCES

[1] M. Abdolmaleki, S. Karimi, and D. K. Salkuyeh. A new block-diagonal preconditioner for a class of 3×3 block saddle point problems. *Mediterr. J. Math.*, 19(1): 43, 15, 2022.

[2] S. S. Ahmad and P. Kanhya. Structured perturbation analysis of sparse matrix pencils with s-specified eigenpairs. *Linear Algebra Appl.*, 602:93–119, 2020.

[3] S. S. Ahmad and P. Kanhya. Backward error analysis and inverse eigenvalue problems for Hankel and Symmetric-Toeplitz structures. *Appl. Math. Comput.*, 406: 126288, 15, 2021.

[4] S. S. Ahmad and P. Khatun. Structured condition numbers for a linear function of the solution of the generalized saddle point problem. *Electron. Trans. Numer. Anal.*, 60:471–500, 2024.

[5] S. S. Ahmad and P. Khatun. A robust parameterized enhanced shift-splitting preconditioner for three-by-three block saddle point problems. *Journal of Computational and Applied Mathematics*, 459:116358, 2025.

[6] M. Arioli, M. Baboulin, and S. Gratton. A partial condition number for linear least squares problems. *SIAM J. Matrix Anal. Appl.*, 29(2):413–433, 2007.

[7] M. Baboulin and S. Gratton. A contribution to the conditioning of the total least-squares problem. *SIAM J. Matrix Anal. Appl.*, 32(3):685–699, 2011.

[8] Z.-J. Bai and Z.-Z. Bai. On nonsingularity of block two-by-two matrices. *Linear Algebra Appl.*, 439(8):2388–2404, 2013.

[9] Z.-Z. Bai. Block alternating splitting implicit iteration methods for saddle-point problems from time-harmonic eddy current models. *Numer. Linear Algebra Appl.*, 19(6):914–936, 2012.

[10] Z.-Z. Bai. Regularized HSS iteration methods for stabilized saddle-point problems. *IMA J. Numer. Anal.*, 39(4):1888–1923, 2019.

[11] Z.-Z. Bai and J.-Y. Pan. *Matrix Analysis and Computations*. SIAM, Philadelphia, PA, 2021.

[12] Z.-Z. Bai and Z.-Q. Wang. On parameterized inexact Uzawa methods for generalized saddle point problems. *Linear Algebra Appl.*, 428(11-12):2900–2932, 2008.

[13] Z.-Z. Bai, G. H. Golub, and M. K. Ng. Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems. *SIAM J. Matrix Anal. Appl.*, 24(3):603–626, 2003.

[14] Z.-Z. Bai, G. H. Golub, and J.-Y. Pan. Preconditioned Hermitian and skew-Hermitian splitting methods for non-Hermitian positive semidefinite linear systems. *Numer. Math.*, 98(1):1–32, 2004.

[15] Z.-Z. Bai, B. N. Parlett, and Z.-Q. Wang. On generalized successive overrelaxation methods for augmented linear systems. *Numer. Math.*, 102(1):1–38, 2005.

[16] Z.-Z. Bai, J.-F. Yin, and Y.-F. Su. A shift-splitting preconditioner for non-Hermitian positive definite matrices. *J. Comput. Math.*, 24(4):539–552, 2006.

[17] F. B. Balani and M. Hajarian. A new block preconditioner for weighted Toeplitz regularized least-squares problems. *Int. J. Comput. Math.*, 100(12):2241–2250, 2023.

[18] F. B. Balani, M. Hajarian, and L. Bergamaschi. Two block preconditioners for a class of double saddle point linear systems. *Appl. Numer. Math.*, 190:155–167, 2023.

[19] F. Balani Bakrani, L. Bergamaschi, Á. Mart´*i*nez, and M. Hajarian. Some preconditioning techniques for a class of double saddle point problems. *Numer. Linear Algebra Appl.*, page e2551, 2024.

[20] F. P. Beik, C. Greif, and M. Trummer. On the invertibility of matrices with a double saddle-point structure. *Linear Algebra and its Applications*, 699:403–420, 2024.

[21] F. P. A. Beik and M. Benzi. Iterative methods for double saddle point systems. *SIAM J. Matrix Anal. Appl.*, 39(2):902–921, 2018.

[22] A. Ben-Israel. On error bounds for generalized inverses. *SIAM J. Numer. Anal.*, 3: 585–592, 1966.

[23] M. Benzi and X.-P. Guo. A dimensional split preconditioner for Stokes and linearized Navier-Stokes equations. *Appl. Numer. Math.*, 61(1):66–76, 2011.

[24] M. Benzi and M. K. Ng. Preconditioned iterative methods for weighted Toeplitz least squares problems. *SIAM J. Matrix Anal. Appl.*, 27(4):1106–1124, 2006.

[25] M. Benzi and V. Simoncini. On the eigenvalues of a class of saddle point matrices. *Numer. Math.*, 103(2):173–196, 2006.

[26] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numer.*, 14:1–137, 2005.

[27] M. Benzi, M. Ng, Q. Niu, and Z. Wang. A relaxed dimensional factorization preconditioner for the incompressible Navier-Stokes equations. *J. Comput. Phys.*, 230 (16):6185–6202, 2011.

[28] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific. Athena Scientific, Belmont, MA, second edition, 1999.

[29] A. K. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, PA, 1996.

[30] A. Bojanczyk, N. J. Higham, and H. Patel. The equality constrained indefinite least squares problem: theory and algorithms. *BIT*, 43(3):505–517, 2003.

[31] A. W. Bojanczyk. Algorithms for indefinite linear least squares problems. *Linear Algebra Appl.*, 623:104–127, 2021.

[32] M. Bolten, M. Donatelli, P. Ferrari, and I. Furci. Symbol based convergence analysis in multigrid methods for saddle point problems. *Linear Algebra Appl.*, 671:67–108, 2023.

[33] S. Bradley and C. Greif. Eigenvalue bounds for double saddle-point systems. *IMA J. Numer. Anal.*, 43(6):3564–3592, 2023.

[34] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*, volume 15. Springer-Verlag, New York, 1991.

[35] J. R. Bunch. The weak and strong stability of algorithms in numerical linear algebra. *Linear Algebra Appl.*, 88/89:49–66, 1987.

[36] J. R. Bunch, J. W. Demmel, and C. F. Van Loan. The strong stability of algorithms for solving symmetric linear systems. *SIAM J. Matrix Anal. Appl.*, 10(4):494–499, 1989.

[37] Y. Cao. Shift-splitting preconditioners for a class of block three-by-three saddle point problems. *Appl. Math. Lett.*, 96:40–46, 2019.

[38] Y. Cao and L. Petzold. A subspace error estimate for linear systems. *SIAM J. Matrix Anal. Appl.*, 24(3):787–801, 2003.

[39] Y. Cao, M.-Q. Jiang, and Y.-L. Zheng. A splitting preconditioner for saddle point problems. *Numer. Linear Algebra Appl.*, 18(5):875–895, 2011.

[40] Y. Cao, J. Du, and Q. Niu. Shift-splitting preconditioners for saddle point problems. *J. Comput. Appl. Math.*, 272:239–250, 2014.

[41] Y. Cao, S.-X. Miao, and Z.-R. Ren. On preconditioned generalized shift-splitting iteration methods for saddle point problems. *Comput. Math. Appl.*, 74(4):859–872, 2017.

[42] S. Chandrasekaran, P. Dewilde, M. Gu, T. Pals, and A. J. van der Veen. Fast stable solver for sequentially semi-separable linear systems of equations. In *International Conference on High-Performance Computing*, pages 545–554, Springer, 2002.

[43] F. Chen and B.-C. Ren. A modified alternating positive semidefinite splitting preconditioner for block three-by-three saddle point problems. *Electron. Trans. Numer. Anal.*, 58:84–100, 2023.

[44] X. S. Chen, W. Li, X. Chen, and J. Liu. Structured backward errors for generalized saddle point systems. *Linear Algebra Appl.*, 436(9):3109–3119, 2012.

[45] Z. Chen, Q. Du, and J. Zou. Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients. *SIAM J. Numer. Anal.*, 37(5):1542–1570, 2000.

[46] W. Cheney. *Analysis for Applied Mathematics*, volume 208 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2001.

[47] S. Chountasis, V. N. Katsikis, and D. Pappas. Applications of the Moore-Penrose inverse in digital image restoration. *Math. Probl. Eng.*, page 12, 2009.

[48] S. Chountasis, V. N. Katsikis, and D. Pappas. Digital image reconstruction in the spectral domain utilizing the Moore-Penrose inverse. *Math. Probl. Eng.*, page 14, 2010.

[49] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*, volume 93 of *Appl. Math. Sci.* Springer-Verlag, Berlin, Second edition, 1998.

[50] F. Cucker and H. Diao. Mixed and componentwise condition numbers for rectangular structured matrices. *Calcolo*, 44(2):89–115, 2007.

[51] F. Cucker, H. Diao, and Y. Wei. On mixed and componentwise condition numbers for Moore-Penrose inverse and linear least squares problems. *Math. Comp.*, 76(258): 947–963, 2007.

[52] H. Diao and Q. Meng. Structured generalized eigenvalue condition numbers for parameterized quasiseparable matrices. *BIT*, 59(3):695–720, 2019.

[53] H. Diao and Y. Wei. On Frobenius normwise condition numbers for Moore-Penrose inverse and linear least-squares problems. *Numer. Linear Algebra Appl.*, 14(8):603–610, 2007.

[54] H. Diao, L. Li, and Q. Meng. Structured condition numbers for sylvester matrix equation with parameterized quasiseparable matrices. *Comm. Anal. Comp.*, 1(3): 183–213, 2023.

[55] H.-A. Diao, L. Liang, and S. Qiao. A condition analysis of the weighted linear least squares problem using dual norms. *Linear and Multilinear Algebra*, 66(6):1085–1103, 2018.

[56] F. M. Dopico and K. Pomés. Structured eigenvalue condition numbers for parameterized quasiseparable matrices. *Numer. Math.*, 134(3):473–512, 2016.

[57] F. M. Dopico and K. Pomés. Structured condition numbers for linear systems with parameterized quasiseparable coefficient matrices. *Numer. Algorithms*, 73(4):1131–1158, 2016.

[58] Y. Eidelman and I. Gohberg. On a new class of structured matrices. *Integral Equations Operator Theory*, 34(3):293–324, 1999.

[59] H. C. Elman, A. Ramage, and D. J. Silvester. Algorithm 886: IFISS, a Matlab toolbox for modelling incompressible flow. *ACM Trans. Math. Software*, 33(2):Art. 14, 18, 2007.

[60] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. Oxford, second edition, 2014.

[61] H. Fan, Y. Li, H. Zhang, and X. Zhu. Preconditioners based on matrix splitting for the structured systems from elliptic PDE-constrained optimization problems. *Appl. Math. Comput.*, 463:Paper No. 128341, 8, 2024.

[62] J. Fessler and S. Booth. Conjugate-gradient preconditioning methods for shift-variant pet image reconstruction. *IEEE Transactions on Image Processing*, 8(5): 688–699, 1999.

[63] I. Gohberg and I. Koltracht. Mixed, componentwise, and structured condition numbers. *SIAM J. Matrix Anal. Appl.*, 14(3):688–704, 1993.

[64] G. H. Golub, X. Wu, and J.-Y. Yuan. SOR-like methods for augmented systems. *BIT*, 41(1):71–85, 2001.

[65] N. I. M. Gould, M. E. Hribar, and J. Nocedal. On the solution of equality constrained quadratic programming problems arising in optimization. *SIAM J. Sci. Comput.*, 23(4):1376–1395, 2001.

[66] A. Graham. *Kronecker Products and Matrix Calculus: with Applications*. Wiley, New York, 1981.

[67] S. Gratton. On the condition number of linear least squares problems in a weighted Frobenius norm. *BIT*, 36(3):523–530, 1996.

[68] L. Greengard and V. Rokhlin. On the numerical solution of two-point boundary value problems. *Comm. Pure Appl. Math.*, 44(4):419–452, 1991.

[69] C. Greif and Y. He. Block preconditioners for the marker-and-cell discretization of the Stokes-Darcy equations. *SIAM J. Matrix Anal. Appl.*, 44(4):1540–1565, 2023.

[70] L. Grigori, Q. Niu, and Y. Xu. Stabilized dimensional factorization preconditioner for solving incompressible Navier-Stokes equations. *Appl. Numer. Math.*, 146:309–327, 2019.

[71] D. Han and X. Yuan. Local linear convergence of the alternating direction method of multipliers for quadratic programs. *SIAM J. Numer. Anal.*, 51(6):3446–3457, 2013.

[72] D. J. Higham and N. J. Higham. Backward error and condition of structured linear systems. *SIAM J. Matrix Anal. Appl.*, 13(1):162–175, 1992.

[73] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, second edition, 2002.

[74] N. Huang. Variable parameter Uzawa method for solving a class of block three-by-three saddle point problems. *Numer. Algorithms*, 85(4):1233–1254, 2020.

[75] N. Huang and C.-F. Ma. Spectral analysis of the preconditioned system for the 3×3 block saddle point problem. *Numer. Algorithms*, 81(2):421–444, 2019.

[76] N. Huang, Y.-H. Dai, and Q. Hu. Uzawa methods for a class of block three-by-three saddle-point problems. *Numer. Linear Algebra Appl.*, 26(6):e2265, 26, 2019.

[77] N. Huang, Y.-H. Dai, D. Orban, and M. A. Saunders. On GSOR, the generalized successive overrelaxation method for double saddle-point problems. *SIAM J. Sci. Comput.*, 45(5):A2185–A2206, 2023. ISSN 1064-8275,1095-7197.

[78] R. Huang. Accurate solutions of weighted least squares problems associated with rank-structured matrices. *Appl. Numer. Math.*, 146:416–435, 2019.

[79] Y.-M. Huang. A practical formula for computing optimal parameters in the HSS iteration methods. *J. Comput. Appl. Math.*, 255:142–149, 2014.

[80] Z.-G. Huang, L.-G. Wang, Z. Xu, and J.-J. Cui. Parameterized generalized shift-splitting preconditioners for nonsymmetric saddle point problems. *Comput. Math. Appl.*, 75(2):349–373, 2018.

[81] L. A. Imhof and W. J. Studden. E-optimal designs for rational models. *Ann. Statist.*, 29(3):763–783, 2001.

[82] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., 1989.

[83] Y.-F. Ke and C.-F. Ma. Some preconditioners for elliptic PDE-constrained optimization problems. *Comput. Math. Appl.*, 75(8):2795–2813, 2018.

[84] A. N. Langville and W. J. Stewart. The Kronecker product and stochastic automata networks. *J. Comput. Appl. Math.*, 167(2):429–447, 2004.

[85] J. Y. Lee and L. Greengard. A fast adaptive numerical method for stiff two-point boundary value problems. *SIAM J. Sci. Comput.*, 18(2):403–429, 1997.

[86] B. Li and Z. Jia. Some results on condition numbers of the scaled total least squares problem. *Linear Algebra Appl.*, 435(3):674–686, 2011.

[87] H. Li and S. Wang. On the partial condition numbers for the indefinite least squares problem. *Appl. Numer. Math.*, 123:200–220, 2018.

[88] H. Li, S. Wang, and H. Yang. On mixed and componentwise condition numbers for indefinite least squares problem. *Linear Algebra Appl.*, 448:104–129, 2014.

[89] J. Li and Z. Li. A modified preconditioner for three-by-three block saddle point problems. *Jpn. J. Ind. Appl. Math.*, 41(1):659–680, 2024.

[90] X. Li and X. Liu. Structured backward errors for structured KKT systems. *J. Comput. Math.*, 22(4):605–610, 2004.

[91] Z. Li, Q. Xu, and Y. Wei. A note on stable perturbations of Moore-Penrose inverses. *Numer. Linear Algebra Appl.*, 20(1):18–26, 2013.

[92] Z.-Z. Liang and M.-Z. Zhu. On the improvement of shift-splitting preconditioners for double saddle point problems. *J. Appl. Math. Comput.*, 70(2):1339–1363, 2024.

[93] Q. Liu and M. Wang. Algebraic properties and perturbation results for the indefinite least squares problem with equality constraints. *Int. J. Comput. Math.*, 87(1-3):425–434, 2010.

[94] Q. Liu, Z. Jia, and Y. Wei. Multidimensional total least squares problem with linear equality constraints. *SIAM J. Matrix Anal. Appl.*, 43(1):124–150, 2022.

[95] P. Lv. Structured backward errors analysis for generalized saddle point problems arising from the incompressible Navier-Stokes equations. *AIMS Math.*, 8(12):30501–30510, 2023.

[96] P. Lv and B. Zheng. Structured backward error analysis for a class of block three-by-three saddle point problems. *Numer. Algorithms*, 90(1):59–78, 2022.

[97] W. Ma. On normwise structured backward errors for the generalized saddle point systems. *Calcolo*, 54(2):503–514, 2017.

[98] W. Ma, Y. Fan, and X. Xu. Structured backward errors for block three-by-three saddle point systems. *Linear and Multilinear Algebra*, pages 1–21, 2024.

[99] A. N. Malyshev. A unified theory of conditioning for linear least squares and Tikhonov regularization solutions. *SIAM J. Matrix Anal. Appl.*, 24(4):1186–1196, 2003.

[100] L. Meng and J. Li. Condition numbers of generalized saddle point systems. *Calcolo*, 56(2):Article 18, 2019.

[101] L. Meng and B. Zheng. Structured condition numbers for the Tikhonov regularization of discrete ill-posed problems. *J. Comput. Math.*, 35(2):159–186, 2017.

[102] L. Meng, Y. He, and S.-X. Miao. Structured backward errors for two kinds of generalized saddle point systems. *Linear Multilinear Algebra*, 70(7):1345–1355, 2022.

[103] Q. Meng, H. Diao, and Q. Yu. Structured condition number for multiple right-hand side linear systems with parameterized quasiseparable coefficient matrix. *J. Comput. Appl. Math.*, 368:112527, 20, 2020.

[104] H. Mirchi and D. K. Salkuyeh. A new preconditioner for elliptic PDE-constrained optimization problems. *Numer. Algorithms*, 83(2):653–668, 2020.

[105] G. Mühlbach. Interpolation by Cauchy-Vandermonde systems and applications. *J. Comput. Appl. Math.*, 122(1-2):203–222, 2000.

[106] M. K. Ng and J. Pan. Weighted Toeplitz regularized least squares computation for image restoration. *SIAM J. Sci. Comput.*, 36(1):B94–B121, 2014.

[107] J. W. Pearson and A. J. Wathen. A new approximation of the Schur complement in preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 19(5):816–829, 2012.

[108] J. W. Pearson, M. Stoll, and A. J. Wathen. Regularization-robust preconditioners for time-dependent PDE-constrained optimization problems. *SIAM J. Matrix Anal. Appl.*, 33(4):1126–1152, 2012.

[109] J. W. Pearson, M. Stoll, and A. J. Wathen. Preconditioners for state-constrained optimal control problems with Moreau-Yosida penalty function. *Numer. Linear Algebra Appl.*, 21(1):81–97, 2014.

[110] I. Perugia and V. Simoncini. Block-diagonal and indefinite symmetric preconditioners for mixed finite element formulations. *Numer. Linear Algebra Appl.*, 7(7-8): 585–616, 2000.

[111] A. Ramage and E. C. Gartland, Jr. A preconditioned nullspace method for liquid crystal director modeling. *SIAM J. Sci. Comput.*, 35(1):B226–B247, 2013.

[112] C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications*. John Wiley & Sons, Inc., New York-London-Sydney, 1971.

[113] T. Rees. Github - tyronerees/poisson-control. Accessed: 2022-3-21, 2010. https://github.com/tyronerees/poisson-control.

[114] T. Rees and M. Stoll. Block-triangular preconditioners for PDE-constrained optimization. *Numer. Linear Algebra Appl.*, 17(6):977–996, 2010.

[115] T. Rees, H. S. Dollar, and A. J. Wathen. Optimal solvers for PDE-constrained optimization. *SIAM J. Sci. Comput.*, 32(1):271–298, 2010.

[116] J. R. Rice. A theory of condition. *SIAM J. Numer. Anal.*, 3:287–310, 1966.

[117] J. L. Rigal and J. Gaches. On the compatibility of a given solution with the data of a linear system. *J. Assoc. Comput. Mach.*, 14:543–548, 1967.

[118] J. Rohn. New condition numbers for matrices and linear systems. *Computing*, 41 (1-2):167–169, 1989.

[119] M. Rozložn´*i*k and V. Simoncini. Krylov subspace methods for saddle point problems with indefinite preconditioning. *SIAM J. Matrix Anal. Appl.*, 24(2):368–391, 2002.

[120] S. M. Rump. Structured perturbations. I. Normwise distances. *SIAM J. Matrix Anal. Appl.*, 25(1):1–30, 2003.

[121] S. M. Rump. Structured perturbations. II. Componentwise distances. *SIAM J. Matrix Anal. Appl.*, 25(1):31–56, 2003.

[122] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, second edition, 2003.

[123] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[124] D. K. Salkuyeh, M. Masoudi, and D. Hezari. On the generalized shift-splitting preconditioner for saddle point problems. *Appl. Math. Lett.*, 48:55–61, 2015.

[125] D. K. Salkuyeh, H. Aslani, and Z.-Z. Liang. An alternating positive semidefinite splitting preconditioner for the three-by-three block saddle point problems. *Math. Commun.*, 26(2):177–195, 2021.

[126] V. Sarin and A. Sameh. An efficient iterative method for the generalized Stokes problem. *SIAM J. Sci. Comput.*, 19(1):206–226, 1998.

[127] R. D. Skeel. Scaling for numerical stability in Gaussian elimination. *J. Assoc. Comput. Mach.*, 26(3):494–526, 1979.

[128] G. W. Stewart and J. G. Sun. *Matrix Perturbation Theory*. Academic Press, New York, 1990.

[129] J.-G. Sun. Structured backward errors for KKT systems. *Linear Algebra Appl.*, 288 (1-3):75–88, 1999.

[130] R. Vandebril, M. Van Barel, and N. Mastronardi. A note on the representation and definition of semiseparable matrices. *Numer. Linear Algebra Appl.*, 12(8):839–858, 2005.

[131] R. Vandebril, M. Van Barel, and N. Mastronardi. *Matrix Computations and Semiseparable Matrices*, volume 1. Johns Hopkins University Press, Baltimore, MD, 2008.

[132] R. Vandebril, M. Van Barel, and N. Mastronardi. *Matrix Computations and Semiseparable Matrices*, volume 2. Johns Hopkins University Press, Baltimore, MD, 2008.

[133] G. Wang, Y. Wei, and S. Qiao. *Generalized Inverses: Theory and Computations*. Springer, Singapore, second edition, 2018.

[134] L. Wang and K. Zhang. Generalized shift-splitting preconditioner for saddle point problems with block three-by-three structure. *Open Access Library Journal*, 6(12): 1–14, 2019.

[135] N.-N. Wang and J.-C. Li. A class of new extended shift-splitting preconditioners for saddle point problems. *J. Comput. Appl. Math.*, 357:123–145, 2019.

[136] N.-N. Wang and J.-C. Li. On parameterized block symmetric positive definite preconditioners for a class of block three-by-three saddle point problems. *J. Comput. Appl. Math.*, 405:113959, 2022.

[137] S. Wang and L. Meng. A contribution to the conditioning theory of the indefinite least squares problems. *Appl. Numer. Math.*, 177:137–159, 2022.

[138] X.-F. Wang and X.-G. Liu. Comparing condition numbers with structured condition numbers for KKT systems. *Math. Numer. Sin.*, 28(2):211–223, 2006.

[139] A. J. Wathen. Preconditioning. *Acta Numerica*, 24:329–376, 2015.

[140] Y. Wei and D. Wang. Condition numbers and perturbation of the weighted Moore-Penrose inverse and weighted linear least squares problem. *Appl. Math. Comput.*, 145(1):45–58, 2003.

[141] Y. Wei, W. Xu, S. Qiao, and H. Diao. Componentwise condition numbers for generalized matrix inversion and linear least squares. *Numer. Math. J. Chinese Univ.*, 14(3):277–286, 2005.

[142] Y. Wei, H. Diao, and S. Qiao. Condition number for weighted linear least squares problem. *J. Comput. Math.*, 25(5):561–572, 2007.

[143] J. H. Wilkinson. *The algebraic eigenvalue problem*. Clarendon Press, Oxford, 1965.

[144] S.-L. Wu and D. K. Salkuyeh. A shift-splitting preconditioner for asymmetric saddle point problems. *Comput. Appl. Math.*, 39(4):Paper No. 314, 17, 2020.

[145] X. Wu. *Structured condition numbers for least squares problems with parameterized quasiseparable coefficient matrices*. Northeast Normal University, Master thesis, under the supervision of H. Diao, 2020.

[146] H. Xiang and Y. Wei. On normwise structured backward errors for saddle point sytems. *SIAM J. Matrix Anal. Appl.*, 29(3):838–849, 2007.

[147] H. Xiang, Y. Wei, and H. Diao. Perturbation analysis of generalized saddle point systems. *Linear Algebra Appl.*, 419(1):8–23, 2006.

[148] X. Xie and H.-B. Li. A note on preconditioning for the 3×3 block saddle point problem. *Comput. Math. Appl.*, 79(12):3289–3296, 2020.

[149] Z.-J. Xie, W. Li, and X.-Q. Jin. On condition numbers for the canonical generalized polar decomposition of real matrices. *Electron. J. Linear Algebra*, 26:842–857, 2013.

[150] X. Xiong and J. Li. A simplified relaxed alternating positive semi-definite splitting preconditioner for saddle point problems with three-by-three block structure. *J. Appl. Math. Comput.*, 69(3):2295–2313, 2023.

[151] W. Xu and W. Li. New perturbation analysis for generalized saddle point systems. *Calcolo*, 46(1):25–36, 2009.

[152] W. Xu, Y. Wei, and S. Qiao. Condition numbers for structured least squares problems. *BIT*, 46(1):203–225, 2006.

[153] W.-W. Xu, M.-M. Liu, L. Zhu, and H.-F. Zuo. New perturbation bounds analysis of a kind of generalized saddle point systems. *East Asian J. Appl. Math.*, 7(1): 116–124, 2017.

[154] A.-L. Yang, J.-L. Zhu, and W. Yu-Jiang. Multi-parameter dimensional split preconditioner for three-by-three block system of linear equations. *Numer. Algorithms*, 95 (2):721–745, 2024.

[155] Z. Yang. Accurate computations of eigenvalues of quasi-Cauchy-Vandermonde matrices. *Linear Algebra Appl.*, 622:268–293, 2021.

[156] L. Yin, Y. Huang, and Q. Tang. Extensive generalized shift-splitting preconditioner for 3×3 block saddle point problems. *Appl. Math. Lett.*, 143:108668, 7, 2023.

[157] M. L. Zeng. A circulant-matrix-based new accelerated GSOR preconditioned method for block two-by-two linear systems from image restoration problems. *Appl. Numer. Math.*, 164:245–257, 2021.

[158] G.-F. Zhang, J.-L. Yang, and S.-S. Wang. On generalized parameterized inexact Uzawa method for a block two-by-two linear system. *J. Comput. Appl. Math.*, 255: 193–207, 2014.

[159] K. Zhang and Y. Su. Structured backward error analysis for sparse polynomial eigenvalue problems. *Appl. Math. Comput.*, 219(6):3073–3082, 2012.

[160] N. Zhang, R.-X. Li, and J. Li. Lopsided shift-splitting preconditioner for saddle point problems with three-by-three structure. *Comput. Appl. Math.*, 41(6):Paper No. 261, 16, 2022.

[161] X. Zhang and Y. Huang. On block preconditioners for PDE-constrained optimization problems. *J. Comput. Math.*, 32(3):272–283, 2014.

[162] B. Zheng and P. Lv. Structured backward error analysis for generalized saddle point problems. *Adv. Comput. Math.*, 46(2):34–27, 2020.

[163] M. Z. Zhu, Y. E. Qi, and G. F. Zhang. On local circulant and residue splitting iterative method for Toeplitz-structured saddle point problems. *IAENG Int. J. Appl. Math.*, 48(2):221–227, 2018.