FUNCTIONS OF STRUCTURED MATRICES

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

September 2016

Bahar Arslan
School of Mathematics
5.4.3 Numerical tests for the structured sign decomposition 120

6 Perturbation Analysis for the Canonical Generalized Polar Decomposition 123

6.1 Introduction 123
6.2 The canonical generalized polar decomposition 125
6.3 Additive perturbation analysis 127
  6.3.1 Real case 127
  6.3.2 Complex case 129
6.4 Multiplicative perturbation analysis 132
  6.4.1 Left multiplicative perturbation 132
  6.4.2 Right multiplicative perturbation 134
6.5 Using standard techniques 135
6.6 Numerical experiments 138

7 Conclusions 143

Bibliography 146
## List of Tables

1.1 Examples of structured matrices associated with the scalar product $\langle \cdot, \cdot \rangle_M$. 28

2.1 The structure of the matrices in the automorphism group and Lie algebra. 40

2.2 The cost analysis of taking a square root of $T$ given in the Schur-like form [35, Table C.1]. ................................................................. 46

2.3 The cost analysis of structure preserving iterations for taking a square root of a full symplectic matrix $A \in \mathbb{R}^{2n \times 2n}$. ......................................................... 48

2.4 The cost of the multiplication $Z_kY_k$ when $Z_k \in \mathbb{R}^{2n \times 2n}$ and $Y_k \in \mathbb{R}^{2n \times 2n}$ are in the Schur-like form. ......................................................... 49

2.5 The cost analysis of iteration (2.10) when $Y_k \in \mathbb{R}^{2n \times 2n}$ and $Z_k \in \mathbb{R}^{2n \times 2n}$ are in the Schur-like form. ......................................................... 49

2.6 Maximal values $\theta_m$ of $\|A-I\|$ ensures that the bound $\|r_m(A-I)-\log A\|$ does not exceed $u = 2^{-53}$ [35, Table 11.1]. ......................................................... 51

2.7 The analysis of Algorithm 11.9 and Algorithm 11.10 from [35, Alg. 11.9, Alg. 11.10] in terms of structure preservation and the accuracy. ........ 59

2.8 The logarithm of the symplectic matrices is computed by Algorithm 2.17 and the square root is evaluated by Algorithm 2.11. ......................... 60

2.9 The logarithm of the symplectic matrices is computed by using Algorithm 2.18 and the square root is evaluated by iteration (2.10). .............. 61

2.10 The logarithm of the symplectic matrices given in the Schur-like form is computed by Algorithm 2.19 and the square root is evaluated by iteration (2.10). ......................................................... 61
List of Figures

1.1 The columns of $K_f(A)$ are formed by evaluating the Fréchet derivative of $f$ at $A$. .......................................................... 24

1.2 The differential of $f$ is a map between tangent spaces $df(A) : T_A\mathcal{M} \to T_{f(A)}\mathcal{N}$. .......................................................... 33

2.1 Comparison of the algorithms ............................................... 62

4.1 Restriction the perturbation to a smaller set. ......................... 81

4.2 Cumulative distribution function for $\rho_{\log}$, empirically estimated for random $10 \times 10$ symplectic matrices. ............................ 99

4.3 Cumulative distribution function for $\rho_{1/2}$, empirically estimated for random $10 \times 10$ symplectic matrices. ............................ 100

4.4 Cumulative distribution function for $\rho_{\log}$ (on the left) and $\rho_{1/2}$ (on the right), empirically estimated for random $10 \times 10$ symplectic–orthogonal matrices ($\kappa_2(X) = 1$). ........................................ 101

4.5 Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal logarithm of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. ................................. 102

4.6 Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. ................................. 103
4.7 Error bounds on the relative error \( \| f(X + E) - f(X) \|_F / \| f(X) \|_F \) for the logarithm (left plot) and the square root (right plot) of 50 randomly generated 12 \times 12 \) symplectic matrices \( X \) with \( \kappa_2(X) = 10^2 \). . . . . . . 104

4.8 Error bounds on the relative error \( \| f(X + E) - f(X) \|_F / \| f(X) \|_F \) for the logarithm (left plot) and the square root (right plot) of 50 randomly generated 12 \times 12 \) symplectic matrices \( X \) with \( \kappa_2(X) = 10^9 \). . . . . . . 104

5.1 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( W \) in plot (a) and the factor \( S \) in plot (b) of the generalized polar decomposition of randomly generated pseudo-orthogonal matrices \( A \in \mathbb{R}^{10 \times 10} \) \( (A \in \mathbb{G}_M) \) with increasing condition number \( \kappa_2(A) \). . . . . . . . . . . . . . . . . . . . 116

5.2 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( W \) in plot (a) and the factor \( S \) in plot (b) of the generalized polar decomposition of randomly generated Hamiltonian matrices \( A \in \mathbb{R}^{10 \times 10} \) \( (A \in \mathbb{L}_M) \) with \( \kappa_2(A) \) changing between 10 to 10\(^5\). . . . . . . . . . . . . . . . . . . . 116

5.3 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( W \) in plot (a) and the factor \( S \) in plot (b) of the generalized polar decomposition of randomly generated skew-Hamiltonian matrices \( A \in \mathbb{R}^{10 \times 10} \) \( (A \in \mathbb{J}_M) \) with \( \kappa_2(A) \) changing between 10 to 10\(^5\). . . . . . . . . . . . . . . . . . . . 117

5.4 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( U \) in plot (a) and the factor \( H \) in plot (b) of the structured polar decomposition of randomly generated symplectic matrices \( A \in \mathbb{R}^{10 \times 10} \) \( (A \in \mathbb{G}_M) \) with increasing condition number \( \kappa_2(A) \). . . . . . . . . . . . . . . . . . . . 118
5.5 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( U \) in plot (a) and the factor \( H \) in plot (b) of the structured polar decomposition of randomly generated pseudo-orthogonal matrices \( A \in \mathbb{R}^{10 \times 10} \) (\( A \in \mathbb{G}_M \)) with increasing condition number \( \kappa_2(A) \). ........................................ 118

5.6 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the structured polar decomposition of randomly generated skew-symmetric and Hamiltonian matrices \( A \in \mathbb{L}_M \), representing the condition number of the factor \( U \) in plot (a) and (b). .................................................. 119

5.7 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the structured polar decomposition of randomly generated symmetric and skew-Hamiltonian matrices \( A \in \mathbb{J}_M \), representing the condition number of the factor \( U \) in plot (a) and (b). .................................................. 119

5.8 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( S \) in plot (a) and the factor \( N \) in plot (b) of the sign factorization of randomly generated symplectic matrices \( A \in \mathbb{R}^{10 \times 10} \) (\( A \in \mathbb{G}_M \)) with increasing condition number \( \kappa_2(A) \). ........................................ 120

5.9 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( S \) in plot (a) and the factor \( N \) in plot (b) of the sign factorization of randomly generated perplectic matrices \( A \in \mathbb{R}^{10 \times 10} \) (\( A \in \mathbb{G}_M \)) with increasing condition number \( \kappa_2(A) \). ........................................ 121

5.10 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor \( S \) in plot (a) and the factor \( N \) in plot (b) of the structured sign factorization of randomly generated Hamiltonian matrices \( A \in \mathbb{R}^{10 \times 10} \) (\( A \in \mathbb{L}_M \)) with \( \kappa_2(A) \) changing between 10 to \( 10^5 \). ........................................ 121
5.11 Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $S$ in plot (a) and the factor $N$ in plot (b) of the structured sign factorization of randomly generated skew-Hamiltonian matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in J_M$) with $\kappa_2(A)$ changing between 10 to $10^5$.

6.1 The bounds are given for random $A \in \mathbb{R}^{12 \times 8}$ matrices with $\kappa_2(A) = 10^6$. 139

6.2 The bounds are given for $A \in \mathbb{R}^{10 \times 8}$ with $M = I_{10}$ and $N = I_8$. 140

6.3 The bounds are given for $A \in \mathbb{R}^{12 \times 8}$ with $\kappa_2(A) = 10^6$, $M = R_{12} \in \mathbb{R}^{12 \times 12}$ and $N = R_8 \in \mathbb{R}^{8 \times 8}$. 141

6.4 The bounds are given for $A \in \mathbb{R}^{12 \times 8}$ with $\kappa_2(A) = 10^5$, $M = \Sigma_{8,4}$ and $N = \Sigma_{6,2}$. 142
The growing interest in computing structured matrix functions stems from the fact that preserving and exploiting the structure of matrices can help us gain physically meaningful solutions with less computational cost and memory requirement. The work presented here is divided into two parts. The first part deals with the computation of functions of structured matrices. The second part is concerned with the structured error analysis in the computation of matrix functions.

We present algorithms applying the inverse scaling and squaring method and using the Schur-like form of the symplectic matrices as an alternative to the algorithms using the Schur decomposition to compute the logarithm of symplectic matrices. There are two main calculations in the inverse scaling and squaring method: taking a square root and evaluating the Padé approximants. Numerical experiments suggest that using the Schur-like form with the structure preserving iterations for the square root helps us to exploit the Hamiltonian structure of the logarithm of symplectic matrices.

Some type of matrices are nearly structured. We discuss the conditions for using the nearest structured matrix to the nearly structured one by analysing the forward error bounds. Since the structure preserving algorithms for computing the functions of matrices provide advantages in terms of accuracy and data storage we suggest to compute the function of the nearest structured matrix. The analysis is applied to the nearly unitary, nearly Hermitian and nearly positive semi-definite matrices for the matrix logarithm, square root, exponential, cosine and sine functions.

It is significant to investigate the effect of the structured perturbations in the sensitivity analysis of matrix functions. We study the structured condition number of matrix functions defined between smooth square matrix manifolds. We develop algorithms computing and estimating the structured condition number. We also present the lower and upper bounds on the structured condition number, which are cheaper to compute than the “exact” structured condition number. We observe that the lower bounds give a good estimation for the structured condition numbers. Comparing the structured and unstructured condition number reveals that they can differ by several orders of magnitude.

Having discussed how to compute the structured condition number of matrix functions defined between smooth square matrix manifolds we apply the theory of structured condition numbers to the structured matrix factorizations. We measure the sensitivity of matrix factors to the structured perturbations for the structured polar decomposition, structured sign factorization and the generalized polar decomposition.

Finally, we consider the unstructured perturbation analysis for the canonical generalized polar decomposition by using three different methods. Apart from theoretical aspect of the perturbation analysis, perturbation bounds obtained from these methods are compared numerically and our findings show an improvement on the sharpness of the perturbation bounds in the literature.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy.
(see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s Policy on Presentation of Theses.
Acknowledgements

It is a great honour to be a part of Numerical Linear Algebra group in the University of Manchester. Despite tough times in my PhD I had enormous support from lots of people during that period.

First, I would like to thank my supervisors Prof. Nicholas Higham and Prof. Françoise Tisseur for their assistance and endless support in my research. I have learned a lot of things from Prof. Higham’s feedback and his work discipline. I would like to express my deepest gratitude to Prof. Tisseur for her willingness to help me whenever I needed and for the discussions we had about my work. Without their supervision and encouragement this dissertation would not have been possible.

I am also indebted to Vanni Noferini for the excellent cooperation and his guidance. I own my deepest appreciation to him for coming to my talk and suggesting brilliant ideas for the open problem that I have proposed.

I would like to thank my office mates Nataša, Sam and Mary for the interesting discussions in maths and for their help in any problems I faced. I already miss our coffee break chats with Nataša. I am also thankful to Sophia for her encouragement in my PhD and for her nice attempts to make me “Mancunian”. Consequently, I love rain!

I should thank to Dr. James Montaldi for giving his time to answer my questions about manifolds.

Beyond research to my friends who supported me emotionally, I will never forget nice moments that I have shared with Burcin, Gamze, Vildan and my other friends that I could not list here. I would also like to thank my friend Javier for being such a nice flatmate during my PhD.
For the financial support I would also like to thank the Republic of Turkey Ministry of Education office and School of Mathematics in the University of Manchester.

Special thanks go to Mesut for his belief, love and encouragement. Your letters helped me out a lot in difficult times. Can I please get one more letter with a topic “Open when you become a doctor”?

Finally my beloved family, I would like to thank my parents and my brother for their unconditional love and supporting me no matter what happens. Thanks my dear brother for always cheering me up in my stressful times. You are my “superman”. Your support was worth more than I can express here.
Publications

- The material in Chapter 4 and Chapter 5 is based on: Bahar Arslan, Vanni Noferini, Françoise Tisseur. The Structured Condition Number of a Smooth Map Between Matrix Manifolds, with Applications. In preparation.
“Perfection is achieved, not when there is nothing more to add, but when there is nothing left to take away.”

– Antoine de Saint-Exupéry
1

Introduction

In this work a matrix function refers to a mapping \( f \) from a set of matrices into a set of matrices. For example, for a polynomial or rational function with scalar coefficients \( f(t) \) and a scalar argument \( t \), \( f(A) \) can be defined by substituting \( A \) for \( t \).

\[
f(t) = \frac{1 + t}{1 - t}, \quad \Rightarrow \quad f(A) = (I - A)^{-1}(I + A), \quad \text{if } 1 \notin \Lambda(A)
\]

where \( \Lambda(A) \) is the set of eigenvalues of \( A \).

The thesis is devoted to several problems aimed to develop the theory of functions of structured matrices. We can divide the contributions into two parts. The first part explores the structure of \( f(A) \) for a given structured \( A \). The second part contains structured error analysis, which motivates the usage of the structure preserving algorithms in the computation of matrix functions.

The structure of thesis is organized as follows. In this chapter we introduce the background materials for the research problems presented in the thesis.

Chapter 2 provides algorithms computing the logarithm of symplectic matrices by using either the Schur-like form or the full form of matrices. We present the advantage of using the Schur-like form of symplectic matrices in the inverse scaling and squaring method in terms of structure preserving and accuracy.

Apart from computation of matrix functions we are interested in error analysis. We try to answer one of the questions arising in functions of nearly structured matrices in Chapter 3. For a given set of structured matrices \( S \) and \( \min\{\|A - X\| : X \in S\} \) the aim is to investigate the bounds of \( \|f(A) - f(X)\| \) for particular structures and for specific
matrix functions. We consider nearly unitary, nearly Hermitian and nearly Hermitian positive semidefinite matrices. The bound is computed for the matrix logarithm, square root, exponential, cosine and sine functions.

In the computation of matrix functions $A$ is subject to small perturbations, which can be rounding or truncating errors. Another possibility is the data might be uncertain. It is essential to know how these small perturbations affect the outcome. The sensitivity of the functions to the small perturbations is measured by condition numbers. The general theory for the condition numbers was developed by Rice [69] and it was applied to matrix functions by Kenney and Laub [45]. In the general theory of condition number of matrix functions there is no requirement on the structure of the perturbed matrix. In Chapter 4 we extend the theory of condition numbers by restricting perturbations to a smaller set that preserves the structure of the perturbed matrix. We have implemented algorithms computing the structured condition number of a matrix function defined between smooth manifolds. We also estimate the structured condition number by applying the power method, which is cheaper to obtain than the “exact” structured condition number. The upper and the lower bounds on the structured condition number are derived with even less computational cost.

Structure is a significant tool not only in matrix functions but also in matrix factorizations. Therefore the need for the structured condition number arises in the computation of structured matrix factorizations. Chapter 5 includes the application of the theory of the structured condition number to the structured matrix factorization. Specifically we are interested in the generalized polar decomposition, structured polar and structured sign decompositions.

Chapter 6 considers the canonical generalized polar decomposition, which is a generalization of the canonical polar decomposition to general scalar products. It is important to point out that apart from theoretical interest it has applications in the solution of corresponding Procrustes problems as shown by [48]. It also gives a way of computing random automorphisms or partial isometries. We discuss different perturbation analysis by giving a first order perturbation bound. Apart from a theoretical aspect of different perturbation analysis we compare our bounds with previous results in the literature. We see that our results give improvements on the previous results in
1. INTRODUCTION

Lastly, in Chapter 7 we summarize the key findings and draw the conclusions of our work. We also recommend future studies on these topics.

1.1 Story of matrix functions

When we look at the literature for matrix functions we see how the theory develops: Sylvester invented the term “matrix” in 1850 [75]. Cayley analysed the algebraic properties of matrices in his “A Memoir on the Theory of Matrices” [12] in which he evaluated the square root of $2 \times 2$ and $3 \times 3$ matrices. Sylvester gave the interpolating polynomial definition of $f(A)$ with distinct eigenvalues [76]. The definition was generalized to the matrices with multiple eigenvalues using Hermite interpolation by Buchheim [8]. The Cauchy integral representation was first stated by Frobenius [24] and then used by Poincaré [68]. The Jordan canonical form definition of $f(A)$ was first proposed by Giorgi [25]. Rinehart [70] showed the equivalence of all these definitions for all $f(A)$. You can find the detailed history of matrix functions in the monograph [35]. We only introduce three equivalent definitions of matrix functions.

For a given $A \in \mathbb{K}^{n \times n}$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) there is a nonsingular $Z \in \mathbb{K}^{n \times n}$ with positive integers $q$ and $m_1 + m_2 + \cdots + m_q = n$ with the distinct eigenvalues of $A$ $\lambda_1, \ldots, \lambda_q \in \mathbb{F}$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$) such that

$$Z^{-1}AZ = J = \text{diag}(J_1(\lambda_1), J_2(\lambda_2), \ldots, J_q(\lambda_q)), \quad J_k(\lambda_k) \in \mathbb{K}^{m_k \times m_k},$$

where

$$J_k(\lambda_k) = \begin{bmatrix} \lambda_k & 1 & & \\ & \lambda_k & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix}.$$

The function $f$ is defined on the spectrum of $A$ if the values

$$f^{(j)}(\lambda_i), \quad j = 0 : n_i - 1, \quad i = 1 : q$$

exist, where $f^{(j)}$ denotes the $j$-th derivative of $f$ and $n_i$ is the dimension of the largest Jordan block including $\lambda_i$. Under these assumptions $f(A)$ can be defined via the
Jordan canonical form as

\[ f(A) := Z f(J) Z^{-1} = Z \text{diag}(f(J_k)) Z^{-1}, \]

where \( f(J_k) \) is given as

\[
\begin{bmatrix}
    f(\lambda_k) & f'(\lambda_k) & \cdots & \frac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\
    0 & f(\lambda_k) & \cdots & \vdots \\
    \vdots & \ddots & \ddots & f'(\lambda_k) \\
    0 & \cdots & 0 & f(\lambda_k)
\end{bmatrix}.
\]

The matrix function definition is independent of the particular Jordan canonical form [35, Prob. 1.1]. In this work we are interested in primary matrix functions so it is necessary to choose the same branch of the function for each Jordan block. A nonprimary matrix function is obtained when \( f \) maps two equal eigenvalues in different Jordan blocks to different values.

A second definition is given via the interpolating polynomial. The unique monic polynomial \( \phi(t) \) of minimum degree that annihilates \( A \) is called the minimal polynomial of \( A \) [40, Thm. 3.3.6]. It has the form

\[ \phi(t) = \prod_{i=1}^{q} (t - \lambda_i)^{n_i} \]

with the distinct eigenvalues of \( A \), \( \lambda_1, \ldots, \lambda_q \). Assuming that \( f \) is defined on the spectrum of \( A \) and the unique polynomial \( p \) with a degree less than \( \sum_{i=1}^{q} n_i = \deg \phi \) satisfies the interpolation conditions

\[ f^{(j)}(\lambda_i) = p^{(j)}(\lambda_i), \quad j = 0 : n_i - 1, \quad i = 1 : q, \]

then \( f(A) := p(A) \). If there are two polynomials \( p(t) \) and \( s(t) \) satisfying the interpolating conditions then \( p(A) = s(A) \).

In the third definition Cauchy integral theorem is used. Let \( \Gamma \subset \mathbb{C} \) be a closed contour that encloses the spectrum of \( A \) and let the function \( f \) be analytic on and inside \( \Gamma \). Then for a given \( A \in \mathbb{K}^{n \times n} \) \( f(A) \) is defined as

\[
  f(A) := \frac{1}{2\pi i} \int_{\Gamma} f(t)(tI - A)^{-1} dt.
\]
It is shown that all these three definitions are equivalent each other [70]. Some basic properties of matrix functions are given in the next theorem. We omit the proofs which can be found in [35, Thm. 1.13].

**Theorem 1.1.** Let $A \in \mathbb{K}^{n \times n}$ and let the functions $f$, $g$ and $h$ which are defined on the spectrum of $A$ be given. Then

a) $A$ commutes with $f(A)$.

b) $f(A^T) = f(A)^T$.

c) $f(XAX^{-1}) = Xf(A)X^{-1}$ holds for any nonsingular matrix $X$.

d) If $XA = AX$, then $Xf(A) = f(A)X$.

e) If $A$ is diagonalizable and $A = Z\text{diag}(\lambda_1, \ldots, \lambda_n)Z^{-1}$, then $f(A)$ is diagonalizable and $f(A) = Z\text{diag}(f(\lambda_1), \ldots, f(\lambda_n))Z^{-1}$.

f) If $A = \text{diag}(A_{11}, \ldots, A_{qq})$ is block diagonal then $f(A) = \text{diag}(f(A_{11}), \ldots, f(A_{qq}))$.

g) $f(A)$ has the same block structure as $A$; that is, if $A = (A_{ij})$ is block triangular then $f(A)$ is block triangular.

Before the sensitivity analysis we present the vector and matrix norms.

**Vector and matrix norms**

A norm is a function $\| \cdot \| : \mathbb{K}^n \rightarrow \mathbb{R}$ that satisfies

- $\|x\| \geq 0$ and $\|x\| = 0$ only if $x = 0$,
- $\|x + y\| \leq \|x\| + \|y\|$,
- $\|\alpha x\| = |\alpha|\|x\|$,

for all vectors $x, y \in \mathbb{K}^n$ and $\alpha \in \mathbb{K}$. Some of the vector norms are defined as

$$
\|x\|_1 = \sum_{i=1}^{n} |x_i|,
$$

$$
\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2},
$$

$$
\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|,
$$

$$
\|x\|_{p} = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad 1 \leq p < \infty.
$$
A matrix \( A \in \mathbb{K}^{m \times n} \) can be viewed as a vector in an \( mn \) dimensional space. Given vector norms \( \| \cdot \|_{(m)} \) and \( \| \cdot \|_{(n)} \) on the range and the domain of \( A \), respectively, the induced matrix norm \( \| A \|_{(m,n)} \) is defined by

\[
\| A \|_{(m,n)} = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\| Ax \|_{(m)}}{\| x \|_{(n)}}.
\]

Let \( q_i \) and \( r_i \) denote the \( i \)-th column and \( i \)-th row of \( A \), respectively. Some of the matrix norms are given by

\[
\begin{align*}
\| A \|_1 &= \max_{1 \leq i \leq n} \| q_i \|_1, & \| A \|_\infty &= \max_{1 \leq i \leq m} \| r_i \|_1, \\
\| A \|_2 &= \sup_{x \neq 0} \{ \| Ax \|_2 / \| x \|_2 \}, & \| A \|_F &= \left( \sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}.
\end{align*}
\]

**Definition 1.2.** A matrix norm \( \| \cdot \| \) is called unitarily invariant if \( \| UAV \| = \| A \| \) for any \( A \in \mathbb{K}^{m \times n} \) and unitary matrices \( U \in \mathbb{K}^{m \times m} \) and \( V \in \mathbb{K}^{n \times n} \).

The condition number examines the changes when we compute \( f(A + \Delta A) \) instead of \( f(A) \). For a matrix function \( f : \mathbb{K}^{n \times n} \to \mathbb{K}^{n \times n} \) and a matrix \( A \) such that \( f(A) \) is defined, the absolute condition number of \( f(A) \) is given as

\[
\text{cond}(f, A) = \lim_{\epsilon \to 0} \sup_{\| E \| \leq \epsilon} \frac{\| f(A + E) - f(A) \|}{\epsilon}, \tag{1.1}
\]

where \( \| \cdot \| \) is any matrix norm. It follows from this definition that

\[
\| f(A + E) - f(A) \| \leq \text{cond}(f, A) \| E \| + o(\| E \|),
\]

which provides a perturbation bound for small \( \| E \| \). The relative condition number of \( f \) at \( A \) can be expressed in terms of the absolute one as

\[
\text{cond}_{rel}(f, A) = \lim_{\epsilon \to 0} \sup_{\| E \| \leq \epsilon\| A \|} \frac{\| f(A + E) - f(A) \|}{\epsilon \| f(A) \|} = \text{cond}(f, A) \| A \|/\| f(A) \|. \tag{1.2}
\]

The condition number (1.1) can be expressed in terms of the Fréchet derivative of \( f \) at \( A \), which is a linear map \( L_f(A, E) : \mathbb{K}^{n \times n} \to \mathbb{K}^{n \times n} \) such that

\[
\| f(A + E) - f(A) - L_f(A, E) \| = o(\| E \|) \tag{1.2}
\]

for all \( E \in \mathbb{K}^{n \times n} \). When the Fréchet derivative of \( f \) at \( A \) exists, it is unique. In that case we have that [35, Thm. 3.1]

\[
\text{cond}(f, A) = \max_{E \neq 0} \frac{\| L_f(A, E) \|}{\| E \|} = \| L_f(A) \|. 
\]
1. INTRODUCTION

An explicit formula for the Fréchet derivative is usually difficult to obtain. In the literature there are some numerical algorithms evaluating the Fréchet derivative for the exponential [47], [2], logarithm [3], fractional power [9], [36]. By using Daleckii-Krein formula Noferini gave an explicit expression for the Fréchet derivative of generalized matrix functions [66]. In this work we assume that we have a numerical method to evaluate $L_f(A, E)$ for a given $E$. Since the Fréchet derivative is linear in $E$, applying the vec operator to $L_f(A, E)$ gives

$$\text{vec}(L_f(A, E)) = K_f(A)\text{vec}(E),$$  \hspace{1cm} (1.3)

where $K_f(A) \in \mathbb{K}^{n^2 \times n^2}$ is called the Kronecker form of the Fréchet derivative as shown in Figure 1.1. If we specialize to the Frobenius norm we have that

$$\text{cond}(f, A) = \max_{E \neq 0} \frac{\|L_f(A, E)\|_F}{\|E\|_F} = \max_{E \neq 0} \frac{\|\text{vec}(L_f(A, E))\|_2}{\|\text{vec}(E)\|_2} = \|K_f(A)\|_2,$$  \hspace{1cm} (1.4)

where we use the fact that for $A \in \mathbb{K}^{n \times n}$, $\|A\|_F = \|\text{vec}(A)\|_2$ and the vec operator is given in Definition 1.19. The problem of computing $\text{cond}(f, A)$ then reduces to finding the 2-norm of $K_f(A)$. Note that the latter matrix can be constructed exactly by forming one column at a time using (1.3),

$$K_f(A)e_{i+(j-1)n} = \text{vec}(L_f(A, e_i e_j^T)), \quad i = 1: n, \quad j = 1: n,$$  \hspace{1cm} (1.5)

where $e_k$ is a vector of appropriate dimension with the $k$th entry equal to one and zero everywhere else. Numerically constructing $K_f(A)$ this way costs $O(n^5)$ operations assuming that $L_f(A, E)$ can be computed in $O(n^3)$ operations. Computing the 2-norm of $K_f(A)$ costs $O(n^6)$ operations.
1.2. STRUCTURED MATRICES

Instead of forming $K_f(A)$ we can estimate $\text{cond}(f, A)$, which is cheaper. A lower bound for $\text{cond}(f, A)$ can be computed by the power method applied to $K_f(A)^*K_f(A)$ and which, for a nonzero matrix $E_0 \in \mathbb{K}^{n \times n}$ constructs the iterates

$$Z_{k+1} = L_f(A, E_k), \quad E_{k+1} = L_f^*(A, Z_{k+1}), \quad \gamma_{k+1} = \|E_{k+1}\|_F/\|Z_{k+1}\|_F, \quad k > 0,$$

with $\gamma_k$ such that $\gamma_k \leq \|L_f(A)\|_F$ and $\gamma_k \to \|L_f(A)\|_F$ as $k \to \infty$. In (1.6), $L_f^*$ is the adjoint of $L_f$ and is given by

$$L_f^*(A, E) = \begin{cases} L_f(A^T, E) & \mathbb{K} = \mathbb{R}, \\ L_f(A^*, E) & \mathbb{K} = \mathbb{C}, \end{cases}$$

where $\bar{f}(z) := f(\bar{z})$. For example if $f$ has a power series representation then $\bar{f}$ is obtained by taking the complex conjugates of the coefficients of the power series. The computation of the $k$th iteration in (1.6) costs $O(n^3)$ operations. The convergence test for the power method is given by

$$k > \max_{\text{it}}, \quad \text{or} \quad |\gamma_{k+1} - \gamma_k| \leq \text{tol} \cdot \gamma_{k+1},$$

where $\max_{\text{it}}$ is a prescribed maximal number of iterations and $\text{tol}$ is a fixed tolerance. For more details about the power method and its convergence see [27, Chapt. 7].

1.2 Structured matrices

In this thesis we consider structured matrices arising in the context of a nondegenerate scalar product on $\mathbb{K}^n$, which we define in this section.

**Definition 1.3.** [40, Def. 5.1.3] Let $V$ be a vector space over a field $\mathbb{K}$. A function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{K}^n$ is a scalar product if for all $x, y, z \in V$ and all $\alpha \in \mathbb{K}$,

i) $\langle x, x \rangle \geq 0$.

ii) $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$.

iii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ and if $\mathbb{K} = \mathbb{C}$ then $\langle \alpha x, y \rangle = \bar{\alpha} \langle x, y \rangle$, where the bar indicates the complex conjugate.
iv) \( \langle x, y \rangle = \overline{\langle y, x \rangle} \).

If \( \langle x, y \rangle = 0 \) for all \( y \) implies \( x = 0 \) then the scalar product is called nondegenerate, otherwise it is degenerate.

The scalar product on \( \mathbb{K}^n \) is called bilinear if it is linear in each argument, that is

\[
\langle a_1 x_1 + a_2 x_2, y \rangle = a_1 \langle x_1, y \rangle + a_2 \langle x_2, y \rangle,
\]

\[
\langle x, b_1 y_1 + b_2 y_2 \rangle = b_1 \langle x, y_1 \rangle + b_2 \langle x, y_2 \rangle,
\]

for all \( a_i, b_i \in \mathbb{K} \) and \( x, y, x_i, y_i \in \mathbb{K}^n, i = 1, 2 \). If \( \mathbb{K} = \mathbb{C} \) and the scalar product is conjugate linear in the first argument and linear in the second,

\[
\langle a_1 x_1 + a_2 x_2, y \rangle = \overline{a_1} \langle x_1, y \rangle + \overline{a_2} \langle x_2, y \rangle,
\]

\[
\langle x, b_1 y_1 + b_2 y_2 \rangle = b_1 \langle x, y_1 \rangle + b_2 \langle x, y_2 \rangle,
\]

for all \( \overline{a_i}, b_i \in \mathbb{K} \) and \( x, y, x_i, y_i \in \mathbb{K}^n, i = 1, 2 \), then it is called a sesquilinear form.

**Proposition 1.4.** [62] Given a nondegenerate bilinear or sesquilinear form on \( \mathbb{K}^n \), there exists a unique nonsingular \( M \in \mathbb{K}^{n \times n} \) such that for all \( x, y \in \mathbb{K}^n \),

\[
\langle x, y \rangle_M = \begin{cases} 
  x^T My, & \text{for bilinear forms}, \\
  x^* My, & \text{for sesquilinear forms}.
\end{cases}
\]

(1.8)

\( M \) is called the matrix associated with the form. We will denote \( \langle x, y \rangle \) by \( \langle x, y \rangle_M \) to prevent confusion. Throughout the thesis we assume that the scalar product is nondegenerate.

**Definition 1.5.** Let \( \langle x, y \rangle_M \) be a nondegenerate scalar product on \( \mathbb{K}^n \). For any matrix \( A \in \mathbb{K}^{n \times n} \), the adjoint of \( A \) with respect to \( \langle x, y \rangle_M \) denoted by \( A^* \), is uniquely defined by

\[
\langle Ax, y \rangle_M = \langle x, A^* y \rangle_M, \quad \forall x, y \in \mathbb{K}^n.
\]

(1.9)

From the definition of \( \langle x, y \rangle_M \), we can write the adjoint matrix explicitly,

\[
A^* = \begin{cases} 
  M^{-1} A^T M, & \text{for bilinear forms}, \\
  M^{-1} A^* M, & \text{for sesquilinear forms}.
\end{cases}
\]
1.2. STRUCTURED MATRICES

Associated with the scalar product $\langle \cdot, \cdot \rangle_M$, there are three important classes of structured matrices. The automorphism group of matrices that preserve the value of the scalar product is defined by

\[ G_M := \{ G \in \mathbb{K}^{n \times n} : \langle Gx, Gy \rangle_M = \langle x, y \rangle_M \} = \{ G \in \mathbb{K}^{n \times n} : G^* = G^{-1} \}. \]

The Lie algebra comprises of the skew-adjoint matrices with respect to the scalar product and it is defined by

\[ L_M := \{ K \in \mathbb{K}^{n \times n} : \langle Kx, y \rangle_M = -\langle x, Ky \rangle_M \} = \{ K \in \mathbb{K}^{n \times n} : K^* = -K \}. \]

The Jordan algebra comprises of the matrices $S$ that are self-adjoint with respect to the scalar product and it is defined by

\[ J_M := \{ S \in \mathbb{K}^{n \times n} : \langle Sx, y \rangle_M = \langle x, Sy \rangle_M \} = \{ S \in \mathbb{K}^{n \times n} : S^* = S \}. \]

In our work we require two properties of scalar products called unitary and orthosymmetric.

**Definition 1.6.** The scalar product $\langle \cdot, \cdot \rangle_M$ is unitary if $\alpha M$ is unitary for some $\alpha > 0$. It is orthosymmetric if

\[ M = \begin{cases} \mu M^T, & \mu = \pm 1, \text{ for bilinear forms}, \\ \mu M^*, & |\mu| = 1, \text{ for sesquilinear forms}. \end{cases} \]

See [63] for a list of equivalent properties for the unitary and orthosymmetric scalar products. Examples of structured matrices associated with some scalar products $\langle \cdot, \cdot \rangle_M$ and their adjoints are given in Table 1.1 [62, Table 2.1].
Table 1.1: Examples of structured matrices associated with the scalar product $\langle \cdot , \cdot \rangle_M$.

### Bilinear forms

<table>
<thead>
<tr>
<th>Space</th>
<th>M</th>
<th>Adjoint</th>
<th>Automorphism Group</th>
<th>Jordan Algebra</th>
<th>Lie Algebra</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{R}^n$</td>
<td>$I$</td>
<td>$A^* = A^T$</td>
<td>Real orthogonals</td>
<td>Symmetrics</td>
<td>Skew-symmetrics</td>
</tr>
<tr>
<td>$\mathbb{C}^n$</td>
<td>$I$</td>
<td>$A^* = A^T$</td>
<td>Complex orthogonals</td>
<td>Complex symmetrics</td>
<td>Complex skew-symmetrics</td>
</tr>
<tr>
<td>$\mathbb{R}^n$</td>
<td>$\Sigma_{p,q}$</td>
<td>$A^* = \Sigma_{p,q} A^T \Sigma_{p,q}$</td>
<td>Pseudo-orthogonals</td>
<td>Pseudo-symmetrics</td>
<td>Pseudo skew-symmetrics</td>
</tr>
<tr>
<td>$\mathbb{C}^n$</td>
<td>$\Sigma_{p,q}$</td>
<td>$A^* = \Sigma_{p,q} A^T \Sigma_{p,q}$</td>
<td>Complex pseudo-orthogonals</td>
<td>Complex pseudo-symmetrics</td>
<td>Complex pseudo-skew-symmetrics</td>
</tr>
<tr>
<td>$\mathbb{R}^n$</td>
<td>$R$</td>
<td>$A^* = RA^T R$</td>
<td>Real perplectics</td>
<td>Persymmetric</td>
<td>Perskew-symmetrics</td>
</tr>
<tr>
<td>$\mathbb{R}^{2n}$</td>
<td>$J$</td>
<td>$A^* = -JA^T J$</td>
<td>Real symplectics</td>
<td>Skew-Hamiltonians</td>
<td>Hamiltonians</td>
</tr>
<tr>
<td>$\mathbb{C}^{2n}$</td>
<td>$J$</td>
<td>$A^* = -JA^T J$</td>
<td>Complex symplectics</td>
<td>J-skew-symmetric</td>
<td>J-symmetric</td>
</tr>
</tbody>
</table>

### Sesquilinear forms

<table>
<thead>
<tr>
<th>Space</th>
<th>M</th>
<th>Adjoint</th>
<th>Automorphism Group</th>
<th>Hermitian</th>
<th>Skew Hermitian</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{C}^n$</td>
<td>$I$</td>
<td>$A^* = A^*$</td>
<td>Unitaries</td>
<td>Hermitian</td>
<td>Skew Hermitian</td>
</tr>
<tr>
<td>$\mathbb{C}^n$</td>
<td>$\Sigma_{p,q}$</td>
<td>$A^* = \Sigma_{p,q} A^* \Sigma_{p,q}$</td>
<td>Pseudo-unitaries</td>
<td>Pseudo Hermitian</td>
<td>Pseudo skew-Hermitian</td>
</tr>
<tr>
<td>$\mathbb{C}^{2n}$</td>
<td>$J$</td>
<td>$A^* = -JA^* J$</td>
<td>Conjugate symplectics</td>
<td>J-skew-Hermitian</td>
<td>J-Hermitian</td>
</tr>
</tbody>
</table>

\[
\Sigma_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}, \quad p + q = n, \quad R = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \end{bmatrix}, \quad \text{and} \quad J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}. \tag{1.10}
\]
Now suppose that $A$ belongs to an automorphism group, Lie algebra, or Jordan algebra. When does $f(A)$ inherit the structure? It is shown in [38, Thm. 3.1] that, for bilinear forms,

$$f(A)^* = f(A^*)$$

(1.11)

holds for all matrix functions $f$ assuming that $f$ is defined at $A$ and $A^*$. For sesquilinear forms, (1.11) holds when, for example, the function $f$ has a convergent power series representation

$$f(A) = \sum_{k=0}^{\infty} \alpha_k A^k,$$

with $\alpha_k \in \mathbb{R}$. Assuming that $f$ is defined at the indicated arguments and that $f$ satisfies (1.11), we have that

(i) if $A \in J_M$ then $f(A) \in J_M$ since $f(A)^* = f(A)$,

(ii) if $A \in L_M$ then $f(A)^* = f(-A)$ so that

- $f(A) \in J_M$ if $f$ is an even function [35, Problem 1.20],
- $f(A) \in L_M$ if $f$ is an odd function [35, Problem 1.20],
- $f(A) \in G_M$ if $f(-A) = f(A)^{-1}$ (e.g., the matrix exponential),

(iii) if $A \in G_M$ then $f(A)^* = f(A^{-1})$ so that

- $f(A) \in L_M$ if $f(A^{-1}) = -f(A)$ (e.g., the matrix logarithm),
- $f(A) \in G_M$ if $f(A^{-1}) = f(A)^{-1}$ for bilinear forms and $f(A^{-\ast}) = f(A)^{-\ast}$ for sesquilinear forms [38, Thm.3.1] (e.g., the principal square root).

1.3 Preliminaries in differential geometry

In this part, we recall some basic notions in differential geometry that will be useful to study matrix functions mapping between smooth manifolds. More details on the material can be found in [51].

Definition 1.7. Let $\Omega$ be a set and $\mathcal{M}$ be a collection of subsets of $\Omega$ with the following properties
• The empty set $\emptyset \in \mathcal{M}$ and the space $\Omega \in \mathcal{M}$.

• The union of any (finite or infinite) number of sets in $\mathcal{M}$ belongs to $\mathcal{M}$.

• The intersection of any two sets in $\mathcal{M}$ belongs to $\mathcal{M}$.

Then we say that $\mathcal{M}$ is a topology on $\Omega$ and that $(\Omega, \mathcal{M})$ is a topological space.

Before we give the definition of a homeomorphism we need a continuity of functions in topological spaces.

**Definition 1.8.** Let $(\Omega, \mathcal{M})$ and $(\Omega, \mathcal{N})$ be topological spaces, and $f: \mathcal{M} \to \mathcal{N}$ be a function between them. Then $f$ is continuous at a point $p \in \mathcal{M}$ if for every neighbourhood $B_p$ of $f(p)$, there exists a neighbourhood $B$ of $p$ such that $f(B) \subseteq B_p$.

**Definition 1.9.** A homeomorphism is a function $f: \mathcal{M} \to \mathcal{N}$ between two topological spaces $\mathcal{M}$ and $\mathcal{N}$, that is continuous, bijective and has a continuous inverse function $f^{-1}$. Then these two topological spaces $\mathcal{M}$ and $\mathcal{N}$ are said to be homeomorphic.

**Definition 1.10.** Suppose $\mathcal{M}$ is a topological space. We say that $\mathcal{M}$ is a (topological) $n$-manifold if it has the following properties.

• For every pair of distinct points $p, q \in \mathcal{M}$, there are disjoint open subsets $U, V \subseteq \mathcal{M}$ such that $p \in U$ and $q \in V$.

• There exists a countable basis for the topology of $\mathcal{M}$.

• Each point of $\mathcal{M}$ has a neighbourhood that is homeomorphic to an open subset of $\mathbb{R}^n$. That is, it is locally Euclidean of dimension $n$. Then, for every point $p$ on an $n$-dimensional manifold $\mathcal{M}$, the points in the neighbourhood of $p$ can be specified by using $n$ coordinates $(x_1, x_2, \ldots, x_n)$ and there exists a continuous map from that neighbourhood to $\mathbb{R}^n$.

**Definition 1.11.** Let a topological $n$-manifold $\mathcal{M}$ be given. A coordinate chart on $\mathcal{M}$ is a pair $(U, \varphi)$, where $U \subset \mathcal{M}$ is open and $\varphi: U \to \hat{U}$ is a homeomorphism from $U$ to an open subset $\hat{U} = \varphi(U) \subseteq \mathbb{R}^n$. 

Definition 1.12. An \(n\)-dimensional atlas \(A\) on \(\mathcal{M}\) is a collection of coordinate charts \(\{(U_\alpha, \varphi_\alpha)\}\) such that

- The \(U_\alpha\) cover \(\mathcal{M}\).
- For each \(\alpha\), \(\varphi_\alpha(U_\alpha) \subseteq \mathbb{R}^n\).
- Whenever \(U_\alpha \cap U_\beta \neq \emptyset\) the transition map \(\varphi_\alpha \varphi_\beta^{-1}: \varphi_\alpha(U_\alpha \cap U_\beta) \rightarrow \varphi_\beta(U_\alpha \cap U_\beta)\) is \(C^\infty\) (infinitely differentiable) with \(C^\infty\) inverse.

To prove something is a manifold it is enough to find at least one atlas. A smooth manifold is a pair \((\mathcal{M}, A)\), where \(\mathcal{M}\) is a topological space and \(A\) is \(C^\infty\) on \(\mathcal{M}\). We will show that the matrix classes that we define are smooth manifolds. For the Lie and the Jordan algebras it is enough to show them to be vector subspaces so they are smooth manifolds.

Lemma 1.13. The Lie algebra \(L_M \subseteq \mathbb{K}^{n \times n}\) and the Jordan algebra \(J_M \subseteq \mathbb{K}^{n \times n}\) are vector subspaces.

Proof. To show that \(L_M\) and \(J_M\) are vector subspaces we need to show that they are closed under addition and multiplication by real scalars. The defining equations \(M^{-1}A^TM = \pm A\) and \(M^{-1}B^TM = \pm B\) are equivalent to \(A, B \in L_M\) or \(J_M\), according to the choice of sign. Hence, for any \(\alpha \in \mathbb{R}\)

\[
M^{-1}(A + \alpha B)^TM = \pm(A + \alpha B) .
\]

To prove that an automorphism group is a smooth manifold we need to find at least one smooth atlas. Before showing that we need to state the following theorem. For \(r > 0\), denote

\[
B_r := \{W \in \mathbb{K}^{n \times n} : \|W\| < r\}.
\]

Theorem 1.14. [77, Thm. 7.1] Let \(G_M\) and \(L_M\) be the automorphism group and Lie algebra of a scalar product \(\langle \cdot, \cdot \rangle_M\).

1) For all \(Y \in L_M\), \(\exp(Y) \in G_M\).
2) For sufficiently small $r > 0$, $V := \exp(B_r \cap \mathbb{L}_M)$ is a neighbourhood of $I$ in $G_M$, and the restriction

$$\exp : B_r \cap \mathbb{L}_M \to V$$

is a homeomorphism.

A diffeomorphism $f : U \to V$ is a homeomorphism such that both $f$ and $f^{-1}$ are differentiable functions.

**Theorem 1.15.** The automorphism group $G_M$ of dimension $n$ is a real submanifold of $\mathbb{K}^{n \times n}$. Furthermore, when $M$ defines a complex bilinear form, $G_M$ is also a complex submanifold of $\mathbb{C}^{n \times n}$. 

**Proof.** The first part of the proof is taken from [77, Thm. 7.17]. Let $G_M$ be an automorphism group of dimension $n$ with Lie algebra $\mathbb{L}_M$. Choose $r > 0$ as in Theorem 1.14. Then $V = \exp(B_r \cap \mathbb{L}_M)$ is a neighbourhood of $I$ in $G_M$, and the restriction $\exp : B_r \cap \mathbb{L}_M \to V$ is a diffeomorphism.

Let $g \in G_M$ be arbitrary. Define the chart $\mathcal{L}_g : \mathbb{K}^{n \times n} \to \mathbb{K}^{n \times n}$ as

$$\mathcal{L}_g(A) := g.A$$

which restricts $\mathcal{L}_g$ to a diffeomorphism from $G_M$ to $G_M$. Then $\mathcal{L}_g(V)$ is a neighbourhood of $g$ in $G_M$ and

$$(\mathcal{L}_g \circ \exp) : B_r \cap \mathbb{L}_M \to \mathcal{L}_g(V)$$

is a diffeomorphism at $g$.

For the second part of the proof note that $G_M$ is the set of solutions of the quadratic matrix equation $M = X^*MX$. Hence the equation is equivalent to either $n^2$ complex polynomial equations or $2n^2$ real polynomial equations. It means $G_M$ is either a complex or a real algebraic variety. A complex (resp. real) algebraic variety is a complex (resp. real) manifold if and only if it does not contain singular points, i.e., points such that the rank of the Jacobian is locally smaller than at a generic point.

We also need the definition of tangent spaces to give the differential of the maps.
Definition 1.16. Let $\mathcal{M} \subset \mathbb{K}^n$ be a smooth $n$-manifold and fix a point $p \in \mathcal{M}$. A vector $v \in \mathbb{K}^n$ is called a tangent vector of $\mathcal{M}$ at $p$ if there is a smooth curve $\gamma : \mathbb{K} \to \mathcal{M}$ such that $\gamma(0) = p$, $\gamma'(0) = v$. The set

$$T_p\mathcal{M} = \{\gamma'(0) = v \mid \gamma : \mathbb{K} \to \mathcal{M} \text{ is smooth}, \gamma(0) = p\}$$

of tangent vectors of $\mathcal{M}$ at $p$ is called the tangent space of $\mathcal{M}$ at $p$.

Definition 1.17 (Differential of smooth maps). Let $\mathcal{M}$, $\mathcal{N}$ be smooth manifolds and suppose one has a differentiable map $f : U \to V$, where $U \subset \mathcal{M}$ and $V \subset \mathcal{N}$. Under the hypotheses of Definition 1.16 the map $df_p$ is a well-defined linear function from $T_p\mathcal{M}$ to $T_{f(p)}\mathcal{N}$ and

$$df_p(\gamma'(0)) = (f \circ \gamma)'(0). \quad (1.12)$$

By taking smooth square matrix manifolds and $p = A$ we illustrate the differential of $f$ between tangent spaces in Figure 1.2.

Figure 1.2: The differential of $f$ is a map between tangent spaces $df(A) : T_A\mathcal{M} \to T_{f(A)}\mathcal{N}$.

We claim that the differential of $f$ plays the same role as the Fréchet derivative. In the computation of condition number, instead of using the Fréchet derivative of $f$ we can use $df(A)$ defined between tangent spaces to the related smooth matrix manifolds.
1.4 Some useful properties

We will give some definitions and properties that we use in our analysis throughout the thesis.

Kronecker product and vec operator

**Definition 1.18.** The Kronecker product of a matrix $A \in \mathbb{K}^{m \times n}$ with the matrix $B \in \mathbb{K}^{n \times z}$ is defined as

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}. $$

**Definition 1.19.** For any matrix $A \in \mathbb{K}^{m \times n}$ the vec operator stacks the columns of a matrix into one column vector \[49\],

$$\text{vec}(A) = [a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{m2}, a_{1n}, \ldots, a_{mn}]^T.$$

We list some useful properties of vec operator and the Kronecker product \[49\].

**Theorem 1.20.** For all $A, B, C, D, X \in \mathbb{K}^{n \times n}$ and $\alpha \in \mathbb{K}$,

1. $(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha (A \otimes B)$.

2. $(A \otimes B)^* = A^* \otimes B^*$.

3. $(A \otimes B)(C \otimes D) = (AC \otimes BD)$.

4. $\text{vec}(AXB) = (B^T \otimes A)\text{vec}(X)$.

We give the properties for square matrices but they are valid for rectangular matrices of appropriate dimensions as well.
Lemma 1.21. The vec permutation matrix $P$ satisfying \( \text{vec}(A^T) = P\text{vec}(A) \) \cite[Sec. 16.6]{34} is given explicitly by

\[
P = \sum_{i,j}^n (e_i e_j^T) \otimes (e_j e_i^T),
\]

where $e_k$ is a vector of appropriate dimension with $k$th entry equal to one and zero everywhere else.

A pseudoinverse of $A$ is a generalization of matrix inverse and the most commonly used one is the Moore-Penrose pseudoinverse.

Definition 1.22 (Moore-Penrose pseudoinverse, \cite{27}). For a given $A \in \mathbb{K}^{m \times n}$, $A^+ \in \mathbb{K}^{n \times m}$ is defined to be the unique matrix satisfying the following Moore-Penrose conditions.

- $AA^+ A = A$
- $(AA^+)^* = AA^+
- $A^+ AA^+ = A^+$
- $(A^+ A)^* = A^+ A$

Padé approximation

One of the ways of approximating a given function $f(x) : U \to V$ is the polynomials. Polynomials have some advantages.

- It is possible to construct an approximation such that $|f(x) - p(x)| < \epsilon$ for all $x \in U$ within an arbitrary tolerance $\epsilon$ (Weierstrass Theorem).
- They are evaluated easily at arbitrary points.
- Derivatives and integrals are easily determined.

However, besides these advantages the polynomials have some drawbacks.

- They do not have finite asymptotes.
- They are finite on the finite real axis and tend to $\pm \infty$ as $x \to \pm \infty$.
- They have a tendency to oscillate.
In contrast, rational functions can be bounded as $x \to \pm \infty$. They have poles and they are also free of oscillation. Therefore, we prefer to approximate a given function $f(x)$ by rational functions. As a rational function we use Padé approximants.

**Definition 1.23.** The rational function

$$r_{mn}(x) = \frac{p_m(x)}{q_n(x)}$$

is a $[m/n]$ Padé approximant of $f(x)$ if $p_m(x)$ is a polynomial in $x$ of degree at most $m$, $q_n(x)$ is a polynomial in $x$ of degree at most $n$ with $q_n(0) = 1$ and

$$f(x) - r_{mn}(x) = O(x^{m+n+1}).$$

**Floating point arithmetic**

For the accuracy of algorithms of numerical linear algebra we need to explain floating point arithmetic. More details can be found in [34] and [79]. Digital computers use a finite number of bits to represent a real number so they can represent a finite subset of real numbers. A floating point number system $F$ is a subset of the real numbers $F \subset \mathbb{R}$. The elements in $F$ have the form

$$y = \pm m \times \beta^{e-t},$$

where $\beta$ is the base, $t$ is the precision, $e$ is the exponent and $m$ is the mantissa satisfying $0 \leq m \leq \beta^t - 1$. For the unique representation of $y \neq 0$ we choose $m \geq \beta^{t-1}$. The range of the exponent is given as $e_{\min} \leq e \leq e_{\max}$. In IEEE double precision arithmetic system [1] we choose $\beta = 2$, $t = 53$, $e_{\min} = -1021$ and $e_{\max} = 1024$. The nonzero floating point numbers in $F$ have the range as $\beta^{e_{\min}-1} \leq |y| \leq \beta^{e_{\max}}(1 - \beta^{-t})$. IEEE double precision arithmetic permits numbers to be as large as $1.80 \times 10^{308}$ and as small as $2.23 \times 10^{-308}$.

Let us give the machine precision as $u = \frac{1}{2} \beta^{1-t}$. In IEEE double precision arithmetic the machine precision is specified to be $2^{-53} \approx 1.11 \times 10^{-16}$. The following theorem states that the difference between a real number and its closest floating point approximation is always smaller than the machine precision in relative terms.
Theorem 1.24. Let \( f_l : \mathbb{R} \to F \) be a function giving the closest floating point approximation to a real number. For all \( x \in \mathbb{R} \), there exists \( \delta \) with \( |\delta| \leq u \) such that
\[
fl(x) = x(1 + \delta).
\]

After showing the presentation of floating point numbers we need to consider the operations. All mathematical computations are reduced to certain elementary arithmetic operations. For all \( x, y \in F \), there exists \( \delta \) with \( |\delta| \leq u \) such that
\[
fl(x \text{ op } y) = (x \text{ op } y)(1 + \delta), \quad \text{op} = +, -, \times, /.
\]

Every operation of floating point arithmetic is exact up to a relative error of size at most machine precision.

All computations in this thesis have been carried out in IEEE double precision arithmetic. For the experiments we run MATLAB R2014b or R2013a with the machine precision \( 1.11 \times 10^{-16} \) on a MacBook Pro Retina with an Intel Core i7 2.6 GHz processor.
2

Using the Schur-like Form in Computing the Matrix Logarithm

2.1 Introduction

The matrix logarithm is not only important for being the inverse function of the matrix exponential, it has also many applications. It has been used by engineers in the continuization process. They compute the logarithm of matrices in converting a discrete process into a continuous one [73], [50]. It has also applications to the stability of differential equations [65], [72]. The logarithm of structured matrices has also some applications in the control mechanical systems [67], [16] and in the optometry [28]. Structured matrix logarithm is also used for generalizing Bézier curves to non-Euclidean spaces. Crouch’s algorithm, which generalizes De Casteljau algorithm to find polynomial splines on Riemannian manifolds requires the computation of matrix logarithm when this manifold is a Lie group of matrices [16]. The theory of splines on Lie groups has applications in robotics path planning and air traffic control.

The growing interest in computing structured matrix functions stems from the fact that predicting and preserving the structure of matrices can help us to explain the results physically and geometrically. Also, the development of specialised algorithms that exploit the structure leads to improved performance in terms of lower computational cost and memory requirement.
2.1. INTRODUCTION

We focus on computing the logarithm of a symplectic matrix $A$ with the spectrum $\rho(A)$ such that $\rho(A) \cap \mathbb{R}^- = \emptyset$, for which $W = \log A$ is Hamiltonian as stated in (iii) on page 29 and in Table 1.1. In the computation of matrix logarithm, we use the inverse scaling and squaring method, proposed by Kenney and Laub [45] and which is based on the relation

$$\log(A) = 2^k \log(A^{1/2^k}).$$

There are two important calculations in the inverse scaling and squaring method. The first one is taking a square root of a symplectic matrix and the second is the Padé approximation. We will analyse the matrix square root and Padé approximation in terms of their structure preservation.

The inverse scaling and squaring method can be applied to $A$ directly or it can be used with the Schur decomposition. However, we show that the latter case does not preserve the structure of the symplectic matrices in an automorphism group. The main contribution of this chapter is using the Schur-like form of the symplectic matrices that preserves the structure with the structure preserving iterations for the square root in the inverse scaling and squaring method.

The chapter is given as follows. In Section 2.2 we present the Schur-like form of the symplectic matrices with necessary conditions for the existence and the uniqueness. In Section 2.3 we give the matrix logarithm and the basic idea of the inverse scaling and squaring method. There are two main computations in this method: taking a square root and Padé approximation which are analysed in Section 2.4 and 2.5, respectively. Three different structure preserving methods are given for the computation of square root of symplectic matrices in Section 2.4. In Section 2.6 we use these three methods in the inverse scaling and squaring method. We discuss the advantage of using the Schur-like form in terms of structure preserving and computational cost. In Section 2.7 we show the error analysis used in the numerical experiments to measure the relative error and the distance from the proposed structure. In the final section we illustrate the results with numerical tests and compare these three algorithms presented.
Table 2.1: The structure of the matrices in the automorphism group and Lie algebra.

<table>
<thead>
<tr>
<th>Family</th>
<th>Definition</th>
<th>Block Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthogonal</td>
<td>$Q \in \mathbb{R}^{2n \times 2n}$ is orthogonal symplectic if $Q$ is orthogonal and $JQ = QJ$.</td>
<td>$Q = \begin{bmatrix} Q_1 &amp; Q_2 \ -Q_2 &amp; Q_1 \end{bmatrix}$, $Q_1^T Q_2$ is symmetric and $Q_1^T Q_1 + Q_2^T Q_2 = I$.</td>
</tr>
<tr>
<td>Symplectic</td>
<td>$S \in \mathbb{R}^{2n \times 2n}$ is symplectic if $JS = S^{-T} J$.</td>
<td>$S = \begin{bmatrix} S_1 &amp; S_2 \ S_3 &amp; S_4 \end{bmatrix}$, $S_4^T S_3$, $S_4^T S_2$ are symmetric and $S_4^T S_4 = I + S_3^T S_2$.</td>
</tr>
<tr>
<td>Hamiltonian</td>
<td>$H \in \mathbb{R}^{2n \times 2n}$ is Hamiltonian if $JH = (JH)^T$.</td>
<td>$H = \begin{bmatrix} A &amp; G \ F &amp; -A^T \end{bmatrix}$, $F$ and $G$ are symmetric.</td>
</tr>
<tr>
<td>Skew-Hamiltonian</td>
<td>$K \in \mathbb{R}^{2n \times 2n}$ is skew-Hamiltonian if $JK = -(JK)^T$.</td>
<td>$K = \begin{bmatrix} A &amp; G \ F &amp; A^T \end{bmatrix}$, $F$ and $G$ are skew-symmetric.</td>
</tr>
</tbody>
</table>

\[
J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}
\]

2.2 Schur-like form of symplectic matrices

The structure of symplectic, orthogonal-symplectic, and Hamiltonian matrices [27, Figure 7.8.1] are given in Table 2.1. This helps us to understand the forms of the matrices that we obtain in our calculations. The following definitions are needed to ensure the existence of the Schur-like form. Note that we apply the theory only to real matrices.

**Definition 2.1.** [26, Definition 1.1.] A subspace $V \subseteq \mathbb{R}^n$ is called invariant for $A \in \mathbb{R}^{n \times n}$, or $A$-invariant, if $Ax \in V$ for every vector $x \in V$.

**Definition 2.2.** Two matrices $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ are said to be congruent, denoted by $A \sim^c B$, if there exists a nonsingular matrix $S$ such that $B = SAS^T$.

The conditions for the existence of the Schur-like form of the symplectic matrices are stated in the following theorem.

**Theorem 2.3.** [23, Thm. 2.12] Let $A \in \mathbb{R}^{2n \times 2n}$ be a symplectic matrix. Then there
exists a real orthogonal and symplectic matrix $Q \in \mathbb{R}^{2n \times 2n}$ such that

$$Q^T A Q = T = \begin{bmatrix} T_1 & T_2 \\ 0 & T_1^{-T} \end{bmatrix}, \quad (2.1)$$

where $T_1 \in \mathbb{R}^{n \times n}$ is a quasi upper triangular matrix if and only if every eigenvalue $\lambda$ of $A$ with modulus 1 has even algebraic multiplicity, say $2k$, and any matrix $\chi_k \in \mathbb{C}^{2n \times 2k}$ with the property that its columns span a basis of the maximal $A$-invariant subspace corresponding to $\lambda$ satisfies

$$\chi_k^* J \chi_k \sim c \begin{bmatrix} 0 & I_k \\ -I_k & 0 \end{bmatrix}.$$ 

Moreover, $Q$ can be chosen such that all the eigenvalues of $T_1$ are inside the closed unit disk.

In the numerical experiments, we will use the Schur-like form of real symplectic matrices given either in the form (2.1) or in the following form [62, Table 3.1]

$$T = \begin{bmatrix} T_1 & T_1 S \\ 0 & T_1^{-T} \end{bmatrix}, \quad \text{and} \quad S = S^T, \quad (2.2)$$

where $T_1$ is a quasi upper triangular matrix. It is important to show that the matrix given in equation (2.2) is also symplectic.

**Lemma 2.4.** Let $T \in \mathbb{R}^{2n \times 2n}$ be given in the form

$$T = \begin{bmatrix} T_1 & T_1 S \\ 0 & T_1^{-T} \end{bmatrix}, \quad (2.3)$$

where $S \in \mathbb{R}^{n \times n}$ is symmetric and $T_1 \in \mathbb{R}^{n \times n}$ is quasi upper triangular. Then $T$ is a symplectic matrix. Besides, any symplectic matrix in the form (2.1) can be rewritten in the new form (2.3).

**Proof.** If $T$ has the form given in equation (2.3) then

$$T^T J T = \begin{bmatrix} T_1 & T_1 S \\ 0 & T_1^{-T} \end{bmatrix}^T \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \begin{bmatrix} T_1 & T_1 S \\ 0 & T_1^{-T} \end{bmatrix}$$

$$= \begin{bmatrix} T_1^T & 0 \\ 0 & T_1^{-T} \end{bmatrix} \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \begin{bmatrix} T_1 & T_1^{-T} \\ 0 & T_1^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} S T_1^T & T_1^{-1} \\ 0 & -T_1 S \end{bmatrix}$$

$$= \begin{bmatrix} -I_n & 0 \\ 0 & I_n \end{bmatrix}$$

$$= J.$$
Hence $T$ is symplectic. It is left to prove that any symplectic matrix $T \in \mathbb{R}^{2n \times 2n}$ given in the form (2.1) is of the form (2.3). The matrix

$$C = \begin{bmatrix} I & T_1^{-1}T_2 \\ 0 & I \end{bmatrix} \tag{2.4}$$

is symplectic because it is the product of two symplectic matrices.

$$\begin{bmatrix} I & T_1^{-1}T_2 \\ 0 & I \end{bmatrix} = \begin{bmatrix} T_1^{-1} & 0 \\ 0 & T_1^T \end{bmatrix} \begin{bmatrix} T_1 & T_2 \\ 0 & T_1^{-T} \end{bmatrix}. \tag{2.5}$$

According to Table 2.1, $C$ is symplectic if and only if $T_1^{-1}T_2$ is symmetric. If we write $T_1^{-1}T_2 = S$ then $T_2 = T_1S$ where $S = S^T$. □

### 2.3 Logarithm of symplectic matrices

For a given $A \in \mathbb{C}^{n \times n}$ a logarithm of $A$ is any matrix $W$ such that $e^W = A$. We assume that $A$ has no eigenvalues on $\mathbb{R}^-$ so that the existence of a unique principal logarithm is assured as shown in the following theorem.

**Theorem 2.5.** [35, Thm. 1.31] Let $A \in \mathbb{C}^{n \times n}$ have no eigenvalues on $\mathbb{R}^-$. There is a unique logarithm $W$ of $A$ all of whose eigenvalues lie in the strip $\{z : -\pi < \text{Im}(z) < \pi\}$. We refer to $W$ as the principal logarithm of $A$ and write $W = \log A$. If $A$ is real then its principal logarithm is real.

A useful expression for the logarithm is the series expansion, which is given for a scalar $a \in \mathbb{C}$

$$\log(1 + a) = \sum_{k=1}^{\infty} (-1)^{k+1}\frac{a^k}{k}, \quad |a| < 1,$$

which is obtained by integrating $(1 + x)^{-1}$ between 0 and $a$. The corresponding series for matrices is

$$\log(I + A) = \sum_{k=1}^{\infty} (-1)^{k+1}\frac{A^k}{k}, \quad \rho(A) < 1,$$

where $\rho(A)$ denotes the spectrum of $A$, which is the set of all the eigenvalues of $A$.

The function logarithm maps a symplectic matrix to a Hamiltonian matrix. The question we need to deal with is whether the square root function and the Padé
approximants preserve the structure or not. Before we give the details of matrix square root and Padé approximation we need to give the structure of the matrix logarithm.

**Theorem 2.6.** [11, Thm. 2.1] If \( A \in \mathbb{R}^{2n \times 2n} \) is a symplectic matrix and \( \rho(A) \cap \mathbb{R}^- = \emptyset \) then \( \log A = W \) is Hamiltonian.

**Proof.** It is a particular case of (iii) on page 29. Since \( A \) is symplectic it satisfies

\[
A^T = -JA^{-1}J = J^{-1}A^{-1}J.
\]

Then,

\[
\log(A^T) = J^{-1} \log A^{-1}J.
\]

But, \( \log(A^T) = (\log A)^T \) as stated in Theorem 1.1-b) and \( \log A^{-1} = - \log A \) [35, Thm. 11.2]. Therefore,

\[
(\log A)^T = -J^{-1} \log A J
= J \log A J.
\]

The equality shows that \( \log A \) is Hamiltonian. \( \square \)

The basic idea of the inverse scaling and squaring method is to take the repetitive square roots of \( A \) so the result is close to the identity matrix and then use Padé approximation. We evaluate the Padé approximant \( r_m(A^{1/2^k} - I) \) to \( \log(A^{1/2^k}) \) and obtain

\[
\log A \approx 2^k r_m(A^{1/2^k} - I).
\]

**Algorithm 2.7.** [45] Given \( A \in \mathbb{C}^{n \times n} \) with no eigenvalues on \( \mathbb{R}^- \) this algorithm employs the inverse scaling and squaring method to compute \( W = \log A \).

1. Bring \( A \) close to an identity matrix by taking \( k \) repetitive square root of \( A \)
2. Decide the order of \( r_m(A^{1/2^k} - I) \) by minimising the cost and maximising the accuracy
3. Approximate \( \log(A^{1/2^k}) \) by using \( r_m(A^{1/2^k} - I) \approx \log(A^{1/2^k}) \)
4. Rescale to obtain \( W \approx 2^k r_m(A^{1/2^k} - I) \)
2.4 Matrix square root

A matrix $X \in \mathbb{C}^{n \times n}$ satisfying the equation $A = X^2$ is said to be a square root of $A$. Any matrix $A$ with no eigenvalues on $\mathbb{R}^-$ has a unique principal square root, defined in the next theorem.

**Theorem 2.8.** [35, Thm. 1.29] Let $A \in \mathbb{C}^{n \times n}$ have no eigenvalues on $\mathbb{R}^-$. There is a unique square root $X$ of $A$ all whose eigenvalues lie in the open right half-plane. We refer to $X$ as the principal square root of $A$ and write $X = A^{1/2}$. If $A$ is real then $X$ is real.

Next theorem shows that the principal square root of matrices in an automorphism group stays in the group.

**Theorem 2.9.** [63] Suppose $G_M$ is the automorphism group of an arbitrary scalar product and $A$ is a matrix that has a principal square root $A^{1/2}$. If $A \in G_M$ then $A^{1/2} \in G_M$.

**Proof.** $A \in G_M \Rightarrow A^* = A^{-1}$. Then,

$$(A^*)^{1/2} = (A^{-1})^{1/2} \Rightarrow (A^{1/2})^* = (A^{1/2})^{-1} \Rightarrow A^{1/2} \in G_M. \quad \Box$$

2.4.1 Structure preserving iterations

We present structure preserving iterations for the square root of matrices in Schur-like form and for the full symplectic matrices. Let $T \in \mathbb{R}^{2n \times 2n}$ and $A \in \mathbb{R}^{2n \times 2n}$ represent the matrix in the Schur-like form (2.2) and the full matrix, respectively. Before the first approach is stated, we need an algorithm to compute the principal square root of an upper quasi triangular matrix.

**Algorithm 2.10.** [35, Alg. 6.7] Given an upper quasi triangular matrix $T_1 \in \mathbb{R}^{n \times n}$ with no eigenvalues on $\mathbb{R}^-$ this algorithm computes $X_1 = \sqrt{T_1}$, where $\sqrt{\cdot}$ denotes any primary square root and $T_1$ is a block $m \times m$ matrix.
2.4. MATRIX SQUARE ROOT

\begin{align*}
1 & \quad (X_1)_{ii} = \sqrt{(T_1)_{ii}}, \quad i = 1:m \\
2 & \quad \text{for } j = 2:m \\
3 & \quad \text{for } i = j - 1: -1:1 \\
4 & \quad \text{Solve } (X_1)_{ii}(X_1)_{ij} + (X_1)_{ij}(X_1)_{jj} = (T_1)_{ij} - \sum_{k=i+1}^{j-1} (X_1)_{ik}(X_1)_{kj} \\
5 & \quad \text{for } (X_1)_{ij} \\
6 & \quad \text{end} \\
\end{align*}

As stated in Theorem 1.1 $X_1$ has the same block structure as $T_1$. The matrices $(X_1)_{ii}$ represent the diagonal blocks and $(X_1)_{ij}$ are the remaining blocks.

**First approach**

Let $T$ be given in the form (2.2). Since the square root of $T$ will be both in the same block triangular form as $T$ and symplectic, we can write

\[
\begin{bmatrix} T_1 & T_1S \\ 0 & T_1^{-T} \end{bmatrix} = \begin{bmatrix} X_1 & X_1G \\ 0 & X_1^{-T} \end{bmatrix} \begin{bmatrix} X_1 & X_1G \\ 0 & X_1^{-T} \end{bmatrix},
\]

where $X_1 \in \mathbb{R}^{n \times n}$ is an upper quasi triangular and $G \in \mathbb{R}^{n \times n}$ is a symmetric matrix.

We compute the square root of $T$ by solving the equations,

\begin{align*}
X_1^2 &= T_1, \\
X_1^2G + X_1GX_1^{-T} &= T_1S.
\end{align*}

We take the principal square root of the upper quasi triangular matrix $T_1$ by using Algorithm 2.10. Since the square root of $T$ is in the Schur-like form, it will be enough to take the inverse transpose of $X_1$ instead of taking one more square root for $X_1^{-T}$.

If we modify the equation (2.7b) by pre-multiplying with $X_1^{-1}$, we get

\[X_1G + GX_1^{-T} = X_1S.\]

This is a Sylvester equation and it is known that it has a unique solution if and only if $X_1$ and $-X_1^{-T}$ have no eigenvalue in common [43]. This is guaranteed for any nonsingular $T$ because when the principal square roots of $T_1$ and $T_1^{-T}$ are taken, their spectra lie in the open right half plane so $X_1$ and $-X_1^{-T}$ cannot have an eigenvalue
Table 2.2: The cost analysis of taking a square root of $T$ given in the Schur-like form [35, Table C.1].

<table>
<thead>
<tr>
<th>Equation</th>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1^2 = T_1$</td>
<td>$n^3/3$</td>
</tr>
<tr>
<td>Computing $X_1^{-T}$</td>
<td>$n^3/3$</td>
</tr>
<tr>
<td>$X_1G + GX_1^{-T} = X_1S$</td>
<td>$3n^3$</td>
</tr>
<tr>
<td>Form $X_1G$</td>
<td>$n^3$</td>
</tr>
</tbody>
</table>

in common. That is, while the spectrum of $X_1$ lies in the open right half plane, the spectrum of $-X_1^{-T}$ lies in the open left half plane. The function nag_lapack_dtrsysl from the NAG library [78] solves the Sylvester matrix equation $AX + XB = C$ for $X \in \mathbb{R}^{m \times n}$, where $A \in \mathbb{R}^{m \times m}$ and $B \in \mathbb{R}^{n \times n}$ are upper or lower quasi triangular matrices. In this computation, the number of floating point operations is approximately $mn(m+n)$.

To solve the Sylvester equation $X_1G + GX_1^{-T} = X_1S$ for the symmetric matrix $G$, we do not need to compute the Schur decomposition of $X_1$ and $X_1^{-T}$ because $X_1$ and $X_1^{-T}$ are upper quasi triangular and lower quasi triangular matrices, respectively. The multiplication $X_1S$ costs $n^3$ flops and solving equation (2.7b) costs $2n^3$ flops. So the total cost of solving equation (2.7b) is $3n^3$ flops in each square root. The cost analysis of taking a square root of $T$ is presented in Table 2.2. Next, we give the algorithm, including the solutions of (2.7a) and (2.7b), for taking a square root of $T$ in the form (2.2).

**Algorithm 2.11.** Given a symplectic matrix $T \in \mathbb{R}^{2n \times 2n}$ in the form (2.2) with no eigenvalues on $\mathbb{R}^-$ this algorithm computes $X = \sqrt{T}$, where $\sqrt{\cdot}$ denotes any primary square root.

1. Compute $X_1 = \sqrt{T_1}$ by using Algorithm 2.10
2. Compute $X_1^{-T}$
3. Solve the Sylvester equation $X_1G + GX_1^{-T} = X_1S$ for $G$
4. Form $X = \begin{bmatrix} X_1 & X_1G \\ 0 & X_1^{-T} \end{bmatrix}$

**Cost:** Taking a square root of $T$ costs $\frac{14}{3}n^3$ in total.
Second approach

In this approach, instead of using the Schur-like form, we give the structure preserving iterative method to compute the square root of full symplectic matrices.

**Theorem 2.12.** [38, Thm. 4.5] Suppose the matrix $A$ has no eigenvalues on $\mathbb{R}^-$, so that $A^{1/2}$ exists. Let $g$ be any matrix function of the form $g(X) = Xh(X^2)$ such that the iteration $X_{k+1} = g(X_k)$ converges to $\text{sign}(X_0)$ whenever $\text{sign}(X_0)$ is defined. Then in the coupled iteration

$$Y_{k+1} = Y_k h(Z_k Y_k), \quad Y_0 = A, \quad \text{(2.8a)}$$

$$Z_{k+1} = h(Z_k Y_k) Z_k, \quad Z_0 = I, \quad \text{(2.8b)}$$

where $Y_k \to A^{1/2}$ and $Z_k \to A^{-1/2}$ as $k \to \infty$, both with order of convergence $m$, $Y_k$ commutes with $Z_k$, and $Y_k = AZ_k$ for all $k$. Moreover, if $g$ is structure preserving for an automorphism group $G_M$, then iteration (2.8) is also structure preserving for $G_M$. That is, $A \in G_M$ implies $Y_k, Z_k \in G_M$ for all $k$.

Theorem 2.12 is applied to the structure preserving matrix sign iterations given in [37] to obtain structure preserving matrix square root iterations.

**Theorem 2.13.** [38, Thm. 6.1] Let $A \in \mathbb{R}^{2n \times 2n}$ and consider the iterations

$$Y_{k+1} = Y_k p_m(I - Z_k Y_k)[\text{rev } p_m(I - Z_k Y_k)]^{-1}, \quad Y_0 = A,$$

$$Z_{k+1} = p_m(I - Z_k Y_k)[\text{rev } p_m(I - Z_k Y_k)]^{-1} Z_k, \quad Z_0 = I,$$

where $p_m(t)$ is the numerator in the $[m/m]$ Padé approximant to $(1 - t)^{-1/2}$, $\text{rev } p_m(t)$ is the polynomial obtained by reversing the order of the coefficients of $p_m(t)$ and $m \geq 1$. Assume that $A$ has no eigenvalues on $\mathbb{R}^-$ and $A \in G_M$, where $G_M$ is an automorphism group. Then $Y_k \in G_M, Z_k \in G_M$ and $Y_k = AZ_k$ for all $k$, and $Y_k \to A^{1/2}, Z_k \to A^{-1/2}$ both with order of convergence $2m + 1$.

As in [38] $p_1(1 - x^2)$ in Theorem 2.13 is obtained by taking the odd part of $(1 + x)^3$ and dividing by $x$. Then cubically converging iteration is obtained as

$$Y_{k+1} = Y_k (3I + Z_k Y_k)(I + 3Z_k Y_k)^{-1}, \quad Y_0 = A, \quad \text{(2.9a)}$$

$$Z_{k+1} = (3I + Z_k Y_k)(I + 3Z_k Y_k)^{-1} Z_k, \quad Z_0 = I. \quad \text{(2.9b)}$$
Table 2.3: The cost analysis of structure preserving iterations for taking a square root of a full symplectic matrix $A \in \mathbb{R}^{2n \times 2n}$.

<table>
<thead>
<tr>
<th>Equation</th>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iteration (2.9)</td>
<td>$64n^3$</td>
</tr>
<tr>
<td>Iteration (2.10)</td>
<td>$176n^3/3$</td>
</tr>
</tbody>
</table>

By rearranging the formula we get

$$Y_{k+1} = \frac{1}{3} Y_k [I + 8(I + 3Z_k Y_k)^{-1}], \quad Y_0 = A,$$  
$$Z_{k+1} = \frac{1}{3} [I + 8(I + 3Z_k Y_k)^{-1}] Z_k, \quad Z_0 = I.$$  

(2.10a)  

(2.10b)

While iteration (2.9) needs 3 matrix multiplications and 1 matrix inversion per iteration, iteration (2.10) needs 1 matrix multiplication and 2 solutions of matrix equations involving coefficient matrices that are transposes of each other. The cost analysis of iteration (2.9) and iteration (2.10) are given in Table 2.3.

Third approach

In the third approach we exploit the symplecticity in each iteration of equation (2.10) by using the Schur-like form of a symplectic matrix $A$. Since the structure is preserved in each iteration of equation (2.10) $Y_k$ and $Z_k$ are obtained in the Schur-like form as

$$Z_k = \begin{bmatrix} H_1^{(k)} & H_2^{(k)} \\ 0 & H_1^{(k)-T} \end{bmatrix}, \quad Y_k = \begin{bmatrix} K_1^{(k)} & K_2^{(k)} \\ 0 & K_1^{(k)-T} \end{bmatrix}$$

where $H_1^{(k)} \in \mathbb{R}^{n \times n}$ and $K_1^{(k)} \in \mathbb{R}^{n \times n}$ are quasi upper triangular matrices. This helps us to reduce the cost of multiplication $Z_k Y_k$, which requires 3 multiplications and 1 matrix inversion. The cost analysis is given in Table 2.4. Iteration (2.10) also requires 2 solutions of matrix equations involving coefficient matrices that are transposes of each other. Since $Z_k$ and $Y_k$ are in the Schur-like form the cost of solving each of these matrix equations is $\frac{8}{3}n^3$ flops. As it is shown in Table 2.5 if we use the Schur-like form of the symplectic matrix in equation (2.10) the cost of each iteration in the square root is $8n^3$ flops.
Table 2.4: The cost of the multiplication $Z_k Y_k$ when $Z_k \in \mathbb{R}^{2n \times 2n}$ and $Y_k \in \mathbb{R}^{2n \times 2n}$ are in the Schur-like form.

<table>
<thead>
<tr>
<th>Equation</th>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1^{(k)} K_1^{(k)}$</td>
<td>$n^3/3$</td>
</tr>
<tr>
<td>$H_1^{(k)} K_2^{(k)}$</td>
<td>$n^3$</td>
</tr>
<tr>
<td>$H_2^{(k)} K_1^{(k)} - T$</td>
<td>$n^3$</td>
</tr>
<tr>
<td>$(H_1^{(k)} K_1^{(k)}) - T$</td>
<td>$n^3/3$</td>
</tr>
</tbody>
</table>

Table 2.5: The cost analysis of iteration (2.10) when $Y_k \in \mathbb{R}^{2n \times 2n}$ and $Z_k \in \mathbb{R}^{2n \times 2n}$ are in the Schur-like form.

<table>
<thead>
<tr>
<th>Equation</th>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z_k Y_k$</td>
<td>$8n^3/3$</td>
</tr>
<tr>
<td>$(3/8)(I + 3Z_k Y_k)^T Y_k^T$</td>
<td>$8n^3/3$</td>
</tr>
<tr>
<td>$(3/8)(I + 3Z_k Y_k)^T Z_k$</td>
<td>$8n^3/3$</td>
</tr>
</tbody>
</table>

2.5 Padé approximation

In the inverse scaling and squaring method we compute the diagonal Padé approximants $r_m(x)$ to $\log(1 + x)$ by using the partial fraction form, which is proved to be the best method overall in [33] in terms of accuracy and computational cost.

$$r_m(x) = \sum_{j=1}^{m} \frac{\alpha_j^{(m)} x}{1 + \beta_j^{(m)} x}, \quad (2.11)$$

where the $\alpha_j^{(m)}$ are the weights and the $\beta_j^{(m)}$ are the nodes of $m$-point Gauss Legendre quadrature rule on $[0, 1]$ [21]. It is important to note that the Padé approximation in the inverse scaling and squaring method preserves the structure of the symplectic matrices in the sense that the Padé approximation to the logarithm of the symplectic matrix is Hamiltonian. To prove this claim, we use the following theorem about the homographic invariance under the argument transformations.

**Theorem 2.14.** [42, Thm. 1.5.2] Let $x$ be a scalar and $f(x)$ have the power series $f(x) = \sum_{k=0}^{\infty} c_k x^k$. Define an origin-preserving linear fractional transformation of the argument

$$y = \frac{ax}{1 + bx},$$
2. SCHUR-LIKE FORM IN MATRIX LOGARITHM

and thereby a new function \( g(y) = f(x) \). Then

\[
  s_m(y) = h_m(x),
\]

where \( s_m(y) \) and \( h_m(x) \) are the Padé approximations to the given functions \( g(y) \) and \( f(x) \), respectively.

**Theorem 2.15.** [19] Let \( r_m(X) \) be the diagonal Padé approximants to \( \log(I + X) \), \( m = 0, 1, \ldots \). Let \( W = \log A, X = A - I \) with \( \rho(X) < 1 \). If \( A \) is symplectic then \( r_m(A - I) \) is Hamiltonian.

**Proof.** We will use the homographic invariance under the argument transformations for this proof. Since \( f(x) = \log x \) does not have a power series we take \( f(x) = \log(1 + x) \). By using the equality \( \log(1 + x) = -\log \left( 1 + \frac{-x}{x + 1} \right) \) and Theorem 2.14 we get \( r_m(x) = -r_m(-x/(x + 1)) \).

For the matrix case this formula yields \( r_m(X) = -r_m(-X(X + I)^{-1}) \). If \( A \) is a symplectic matrix then \( A^{-1} = -JA^*J \). We can write

\[
  r_m(A - I) = -r_m(A^{-1} - I) = -r_m(-JA^*J - I) = -r_m(-J(A^* - I)J).
\]

We get \( r_m(X) = -r_m(-J(X^*)J) = Jr_m(X^*)J = Jr_m(X)^*J \). The definition in Table 2.1 indicates that \( r_m(X) \) is Hamiltonian. \[\square\]

We state in the next theorem that the error in matrix Padé approximation is less than the error in scalar Padé approximation at the norm of the matrix.

**Theorem 2.16.** [44] For \( \|A - I\| < 1 \) and any subordinate matrix norm,

\[
  \|r_m(A - I) - \log A\| \leq |r_m(-\|A - I\|) - \log(1 - \|A - I\|)|. \tag{2.12}
\]

The maximal values \( \theta_m \) of \( \|A - I\| \) such that the error bound \( \|r_m(A - I) - \log A\| \) does not exceed \( u = 2^{-53} \approx 1.1 \times 10^{-16} \) are given in Table 2.6.

In the following section, we will use the structure preserving matrix square root iterations in the inverse scaling and squaring method.
Table 2.6: Maximal values $\theta_m$ of $\|A - I\|$ ensures that the bound $\|r_m(A - I) - \log A\|$ does not exceed $u = 2^{-53}$ [35, Table 11.1].

<table>
<thead>
<tr>
<th>m</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_m$</td>
<td>1.10e-5</td>
<td>1.82e-3</td>
<td>1.62e-2</td>
<td>5.39e-2</td>
<td>1.14e-1</td>
<td>1.87e-1</td>
<td>2.64e-1</td>
<td>3.40e-1</td>
<td>4.11e-1</td>
</tr>
<tr>
<td>m</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>$\theta_m$</td>
<td>4.75e-1</td>
<td>5.31e-1</td>
<td>5.81e-1</td>
<td>6.24e-1</td>
<td>6.62e-1</td>
<td>6.95e-1</td>
<td>7.24e-1</td>
<td>9.17e-1</td>
<td>9.78e-1</td>
</tr>
</tbody>
</table>

2.6 The inverse scaling and squaring method

Let $A$ have a matrix decomposition as $A = Q^*TQ$, where $T$ is in the Schur-like form and $Q$ is orthogonal-symplectic. Then we can compute $\log A$ by

$$\log A = Q^* \log TQ.$$ 

We consider three methods applying the inverse scaling and squaring method. In the first method we use the Schur-like form of the symplectic matrices $T$ given in the form (2.2) and Algorithm 2.11 to compute the matrix square root. In the second one we apply the method to the full form of the symplectic matrix $A$ and use the structure preserving iteration (2.10) for the square root. In the final approach we take $T$ in the Schur-like form (2.1) and we compute the square root by the same structure preserving iteration (2.10).

2.6.1 The first approach in the inverse scaling and squaring method

We adapt the algorithm [35, Alg. 11.9] to the symplectic matrix given in the Schur-like form (2.2). To compute $W = \log T$, we use the inverse scaling and squaring method where the square root is computed by using Algorithm 2.11, which costs $\frac{14}{3}n^3$ flops. To decide the number of square roots and the order of Padé approximants we use the cost of evaluating Padé approximation. The Padé approximants $r_m(T^{1/2} - I)$ are evaluated by using the partial fraction expansion which costs $\frac{8}{3}mn^3$ flops [33], [35, Table C.1]. When we compare the cost of the square root with the cost of Padé approximation, it is only worth taking a square root if it reduces the order of Padé approximation by
at least two. That is, while a two order decrease in the order of Padé approximation
provides $\frac{16}{3}n^3$ flops cost reduction, taking an additional square root costs $\frac{14}{3}n^3$ flops.

Taking a square root of $T$ approximately reduces the distance of $T^{1/2k}$ to the
identity matrix by a half. This is easy to see since

$$(I - T^{1/2k+1})(I + T^{1/2k+1}) = I - T^{1/2k}$$

and $T^{1/2k} \to I$ as $k \to \infty$ it gives

$$\|I - T^{1/2k+1}\| \approx \frac{1}{2}\|I - T^{1/2k}\|.$$  \hspace{1cm} (2.13)

This approximation shows that we need to check the value $\tau = \|I - T^{1/2k}\|$ and take
the square root of $T$ until $\tau \leq \theta_7$, because each square root of $T$ reduces the order
of the Padé approximation by at least 2 for $m > 7$; that is, $\theta_m/2 \leq \theta_{m-2}$ for $m > 7$.
After we obtain $\tau \leq \theta_7$, we need to check the cost reduction for $m = 3: 7$. If taking an
extra square root decreases the order of Padé approximation for at least two, we can
take an extra square root. So we check the value $m_1 - m_2$, where $m_1$ and $m_2$ are the
orders of Padé approximations before and after the extra square root, respectively. If
$m_1 - m_2 \geq 2$ is not satisfied, we need to limit the number of extra square root by one,
since taking two extra square roots costs $\frac{28}{3}n^3$ flops which exceeds the cost of Padé
approximation even if it reduces the order of Padé approximation by two.

We give the algorithm of the inverse scaling and squaring method using the Schur-
like form of symplectic matrices given in the form (2.2) as follows.

\textbf{Algorithm 2.17.} Given a symplectic matrix $T \in \mathbb{R}^{2n \times 2n}$ in the Schur-like form (2.2)
with no eigenvalues on $\mathbb{R}^-$ this algorithm computes $W = \log T$ by using the inverse
scaling and squaring method and the constants $\theta_m$ given in Table 2.6. It uses Algo-
rithm 2.11 to compute the square root of $T$. This algorithm is intended for IEEE double
precision arithmetic.

1. $k = 0$, $p = 0$
2. while true
3. \hspace{1cm} $\tau = \|T - I\|_1$
4. \hspace{1cm} if $\tau \leq \theta_7$
2.6. THE INVERSE SCALING AND SQUARING METHOD

5 \[ p = p + 1 \]
6 \[ m_1 = \min\{i : \tau \leq \theta_i, \ i = 3:7\} \]
7 \[ m_2 = \min\{i : \tau/2 \leq \theta_i, \ i = 3:7\} \]
8 if \( m_1 - m_2 < 2 \) or \( p = 2 \), \( m = m_1 \) and go to line 13, end
9 end
10 \( T \leftarrow T^{1/2} \) by using Algorithm 2.11
11 \( k = k + 1 \)
12 end
13 Evaluate \( Y = r_m(T - I) \)
14 \( W = 2^kY \)

Here \( m_1 \) and \( m_2 \) are the orders of Padé approximations chosen before and after the extra square root, respectively.

**Cost:** Taking a square root of \( T \) costs \( \frac{14}{3}n^3k \) flops and evaluating the partial fraction form of the Padé approximation costs \( \frac{8}{3}mn^3 \) flops, where \( k \) is the number of square roots and \( m \) is the order of Padé approximation. It is \( \frac{14}{3}n^3k + \frac{8}{3}mn^3 \) flops in total.

2.6.2 The second approach in the inverse scaling and squaring method

The idea behind this approach is similar to the idea given in [35, Alg. 11.10]. However, instead of using the scaled product of Denman-Beavers iteration to take a square root we use the structure preserving iteration. We also use the full form of the symplectic matrices. The cost analysis in Table 2.3 suggests that we use iteration (2.10). Let \( it_j \) be the number of iterations required in each square root. The total cost of taking a square root is \( (\sum_{j=1}^k it_j) \frac{176}{3}n^3 \) flops, where \( k \) is the number of square roots taken in the inverse scaling and squaring method. The Padé approximation costs \( \frac{64}{3}mn^3 \) flops [33], [35, Table C.1], where \( m \) is the order of Padé approximation. The number of iterations required to take a square root of \( A \) typically changes from 16 on the first iterations to 4 for the last few iterations. As seen from Table 2.3 the cost of taking a square root of \( A \) at the last few iterations is \( \frac{704}{3}n^3 \) flops. It is only worth taking an extra square root if it reduces the order of Padé approximation by at least 12. By
using the idea given in equation (2.13), we will take a square root until \( \tau \leq \theta_{16} \) where \( \tau = \| A^{1/2} - I \| \), because decreasing order of Padé approximation by at least 12 can be obtained when \( \tau > \theta_{16} \). We will take only one more square root after \( \tau < \theta_{16} \) is obtained because the cost of two more extra square roots exceeds the cost of the Padé approximation. The following inequality indicates that the reduction in the order of the Padé approximation does not provide enough cost reduction compared with the cost of taking a square root:

\[
\frac{64}{3} (m_1 - m_2) n^3 \leq \frac{176}{3} n^3 \Delta 
\]

When this inequality is satisfied we take an extra square root. Since the cost of taking 2 more extra square roots exceeds the cost of evaluation the Padé approximation we limit the number of extra square root by taking \( p = 2 \) in the algorithm.

Algorithm 2.18. Given a full symplectic matrix \( A \in \mathbb{R}^{2n \times 2n} \) with no eigenvalues on \( \mathbb{R}^- \) this algorithm computes \( W = \log A \) by the inverse scaling and squaring method. It uses the constants \( \theta_m \) given in Table 2.6 and iteration (2.10) to take a square root of \( A \). This algorithm is intended for IEEE double precision arithmetic.

1. \( k = 0, \ p = 0 \)
2. while true
3. \( \tau = \| A - I \|_1 \)
4. if \( \tau < \theta_{16} \)
5. \( p = p + 1 \)
6. \( m_1 = \min\{i : \tau \leq \theta_i, \ i = 3:16\} \)
7. \( m_2 = \min\{i : \tau/2 \leq \theta_i, \ i = 3:16\} \)
8. if \( 4(m_1 - m_2)/11 \leq \Delta \) or \( p = 2, \ m = m_1 \), go to line 13, end
9. end
10. \( A \leftarrow A^{1/2} \) by using iteration (2.10)
11. \( k = k + 1 \)
12. end
13. Evaluate \( Y = r_m(A - I) \)
14. \( W = 2^k Y \)
**2.6. THE INVERSE SCALING AND SQUARING METHOD**

**Cost:** Taking a square root costs \((\sum_{j=1}^{k} it_j)^{176/3} n^3\) flops where \(k\) is the number of square roots and evaluating the partial fraction form of Padé approximation costs about \(\frac{64}{3} mn^3\) flops. It is \((\sum_{j=1}^{k} it_j)^{176/3} n^3 + \frac{64}{3} mn^3\) flops in total.

**2.6.3 The third approach in the inverse scaling and squaring method**

In the second approach, iteration (2.10) is used to compute the square root of full symplectic matrix in the inverse scaling and squaring method. Here we apply the iteration to take a square root of symplectic matrices given in the Schur-like form (2.1) and exploit the symplecticity in each iteration. Let \(it_j\) be the number of iterations required in each square root. From Table 2.5 the total cost of taking a square root is \((\sum_{j=1}^{k} it_j)8n^3\) flops, where \(k\) is the number of square roots taken in the inverse scaling and squaring method. Since the matrix is given in the Schur-like form evaluating the partial fraction form of the Padé approximation with the order \(m\) costs \(\frac{8}{3} mn^3\) flops. The number of iterations required to take a square root of \(T\) typically changes from 16 on the first iterations to 4 for the last few iterations. So the cost of taking a square root of symplectic matrix \(T\) at the last few iterations is \(32n^3\) flops. It is worth only taking one more square root if it reduces the order of Padé approximation by at least 13. That decrease in the order of Padé approximation can only be obtained when \(\|T^{1/2^s} - I\| > \theta_{16}\), where \(\theta_{16}\) is the value given in Table 2.6. We will use the similar idea given in equation (2.13). When \(\|T^{1/2^s} - I\| \leq \theta_{16}\) is obtained, in order to compare the total cost we check the inequality

\[
\frac{8}{3}(m_1 - m_2)n^3 \leq 8n^3it_j
\]  

(2.14)

by assuming the same number of iterations is required. In equation (2.14) \(m_1\) and \(m_2\) are the order of Padé approximants before and after the extra square root, respectively. Since the cost of taking 2 more extra square roots exceeds the cost of evaluating the Padé approximant we limit it by taking \(p = 2\). That is, only one extra square root is taken if it is required. We give the algorithm of the inverse scaling and squaring method using the Schur-like form of the symplectic matrix (2.1) with iteration (2.10).
Algorithm 2.19. Given a symplectic matrix $T \in \mathbb{R}^{2n \times 2n}$ in the Schur-like form (2.1) with no eigenvalues on $\mathbb{R}^-$ this algorithm computes $W = \log T$ by the inverse scaling and squaring method. It uses the constants $\theta_m$ given in Table 2.6 and iteration (2.10) to take a square root of $T$. This algorithm is intended for IEEE double precision arithmetic.

1. $k = 0, p = 0$
2. while true
3. $\tau = \|T - I\|_1$
4. if $\tau < \theta_{16}$
5. $p = p + 1$
6. $m_1 = \min\{i : \tau \leq \theta_i, \ i = 3:16\}$
7. $m_2 = \min\{i : \tau/2 \leq \theta_i, \ i = 3:16\}$
8. if $(m_1 - m_2)/3 \leq it_j$ or $p = 2, m = m_1$, go to line 13, end
9. end
10. $T \leftarrow T^{1/2}$ by using iteration (2.10)
11. $k = k + 1$
12. end
13. Evaluate $Y = r_m(T - I)$
14. $W = 2^k Y$

**Cost:** Taking a square root costs $(\sum_{j=1}^k it_j)8n^3$ flops where $k$ is the number of square roots and evaluating the partial fraction form of Padé approximation costs about $\frac{8}{3}mn^3$ flops. It is $(\sum_{j=1}^k it_j)8n^3 + \frac{8}{3}mn^3$ flops in total.

### 2.7 Error measure

Let $A \in \mathbb{G}_M$ be given. The characterization of the adjoint matrix $A$ is given as $A^* = A^{-1}$. The analysis to measure the departure from $\mathbb{G}$ structure that is less dependent on the conditioning of $A$ is given in [37, p. 8]. We rewrite the relation as $A^*A = I$. Consider $A + \Delta A$, where $A$ is a symplectic matrix and $\|\Delta A\|_2 \leq \epsilon\|A\|_2$. For the symplectic matrix, since $A^* = -JA^*J$ and $J$ is unitary, $\|A^*\|_2 = \|A\|_2$ for all
2.7. ERROR MEASURE

\( A \) and hence,

\[
\|(A + \Delta A)^*(A + \Delta A) - I\|_2 = \|A^*\Delta A + \Delta A^*A + \Delta A^*\Delta A\|_2 \\
\leq 2\|A\|_2\|\Delta A\|_2 + \|\Delta A\|_2^2 \\
\leq (2\epsilon + \epsilon^2)\|A\|_2^2.
\]

We get

\[
\frac{\|(A + \Delta A)^*(A + \Delta A) - I\|_2}{\|A\|_2^2} \leq (2\epsilon + \epsilon^2).
\]

Equation (2.15) shows that the relative measure of departure from \( G \) structure can be measured by

\[
\text{err}_G(A) = \frac{\|A^*A - I\|_2}{\|A\|_2^2}.
\]

To measure the structure error for the matrices in Lie algebra, we will use the characterization of an adjoint matrix in Lie algebra which is \( W^* = -W \). Consider \( W + \Delta W \), where \( W \in L \) and \( \|\Delta W\|_2 \leq \epsilon\|W\|_2 \) for all \( W \). Since \( W^* = JWJ \) and \( J \) is unitary \( \|W^*\|_2 = \|W\|_2 \) for all Hamiltonian matrices. So we get

\[
\|(W + \Delta W)^* + (W + \Delta W)\|_2 = \|\Delta W^* + \Delta W\|_2 \\
\leq \|\Delta W^*\|_2 + \|\Delta W\|_2 \\
\leq 2\epsilon\|W\|_2.
\]

This gives

\[
\frac{\|(W + \Delta W)^* + (W + \Delta W)\|_2}{\|W\|_2} \leq 2\epsilon.
\]

This inequality shows that the appropriate relative measure of departure from \( L \) structure can be computed by

\[
\text{err}_L(W) = \frac{\|W^* + W\|_2}{\|W\|_2}.
\]

The relative error for the computed logarithm \( \hat{W} \) is given by

\[
\text{rel}_\text{err}(\hat{W}) = \frac{\|\hat{W} - W\|_2}{\|W\|_2},
\]

where \( W = \log A \) is the “exact” logarithm. In the next lemma we state the relationship between the relative error and the departure from the Lie algebra structure. Similar analysis for the matrix square root is given in [38, Lem. 7.1].
Lemma 2.20. Assume that \( A \) has no eigenvalues on \( \mathbb{R}^- \). If \( A \in G_M \) then \( \log A \in \mathbb{L}_M \), where \( G_M \) is the automorphism group and \( \mathbb{L}_M \) is the Lie algebra of any scalar product, respectively. Then for a computed matrix logarithm \( \hat{W} \in \mathbb{R}^{2n \times 2n} \) sufficiently close to \( W = \log A \) it holds that

\[
\text{err}_{\mathbb{L}_M}(\hat{W}) \leq 2\text{rel}_\text{err}(\hat{W}) + O(\text{rel}_\text{err}(\hat{W})^2). \tag{2.19}
\]

Proof. Let \( \hat{W} = W + \Delta W \) where \( W = \log A \).

\[
\hat{W}^* + \hat{W} = (W + \Delta W)^* + W + \Delta W
= \Delta W^* + \Delta W.
\]

Since \( \Delta W^* = J\Delta W^* J \) and \( J \) is unitary we have \( \|\Delta W^*\| = \|\Delta W\| \). Then we get

\[
\|\hat{W}^* + \hat{W}\| \leq 2\|\Delta W\|.
\]

Multiplying this by \( \|\hat{W}\|^{-1} = \|W\|^{-1} + O(\|\Delta W\|) \) we obtain,

\[
\frac{\|\hat{W}^* + \hat{W}\|}{\|\hat{W}\|} \leq 2 \frac{\|\hat{W} - \log A\|}{\|\log A\|} + O(\text{rel}_\text{err}(\hat{W})^2). \tag*{\Box}
\]

2.8 Numerical experiments

In our experiments we will compare using the Schur decomposition or the full form of the matrices against the Schur-like form for symplectic matrices in the inverse scaling and squaring method. The experiments are carried out in MATLAB R2013a with \( u = 1.1 \times 10^{-16} \). In the result part, the relative error of the logarithm of the symplectic matrices and the structural error are presented. \( \hat{W} \) represents the computed logarithm and \( W = \log A \) represents the “exact” logarithm computed by using Symbolic Math Toolbox. \( A \) is diagonalized in 100 digit precision as \( A = VDV^{-1} \) and the logarithm is computed by \( \log A = V \log DV^{-1} \). In the experiments the relative condition number \( \text{cond}_{\log}(A) \) of the matrix logarithm is computed by using equation (1.4), in which the Fréchet derivative is computed by \texttt{logm_frechet_pade} from Higham’s Matrix Function Toolbox [30].


2.8. NUMERICAL EXPERIMENTS

Table 2.7: The analysis of Algorithm 11.9 and Algorithm 11.10 from [35, Alg. 11.9, Alg. 11.10] in terms of structure preservation and the accuracy.

<table>
<thead>
<tr>
<th>$\text{condlog}(A)$</th>
<th>$\text{relerr}(\hat{W})$</th>
<th>$\text{err}_{\text{rel}}(\hat{W})$</th>
<th>$\text{relerr}(\hat{W})$</th>
<th>$\text{err}_{\text{rel}}(\hat{W})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.8e+2</td>
<td>2.0e-14</td>
<td>3.0e-14</td>
<td>1.1e-14</td>
<td>1.2e-14</td>
</tr>
<tr>
<td>5.5e+4</td>
<td>2.0e-12</td>
<td>2.3e-12</td>
<td>2.6e-12</td>
<td>2.7e-12</td>
</tr>
<tr>
<td>1.9e+7</td>
<td>2.2e-10</td>
<td>4.7e-11</td>
<td>1.2e-10</td>
<td>1.1e-10</td>
</tr>
<tr>
<td>1.4e+8</td>
<td>2.5e-10</td>
<td>2.6e-9</td>
<td>1.7e-10</td>
<td>2.6e-9</td>
</tr>
<tr>
<td>1.0e+9</td>
<td>6.1e-9</td>
<td>1.9e-8</td>
<td>1.0e-8</td>
<td>1.5e-8</td>
</tr>
<tr>
<td>1.9e+11</td>
<td>2.2e-6</td>
<td>3.2e-6</td>
<td>6.6e-7</td>
<td>1.5e-6</td>
</tr>
<tr>
<td>1.1e+12</td>
<td>3.4e-5</td>
<td>3.5e-5</td>
<td>6.6e-5</td>
<td>4.5e-5</td>
</tr>
</tbody>
</table>

Experiment 1

We use full symplectic test matrices $A \in \mathbb{R}^{24 \times 24}$ formed by using the function `rand_resymp` from Jagger’s MATLAB Toolbox [41]. We compute the logarithm of the symplectic matrices by using Algorithm 11.9 and Algorithm 11.10 [35, Alg. 11.9, Alg. 11.10]. Algorithm 11.9 is computing the logarithm of the symplectic matrices by using the Schur decomposition and using Algorithm 2.10 for the square roots. Algorithm 11.10 applies the inverse scaling and squaring method to the full symplectic matrices without a Schur decomposition and it uses the scaled product Denman-Beavers iteration to take a square root. The results for these algorithms are shown in Table 2.7. The test results in Table 2.7 show that the computation of matrix logarithm by using the Schur decomposition or without Schur decomposition is not a problem for the well conditioned symplectic matrices. However, for the ill-conditioned matrices these algorithms cause a structure loss.

Experiment 2

In this experiment, the Schur-like form of the symplectic matrices is used. The steps to build the test matrices are stated as follows.

- Build the Schur-like form as

$$
T = \begin{bmatrix}
T_1 & T_1 S \\
0 & T_1^{-T}
\end{bmatrix}
$$
Table 2.8: The logarithm of the symplectic matrices is computed by Algorithm 2.17 and the square root is evaluated by Algorithm 2.11.

<table>
<thead>
<tr>
<th>cond_{log}(A)</th>
<th>err_{G}(A)</th>
<th>err_{L}(\hat{W})</th>
<th>rel_{err}(\hat{W})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1e+03</td>
<td>5.7e-16</td>
<td>3.0e-15</td>
<td>4.6e-14</td>
</tr>
<tr>
<td>3.5e+04</td>
<td>6.4e-16</td>
<td>1.9e-15</td>
<td>1.5e-13</td>
</tr>
<tr>
<td>8.1e+05</td>
<td>3.6e-16</td>
<td>2.6e-15</td>
<td>1.2e-12</td>
</tr>
<tr>
<td>9.3e+06</td>
<td>2.1e-16</td>
<td>3.4e-15</td>
<td>2.6e-11</td>
</tr>
<tr>
<td>3.0e+08</td>
<td>5.5e-16</td>
<td>4.1e-15</td>
<td>1.4e-09</td>
</tr>
<tr>
<td>2.4e+09</td>
<td>3.1e-16</td>
<td>2.1e-15</td>
<td>4.0e-08</td>
</tr>
<tr>
<td>2.4e+11</td>
<td>4.8e-16</td>
<td>2.9e-15</td>
<td>3.4e-07</td>
</tr>
<tr>
<td>5.7e+12</td>
<td>1.9e-16</td>
<td>2.6e-15</td>
<td>7.1e-05</td>
</tr>
</tbody>
</table>

where $T_1 \in \mathbb{R}^{n \times n}$ is quasi upper triangular and $S \in \mathbb{R}^{n \times n}$ is symmetric.

- Generate symplectic-orthogonal matrix $Q$ by taking the orthogonal factor of the polar decomposition of a symplectic matrix and build $A \in \mathbb{R}^{24 \times 24}$ as

$$A = QTQ^T.$$  

We test Algorithm 2.17 in terms of structure preservation and the accuracy of the logarithm of the symplectic matrices. The test results of Algorithm 2.17 are presented in Table 2.8 with the relative condition number of matrix logarithm $\text{cond}_{log}(A)$ of $A$. The results presented in Table 2.8 show that while the condition number affects the rounding error, structure loss is not observed even for the ill-conditioned matrices.

**Experiment 3**

In this experiment, we use the full symplectic matrices $A \in \mathbb{R}^{24 \times 24}$ generated by using $\text{rand}_\text{rsymp}$ from Jagger’s MATLAB Toolbox. Algorithm 2.18 is tested in terms of structure preservation and the accuracy. Table 2.9 shows that even though using iteration (2.10) in the inverse scaling and squaring method preserves the structure of the symplectic matrix, there is a structure loss for the ill-conditioned matrices. As we are using the full form of the symplectic matrices, this method is relatively expensive.
Table 2.9: The logarithm of the symplectic matrices is computed by using Algorithm 2.18 and the square root is evaluated by iteration (2.10).

<table>
<thead>
<tr>
<th>cond$_{log}(A)$</th>
<th>err$_{G}(A)$</th>
<th>err$_{L}(\hat{W})$</th>
<th>rel$_{err}(\hat{W})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1e+03</td>
<td>5.7e-16</td>
<td>1.8e-14</td>
<td>2.5e-14</td>
</tr>
<tr>
<td>3.5e+04</td>
<td>6.4e-16</td>
<td>1.9e-11</td>
<td>1.9e-11</td>
</tr>
<tr>
<td>8.1e+05</td>
<td>3.6e-16</td>
<td>5.2e-10</td>
<td>5.2e-10</td>
</tr>
<tr>
<td>9.3e+06</td>
<td>2.1e-16</td>
<td>7.1e-09</td>
<td>7.5e-09</td>
</tr>
<tr>
<td>3.0e+08</td>
<td>5.5e-16</td>
<td>7.2e-08</td>
<td>6.7e-08</td>
</tr>
<tr>
<td>2.4e+09</td>
<td>3.1e-16</td>
<td>4.3e-06</td>
<td>4.0e-06</td>
</tr>
<tr>
<td>2.4e+11</td>
<td>4.8e-16</td>
<td>2.2e-04</td>
<td>2.2e-04</td>
</tr>
<tr>
<td>5.7e+12</td>
<td>1.9e-16</td>
<td>1.4e-02</td>
<td>1.4e-02</td>
</tr>
</tbody>
</table>

Experiment 4

We test Algorithm 2.19 which uses the Schur-like form of the symplectic matrix (2.1) and iteration (2.10) to compute the square root. The test matrices $A \in \mathbb{R}^{24 \times 24}$ are built by using the same steps in Experiment 2. Algorithm 2.19 is tested in terms of the accuracy and the structure preservation of the logarithm of the symplectic matrix. The test results are presented in Table 2.10. Unlike Experiment 3 we do not observe a structure loss even for the ill conditioned matrices. Using the Schur-like form of the symplectic matrices (2.1) in the inverse scaling and squaring method with iteration (2.10) instead of the full forms helps us to reduce the total cost.

Table 2.10: The logarithm of the symplectic matrices given in the Schur-like form is computed by Algorithm 2.19 and the square root is evaluated by iteration (2.10).

<table>
<thead>
<tr>
<th>cond$_{log}(A)$</th>
<th>err$_{G}(A)$</th>
<th>err$_{L}(\hat{W})$</th>
<th>rel$_{err}(\hat{W})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1e+03</td>
<td>5.7e-16</td>
<td>8.5e-14</td>
<td>8.2e-14</td>
</tr>
<tr>
<td>3.5e+04</td>
<td>6.4e-16</td>
<td>2.8e-12</td>
<td>3.2e-12</td>
</tr>
<tr>
<td>8.1e+05</td>
<td>3.6e-16</td>
<td>6.0e-12</td>
<td>6.1e-12</td>
</tr>
<tr>
<td>9.3e+06</td>
<td>2.1e-16</td>
<td>7.7e-12</td>
<td>2.6e-11</td>
</tr>
<tr>
<td>3.0e+08</td>
<td>5.5e-16</td>
<td>4.5e-12</td>
<td>1.4e-09</td>
</tr>
<tr>
<td>2.4e+09</td>
<td>3.1e-16</td>
<td>1.6e-12</td>
<td>4.0e-08</td>
</tr>
<tr>
<td>2.4e+11</td>
<td>4.8e-16</td>
<td>1.7e-11</td>
<td>3.4e-07</td>
</tr>
<tr>
<td>5.7e+12</td>
<td>1.9e-16</td>
<td>1.5e-11</td>
<td>7.1e-05</td>
</tr>
</tbody>
</table>
2. SCHUR-LIKE FORM IN MATRIX LOGARITHM

Experiment 5

In the final experiment Algorithm 2.17, 2.18 and 2.19 are compared in terms of structure loss and the accuracy. In the legends schur-like, schur-like-it and full-form represent the results from Algorithm 2.17, Algorithm 2.19 and Algorithm 2.18, respectively. Figure 2.1 (a) shows the structure loss in the Hamiltonian matrix \( \hat{W} \) ordered by increasing order. Figure 2.1 (b) presents the relative error and the solid line is \( \text{cond}_{\log}(A)u \), where \( u \) is the machine precision. Both figures illustrate the advantage of using the Schur-like form in the inverse scaling and squaring method.

The numerical experiments show that the best structure and the best accuracy for the logarithm of the symplectic matrices are obtained by using the Schur-like form of the symplectic matrices instead of using the Schur decomposition. Moreover, using the Schur-like form is cheaper comparing to using the full symplectic matrices in terms of the computational cost of the matrix square root in the inverse scaling and squaring method. It is worth noting that although Algorithm 2.17 and Algorithm 2.19 are using the Schur-like form of the symplectic matrices given in the form (2.2) and (2.1), respectively, Algorithm 2.17 preserves the structure better.
Error Analysis for Functions of Nearly Structured Matrices

For a given matrix $A$, let $\mathcal{S}$ be a class of structured matrices and let $\min\{\|A - X\| : X \in \mathcal{S}\}$ be small. Our aim is to approximate $f(A)$ by $f(X)$ and investigate bounds for $\|f(A) - f(X)\|$, where $X$ achieves the above minimum. We analyse the error bound for the cases where $\mathcal{S}$ is equal to the class of unitary, Hermitian and Hermitian positive semidefinite matrices.

The reason why we want to compute $f(X)$ instead of $f(A)$ is the advantage of the structure preserving algorithms in terms of accuracy and storage. In the literature computing the functions of structured matrices has been studied by many researchers. The logarithm of symplectic matrices can be found in [19], [20]. Another study [15] considers the logarithm of unitary matrices. Similarly, the logarithm of matrices in an automorphism group is examined in [10]. Structure preserving iterations are investigated in [38] for the matrix square root and matrix sign functions.

The chapter is organised as follows. In Section 3.1 we bound $\|f(A) - f(X)\|$ for the matrix logarithm and the square root of nearly unitary matrices. The same analysis is presented for the logarithm of nearly Hermitian matrix in Section 3.2. The final Section 3.3 focuses on nearly Hermitian positive semi-definite matrices and presents error bounds for the exponential and trigonometric functions.
3. FUNCTIONS OF NEARLY STRUCTURED MATRICES

3.1 Nearly unitary matrices

We first state the following lemma to bound \( \| f(A) - f(X) \| \) for the general case.

**Lemma 3.1.** Let \( S \) be the class of structured matrices and \( X \in S \) be the nearest structured matrix to \( A \). Then we have

\[
\| f(A) - f(X) \| \leq \| L_f(X) \| \| A - X \| + O(\| A - X \|^2),
\]

where \( \| L_f(X) \| = \max_{E \neq 0} \frac{\| L(X, E) \|}{\| E \|} \).

**Proof.** By the definition of Fréchet derivative we have

\[
f(A) = f(X + (A - X)) = f(X) + L_f(X, A - X) + O(\| A - X \|^2),
\]

so that

\[
f(A) - f(X) = L_f(X, A - X) + O(\| A - X \|^2).
\]

For any unitarily invariant norm, we can write

\[
\| f(A) - f(X) \| = \| L_f(X) (A - X) \| + O(\| A - X \|^2)
\leq \| L_f(X) \| \| A - X \| + O(\| A - X \|^2).
\]

Let \( S \) be the class of unitary matrices and let that matrix \( A \) be such that \( \min \{ \| A - U \| : U \in S \} \) is small, i.e., \( A \) is nearly unitary. To compute the logarithm of nearly unitary matrices, Loring [61] suggests computing an approximate unitary polar factor \( U \) of the nearly unitary matrix \( A \) and then approximating \( \log A = \log U \). The backward error bound \( \| e^W - U \| \) is analysed by Loring in [61]. Our aim is to bound the forward error \( \| f(A) - f(U) \| \) for the nearly unitary matrix \( A \).

For a given matrix \( A \in \mathbb{C}^{n \times n} \), the equation \( \log A = W \) has many solutions. However, we are only interested in the principal logarithm which is given in Theorem 2.5. We are interested in the logarithm of unitary matrices so it is worth giving the structure of the logarithm of unitary matrices.

**Theorem 3.2.** If \( A \in \mathbb{C}^{n \times n} \) is a unitary matrix and \( \rho(A) \cap \mathbb{R}^- = \emptyset \), where \( \rho(A) \) is the set of eigenvalues of \( A \), then \( \log A \) is skew-Hermitian.
3.1. NEARLY UNITARY MATRICES

Proof. Since $A$ is unitary it satisfies $A^* = A^{-1}$. Then,

$$\log(A^*) = \log A^{-1}. $$

We know that $\log(A^*) = (\log A)^*$ and $\log A^{-1} = -\log A$ \cite[Thm. 11.2]{35}. Therefore

$$(\log A)^* = -\log A. $$

The equality shows that $\log A$ is skew-Hermitian.

We obtain the nearest matrix to $A$ with orthonormal columns by polar decomposition, which is stated in the next theorem.

**Theorem 3.3.** \cite{22} Let $A \in \mathbb{C}^{n \times n}$ and $A = UH$ where $U$ is unitary and $H$ is Hermitian positive semidefinite. Then for any unitary matrix $Z \in \mathbb{C}^{n \times n}$

$$||A - U|| \leq ||A - Z|| \leq ||A + U|| $$

(3.5)

holds for every unitarily invariant norm.

So if we are happy with a first order bound then for the nearest unitary matrix $U$ the problem reduces to bounding $||A - U||$ and $||L_f(U)||$ as Lemma 3.1 states. Since $U$ is unitary by using \cite[Corol. 3.16]{35} we can write

$$\|L_f(U)\|_F = \max_{i,j} |f[\mu_i, \mu_j]|, $$

(3.6)

where $\mu_i$ are the eigenvalues of $U$. Here, the divided difference $f[\mu_i, \mu_j]$ is given as

$$f[\mu_i, \mu_j] = \begin{cases} 
\frac{f(\mu_i) - f(\mu_j)}{\mu_i - \mu_j}, & \mu_i \neq \mu_j, \\
 f'(\mu_i), & \mu_i = \mu_j.
\end{cases} $$

(3.7)

To bound $||A - U||$ we will use the following theorem.

**Theorem 3.4.** \cite[Thm. 8.17]{35} Let $A \in \mathbb{C}^{m \times n}$ have the polar decomposition $A = UH$. Then for any unitarily invariant norm

$$\|A^*A - I\|_{1 + \sigma_1(A)} \leq ||A - U|| \leq \|A^*A - I\|_{1 + \sigma_n(A)}, $$

(3.8)

where $\sigma_1(A)$ and $\sigma_n(A)$ are the largest and the smallest singular values of $A$, respectively.
Proof. By using the inequality \( \|ABC\| \leq \|A\|_2\|B\|\|C\|_2 \), which is a corollary of the theorem in [40, Thm. 7.4.8.4] and the equality \( A^*A - I = (A - U)^*(A + U) \) we get

\[
\|A^*A - I\| = \|(A - U)^*(A + U)\|
\leq \|I\|_2\|(A - U)^*\||U(H + I)\|_2 \\
\leq \|A - U\|\|(H + I)\|_2 \\
\leq \|A - U\|(1 + \sigma_1(A)),
\]

which provides the lower bound in (3.8).

Since \( A^*A - I = (A - U)^*U(H + I) \) we have

\[(A^*A - I)(H + I)^{-1} = (A - U)^*U.\]

Then by using the same inequality \( \|ABC\| \leq \|A\|_2\|B\|\|C\|_2 \),

\[
\|(A - U)^*U\| = \|A - U\| \leq \|I\|_2\|A^*A - I\|\|(H + I)^{-1}\|_2 \\
\leq \|A^*A - I\|(1 + \sigma_n(A))^{-1}.
\]

Since the singular values of \( A \) are the eigenvalues of \( H \) we can get the upper bound

\[
\|A - U\| \leq \frac{\|A^*A - I\|}{1 + \sigma_n(A)}.
\]

\( \Box \)

So if we substitute equation (3.6) and the upper bound of (3.8) in equation (3.1) then we obtain

\[
\|f(A) - f(U)\|_2 \leq \max_{i,j} |f[\mu_i, \mu_j]| \frac{\|A^*A - I\|_2}{1 + \sigma_n(A)} + O(\|A - U\|^2).
\]

Any \( A \in \mathbb{C}^{m \times n} \) has a singular value decomposition

\[
A = PSV^*, \quad \Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}, \quad p = \min(m, n),
\]

where \( P \in \mathbb{C}^{m \times m} \) and \( V \in \mathbb{C}^{n \times n} \) are unitary and \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0 \). The parameters \( \sigma_i \) are the singular values of \( A \). We can substitute the singular value
decomposition in \( \|A^*A - I\|_2 \) to obtain
\[
\|A^*A - I\|_2 = \|V \Sigma^* P \Sigma V^* - I\|_2 \\
= \|V (\Sigma^* \Sigma - I) V^*\|_2 \\
= \|\Sigma^*\Sigma - I\|_2 \\
= |\sigma_1^2(A) - 1|.
\]
So we get the bound for the 2-norm
\[
\|f(A) - f(U)\|_2 \leq \tau \frac{|\sigma_1^2(A) - 1|}{\sigma_n(A) + 1} + O(\|A - U\|_2^2),
\]
where \( \tau = \max_{i,j} |f[\mu_i, \mu_j]|. \) Now we try to bound \( \tau \) by taking \( f(x) = \log x \) in equation (3.7). Let \( \Lambda(U) \) be the set of eigenvalues of \( U \). Then
\[
\max_{i,j} |f[\mu_i, \mu_j]| = \max \left( \max_{\mu \in \Lambda(U)} \frac{1}{|\mu|}, \max_{\mu_i, \mu_j \in \Lambda(U), \mu_i \neq \mu_j} \frac{|\log \mu_i - \log \mu_j|}{|\mu_i - \mu_j|} \right).
\]
We know that the eigenvalues of the unitary matrices are written in the following form.
\[
\mu_i = e^{i\theta_i}, \quad |\mu_i| = 1 \quad (3.10)
\]
It is trivial that
\[
\max_{\mu \in \Lambda(U)} \frac{1}{|\mu|} = 1.
\]
To bound
\[
\max_{\mu_i, \mu_j \in \Lambda(U), \mu_i \neq \mu_j} \frac{|\log \mu_i - \log \mu_j|}{|\mu_i - \mu_j|},
\]
we take \( \mu_i = e^{i\theta_i} \) and \( \mu_j = e^{i\theta_j} \). Then by assuming \( -\pi < \theta_i \leq \pi \) we get
\[
\frac{|\log e^{i\theta_i} - \log e^{i\theta_j}|}{|e^{i\theta_i} - e^{i\theta_j}|} = \frac{|\theta_i - \theta_j|}{|e^{i\theta_i} - e^{i\theta_j}|}.
\]
We can write \( \theta_i = \log e^{i\theta_i} \) for \( \theta_i \in (-\pi, \pi] \) by using the corollary in [4, Corol. 2.2]. Since the principal matrix logarithm is discontinuous on \( \mathbb{R}^- \) it is stated that all the imaginary parts of the eigenvalues of \( \log A \) lie in the interval \( (-\pi, \pi) \). For the principal matrix logarithm the extension is given for the interval \( (-\pi, \pi] \) in [4] that \( \log A \) is well defined. That is, the first derivative of the logarithm is defined by using the one sided limit.
\[
\log'(x) = \lim_{\substack{h \to 0 \\ \text{lm} h \geq 0}} \frac{\log(x + h) - \log x}{h}
\]
To bound the value $|e^{i\theta_i} - e^{i\theta_j}|$ we use Euler’s formula.

$$e^{i\theta_i} - e^{i\theta_j} = \cos \theta_i - \cos \theta_j + i(\sin \theta_i - \sin \theta_j).$$  \hfill (3.12)

By using the trigonometric formulas given as

$$\cos \theta_i - \cos \theta_j = -2 \sin \left(\frac{\theta_i + \theta_j}{2}\right) \sin \left(\frac{\theta_i - \theta_j}{2}\right),$$
$$\sin \theta_i - \sin \theta_j = 2 \sin \left(\frac{\theta_i - \theta_j}{2}\right) \cos \left(\frac{\theta_i + \theta_j}{2}\right),$$

we rewrite equation (3.12) as

$$e^{i\theta_i} - e^{i\theta_j} = -2 \sin a \sin b + i2 \sin b \cos a$$
$$= 2i \sin b(\cos a + i \sin a)$$
$$= 2i \sin be^{ia},$$

where $a = \frac{\theta_i + \theta_j}{2}$ and $b = \frac{\theta_i - \theta_j}{2}$. Then we obtain

$$|e^{i\theta_i} - e^{i\theta_j}| = |2i \sin be^{ia}| = 2|\sin b|. \quad (3.13)$$

At the end we rewrite equation (3.11) as

$$\frac{|\theta_i - \theta_j|}{|e^{i\theta_i} - e^{i\theta_j}|} = \frac{|2b|}{2|\sin b|} = \frac{|b|}{|\sin b|}.$$ 

When $b$ approaches zero this term goes to 1 which bounds the error nicely.

$$\lim_{b \to 0} \frac{b}{\sin b} = 1.$$ 

However if $b$ approaches $\mp \pi$ then the error bound becomes arbitrarily large.

$$\lim_{b \to \mp \pi} \frac{b}{\sin b} = \infty.$$ 

**Lemma 3.5.** Let $A \in \mathbb{C}^{n \times n}$ ($n > 1$) have no eigenvalues on $\mathbb{R}^-$ and let $U$ be the nearest unitary matrix to $A$, obtained from a polar decomposition. Let the eigenvalues of the unitary matrix $U$ be given as $e^{i\theta_i}$. For the logarithm of $A$

$$\|\log A - \log U\|_2 \leq \tau \frac{\sigma_1^2(A) - 1}{\sigma_n(A) + 1} + O(\|A - U\|_2^2),$$  \hfill (3.14)

where $\sigma_1(A)$ and $\sigma_n(A)$ are the largest and the smallest singular values of $A$, respectively and

$$\tau = \max \left(1, \frac{b}{\sin b}\right),$$

where $b = (\theta_i - \theta_j)/2$. 

3.1. NEARLY UNITARY MATRICES

The error bound \( \| \log A - \log U \|_2 \) for the closest unitary matrix \( U \) depends on the singular values of \( A \) and the arguments of the eigenvalues of \( U \). If the difference of any two arguments of the eigenvalues of \( U \) gets close to \( 2\pi \) then the error bound approaches infinity.

The square root also occurs in many matrix applications. For \( A \in \mathbb{C}^{n \times n} \) with no eigenvalues on \( \mathbb{R}^- \), \( A^{1/2} \) is the unique square root of \( A \) called the principal square root whose spectrum lies in the open right half plane as stated in Theorem 2.8. We are only interested in the principal matrix square root. When we take \( f(x) = x^{1/2} \) in equation (3.7), we obtain

\[
\max_{i,j} |f[\mu_i, \mu_j]| = \max \left( \max_{\mu \in \Lambda(U)} \frac{1}{2|\mu|}, \max_{\mu_i, \mu_j \in \Lambda(U)} \frac{|\mu_i^{1/2} - \mu_j^{1/2}|}{|\mu_i - \mu_j|} \right). \tag{3.15}
\]

By taking \( \mu = e^{i\theta} \) we get

\[
\max_{\mu \in \Lambda(U)} \frac{1}{2|\mu|} = \max \frac{1}{2|e^{i\theta/2}|} = \frac{1}{2}. \tag{3.16}
\]

For \( \mu_i = e^{i\theta_i} \) and \( \mu_j = e^{i\theta_j} \), the second term in \( \tau \) can be simplified as

\[
\frac{|\sqrt{\mu_i} - \sqrt{\mu_j}|}{|\mu_i - \mu_j|} = \frac{1}{|\sqrt{\mu_i} + \sqrt{\mu_j}|} = \frac{1}{|e^{i\theta_i/2} + e^{i\theta_j/2}|}. \tag{3.17}
\]

When we rewrite it by using Euler’s formula we get

\[
\max_{\theta_i \neq \theta_j} \frac{1}{|e^{i\theta_i/2} + e^{i\theta_j/2}|} = \frac{1}{((\cos \theta_i/2 + \cos \theta_j/2)^2 + (\sin \theta_i/2 + \sin \theta_j/2)^2)^{1/2}}
\]

\[
= \frac{1}{((2 + 2 \cos \theta_i/2 \cos \theta_j/2 + 2 \sin \theta_i/2 \sin \theta_j/2)^{1/2}}
\]

\[
= \frac{1}{(2 + 2 \cos(\theta_i - \theta_j)/2)^{1/2}}.
\]

Lemma 3.6. Let \( A \) have no eigenvalues on \( \mathbb{R}^- \) and let \( U \) be the nearest unitary matrix to \( A \), obtained from a polar decomposition. Let the eigenvalues of the unitary matrix \( U \) be given as \( e^{i\theta_i} \). Then

\[
\| A^{1/2} - U^{1/2} \|_2 \leq \tau \frac{|\sigma_1^2(A) - 1|}{\sigma_n(A) + 1} + O(\| A - U \|_2^2) \tag{3.18}
\]

where \( \sigma_1(A) \) and \( \sigma_n(A) \) are the largest and the smallest singular values of \( A \), respectively and

\[
\tau = \max \left( \frac{1}{2}, \max_{b \neq 0} \frac{1}{(2 + 2 \cos(b))^{1/2}} \right). \tag{3.19}
\]

Here \( b = (\theta_i - \theta_j)/2 \).
The error bound of \( \| A^{1/2} - U^{1/2} \|_2 \) depends not only the singular values of \( A \) but also the arguments of the eigenvalues of the closest unitary matrix \( U \). A similar situation occurs here as well: If \( |\theta_i - \theta_j| \approx 2\pi \) the second term in \( \tau \) goes to infinity and the error bound is arbitrarily large.

### 3.2 Nearly Hermitian matrices

Let \( A_H = (A + A^*)/2 \) and \( A_K = (A - A^*)/2 \) be the Hermitian and the skew-Hermitian parts of \( A \), respectively. Let \( S \) be the class of Hermitian matrices and let \( \min\{\|A - X\| : X \in S\} \) be small. The next theorem states that the nearest Hermitian matrix to \( A \) is \( A_H \).

**Theorem 3.7.** \([22]\) For \( A \in \mathbb{C}^{n \times n} \) and any unitarily invariant norm,

\[
\|A - A_H\| = \min\{\|A - X\| : X = X^*\}.
\]

**Proof.** For any Hermitian \( Y \)

\[
\|A - A_H\| = \frac{1}{2}\|(A - Y) + (Y^* - A^*)\| \\
\leq \frac{1}{2} (\|A - Y\| + \|Y^* - A^*\|) \\
\leq \|A - Y\|.
\]

Using the fact that \( \|A - Y\| = \|(A - Y)^*\| \) for any unitarily invariant norm. \(\square\)

We know that for the Hermitian matrix \( A_H \), \( \log A_H \) is also Hermitian.

\[
(\log A_H)^* = \log A_H^* = \log A_H.
\]

To bound \( \|f(A) - f(A_H)\| \) for the matrix logarithm we take \( X = A_H \) in Lemma 3.1 and the problem reduces to bounding \( \|L_f(A_H)\| \) and \( \|A - A_H\| \). Since \( A_H \) is Hermitian we can write

\[
\|L_f(A_H)\|_2 = \max_{i,j} |f[\mu_i, \mu_j]|,
\]

where \( \mu_i \) are the eigenvalues of \( A_H \). To bound \( \|A - A_H\| \) we write

\[
\|A - A_H\| = \left\| A - \frac{A + A^*}{2} \right\| = \frac{\|A - A^*\|}{2}.
\]
3.2. NEARLY HERMITIAN MATRICES

Then we obtain the bound
\[ \| f(A) - f(A_H) \|_2 \leq \max_{i,j} |f[\mu_i, \mu_j]| \frac{\| A - A^* \|_2}{2} + O(\| A - A_H \|_2^2). \] (3.23)

To bound the value \( \max_{i,j} |f[\mu_i, \mu_j]| \) for the matrix logarithm we take \( f(x) = \log x \) in equation (3.7),
\[ \max_{i,j} |f[\mu_i, \mu_j]| = \left( \frac{\max_{\mu \in \Lambda(A_H)} 1}{|\mu|}, \frac{\max_{\mu_i, \mu_j \in \Lambda(A_H)} |\log \mu_i - \log \mu_j|}{|\mu_i - \mu_j|} \right). \] (3.24)

We assume that the spectrum of \( A_H \) lies in the open right half plane. \( A_H \) has no eigenvalues on \( \mathbb{R}^- \) so that the principal logarithm is defined. The eigenvalues of the Hermitian matrix \( A_H \) are real and given as
\[ \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n > 0. \] (3.25)

Since \( \mu_n \) is the smallest eigenvalue of \( A_H \),
\[ \max_{\mu \in \Lambda(A_H)} 1 = \frac{1}{\mu_n}. \] (3.26)

To bound the second term we apply the mean value theorem [74, Chap. 3.8].
\[ \frac{|\log \mu_i - \log \mu_j|}{|\mu_i - \mu_j|} = \frac{1}{|\xi|} \leq \frac{1}{\mu_n}, \quad \xi \in [\mu_i, \mu_j]. \] (3.27)

Then we get \( \max_{\xi \in [\mu_i, \mu_j]} \frac{1}{|\xi|} = \frac{1}{\mu_n}. \)

**Lemma 3.8.** Let \( A_H \) be the nearest Hermitian matrix to \( A \). Assume that the eigenvalues of \( A_H \) are given as
\[ \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n > 0. \]

For the logarithm of \( A \) we get
\[ \| \log A - \log A_H \|_2 \leq \frac{1}{\mu_n} \| A - A^* \|_2 + O(\| A - A_H \|_2^2). \] (3.28)

Lemma 3.8 states that the error bound of \( \| \log A - \log A_H \|_2 \) depends on not only the distance between \( A \) and \( A^* \) but also the minimum eigenvalue of \( A_H \). If the minimum eigenvalue of \( A_H \) is close to zero then the error bound is large.
3. FUNCTIONS OF NEARLY STRUCTURED MATRICES

3.3 Nearly Hermitian positive semidefinite matrices

A matrix $A \in \mathbb{C}^{n \times n}$ is called positive semidefinite if $x^* A x \geq 0$ for all $x \in \mathbb{C}^n$. Let $A_P$ be the nearest Hermitian positive semidefinite matrix to $A$. Our aim is to investigate the error bound of $\|f(A) - f(A_P)\|$ when we compute $f(A_P)$ instead of $f(A)$. We take $X = A_P$ in Lemma 3.1 so that the problem reduces to bounding $\|L_f(A_P)\|$ and $\|A - A_P\|$. To bound $\|A - A_P\|$ we first give the theorem stating the nearest Hermitian positive semidefinite matrix to $A$ in the Frobenius norm.

Theorem 3.9. [32, Thm. 2.1] Let $A \in \mathbb{C}^{n \times n}$, and let $A_H$ and $A_K$ be the Hermitian and the skew-Hermitian parts of $A$, respectively. Let $A_H = UH$ be a polar decomposition ($U^* U = I$, $H = H^* \geq 0$). Then, $A_P = (A_H + H)/2$ is the unique Hermitian positive semidefinite approximation of $A$ in the Frobenius norm, and

$$\|A - A_P\|_F^2 = \sum_{\lambda_i(A_H) < 0} \lambda_i(A_H)^2 + \|A_K\|_F^2. \tag{3.29}$$

Since $A_P$ is a normal matrix we can write

$$\|L_f(A_P)\|_F = \max_{i,j} |f[\mu_i, \mu_j]|, \tag{3.30}$$

where $\mu_i$ and $\mu_j$ are the eigenvalues of $A_P$ given as

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0. \tag{3.31}$$

We can bound $\|A - A_P\|_F$ as follows.

$$\|A - A_P\|_F \leq \left( \sum_{\lambda_i(A_H) < 0} \lambda_i(A_H)^2 + \|A_K\|_F \right)^{1/2}. \tag{3.32}$$

For the exponential function we take $f(x) = e^x$ in equation (3.6) and get

$$\max_{i,j} |f[\mu_i, \mu_j]| = \left( \max_{\mu \in \lambda(A_P)} |e^\mu|, \max_{\mu_i, \mu_j \in \lambda(A_P), \mu_i \neq \mu_j} \left| e^{\mu_i} - e^{\mu_j} \right| \right). \tag{3.33}$$

Since $e^x$ is an increasing function the first term takes its maximum value at $\mu_1$. Therefore,

$$\max_{\mu \in \lambda(A_P)} |e^\mu| = e^{\mu_1}. \tag{3.34}$$
3.3. NEARLY HERMITIAN POSITIVE SEMIDEFINITE MATRICES

To bound the second term we use the mean value theorem and write

\[ \frac{|e^{\mu_i} - e^{\mu_j}|}{|\mu_i - \mu_j|} = |e^\xi| \leq e^{\mu_1}, \quad \xi \in [\mu_i, \mu_j]. \quad (3.34) \]

We obtain max_{\xi \in [\mu_i, \mu_j]} |e^\xi| = e^{\mu_1}.

**Lemma 3.10.** Let \( A_P \) be the nearest Hermitian positive semidefinite matrix to \( A \), given by \( A_P = (A_H + H)/2 \) and let \( \|A_K\|_F = \alpha \). Assume that the eigenvalues of \( A_P \) are given as

\[ \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0. \quad (3.35) \]

For the exponential of \( A \)

\[ \|e^A - e^{A_P}\|_F \leq e^{\mu_1} \left( \sum_{\lambda_i(A_H) < 0} \lambda_i(A_H)^2 + \alpha \right)^{1/2} + O(\|A - A_P\|_F^2). \quad (3.36) \]

If \( A \) is Hermitian then \( \alpha = 0 \) and the upper bound becomes

\[ \|e^A - e^{A_P}\|_F \leq e^{\mu_1} \left( \sum_{\lambda_i(A_H) < 0} \lambda_i(A_H)^2 \right)^{1/2} + O(\|A - A_P\|_F^2). \quad (3.37) \]

For the exponential matrix the error bound for \( \|e^A - e^{A_P}\|_F \) depends on negative eigenvalues of the Hermitian part and the norm of the skew-Hermitian part of \( A \). The error bound can be arbitrarily large depending on the largest eigenvalue of \( A_P \).

The matrix sine and cosine functions also play an important role in matrix functions since they occur in the solutions of second order differential equations. For a given \( A \in \mathbb{C}^{n \times n} \) the sine and the cosine functions are defined by their Maclaurin series.

\[
\sin A = A - \frac{A^3}{3!} + \frac{A^5}{5!} - \frac{A^7}{7!} + \cdots, \\
\cos A = I - \frac{A^2}{2!} + \frac{A^4}{4!} - \frac{A^6}{6!} + \cdots.
\]

We will bound \( \|f(A) - f(A_P)\|_F \) by taking \( f(x) = \sin x \) and \( f(x) = \cos x \) in equation (3.7), respectively. So we can examine what happens if we compute the sine and the cosine functions of the closest Hermitian positive semidefinite matrix instead of the nearly structured one. When we take \( f(x) = \sin x \) in equation (3.7) we get

\[
\max_{i,j} |f[\mu_i, \mu_j]| = \left( \max_{\mu \in \Lambda(A_P)} |\cos \mu|, \max_{\mu_i, \mu_j \in \Lambda(A_P), \mu_i \neq \mu_j} \frac{|\sin \mu_i - \sin \mu_j|}{|\mu_i - \mu_j|} \right).
\]
It is obvious that \( \max_{\mu \in \Lambda(A_P)} |\cos \mu| = 1 \). To bound the second part we apply the mean value theorem and get
\[
\frac{|\sin \mu_i - \sin \mu_j|}{|\mu_i - \mu_j|} = |\cos \xi| \leq 1, \quad \xi \in [\mu_i, \mu_j].
\]
The similar analysis is done for the cosine function. By taking \( f(x) = \cos x \) in equation (3.7) we get
\[
\max_{i,j} |f[\mu_i, \mu_j]| = \left( \max_{\mu \in \Lambda(A_P)} |\sin \mu|, \max_{\mu_i, \mu_j \in \Lambda(A_P), \mu_i \neq \mu_j} \frac{|\cos \mu_i - \cos \mu_j|}{|\mu_i - \mu_j|} \right).
\]
Here, \( \max_{\mu \in \Lambda(A_P)} |\sin \mu| = 1 \). To bound the second part we again apply the mean value theorem.
\[
\frac{|\cos \mu_i - \cos \mu_j|}{|\mu_i - \mu_j|} = |\sin \xi| \leq 1, \quad \xi \in [\mu_i, \mu_j].
\]
In conclusion, for the sine and the cosine functions \( \max_{i,j} |f[\mu_i, \mu_j]| = 1 \), where \( \mu_i \) are the eigenvalues of \( A_P \).

**Lemma 3.11.** Let \( A_P \) be the nearest Hermitian positive semidefinite matrix to \( A \), given by \( A_P = (A_H + H)/2 \) and let \( \|A_K\|_F = \alpha \). For the sine and the cosine functions of the nearly Hermitian positive semidefinite matrix \( A_P \) the bound becomes
\[
\|f(A) - f(A_P)\|_F \leq \left( \sum_{\lambda_i(A_H) < 0} \lambda_i(A_H)^2 + \alpha \right)^{1/2} + O(\|A - A_P\|_F^2). \tag{3.38}
\]
If \( A \) is Hermitian then \( \alpha = 0 \) and the upper bound becomes
\[
\|f(A) - f(A_P)\|_F \leq \left( \sum_{\lambda_i(A_H) < 0} \lambda_i(A_H)^2 \right)^{1/2} + O(\|A - A_P\|_F^2). \tag{3.39}
\]

Lemma 3.11 indicates that there is no coefficient or function magnifying the error bound of \( \|f(A) - f(A_P)\| \) for the sine and the cosine functions. It depends on the eigenvalues of the Hermitian part and the norm of the skew-Hermitian part of \( A \).
4

Structured Condition Number for Matrix Functions

4.1 Introduction

A theory for the condition number of matrix functions was developed by Kenney and Laub [45], in which the structure was not taken into account. The theory was extended to structured matrices in Lie and Jordan algebras of a scalar product in [17]. When the matrix $X$ is structured, the perturbed matrix $X + \Delta X$ is required to have the same structure. Since Lie and Jordan algebras are linear vector subspaces it is guaranteed that $\Delta X$ belongs to the same space as $X$. However for nonlinear spaces we cannot preserve the structure after $X$ is perturbed. In this work the theory is extended to nonlinear vector spaces by defining the differential of a map between smooth square matrix manifolds.

The chapter is organized as follows. In Section 4.2 after defining the structured condition number we show that the differential defined between the tangent spaces to smooth square matrix manifolds plays the same role as the Fréchet derivative, which is used in the computation of the condition number. We analyse the structured condition number for three different classes of smooth square matrix manifolds. We derive an algorithm for computing the structured condition number of a map between smooth
square matrix manifolds and introduce an algorithm to estimate the structured condition number by using the power method. We also give a lower and upper bounds on the structured condition number, which are cheaper to compute than the structured condition number. In Section 4.3 we specialize the theory to automorphism groups, Lie and Jordan algebras associated with nondegenerate bilinear or sesquilinear forms on \( \mathbb{R}^n \) or \( \mathbb{C}^n \) and characterize their tangent spaces. We bound the structured condition number for each matrix group. A case for doubly structured matrices (symplectic-orthogonal) is also analysed. In Section 4.4 we illustrate two examples for the structured condition number of the matrix logarithm and the square root. In Section 4.5 numerical tests are presented and it is shown that the unstructured and structured condition numbers can differ by several orders of magnitude, which motivates the development of structure preserving algorithms for matrix functions. We also compare the structured and unstructured relative error bounds for the matrix logarithm and square root.

### 4.2 Structured condition number

Suppose that \( f : \mathcal{M} \to \mathcal{N} \) is a smooth matrix function, where \( \mathcal{M}, \mathcal{N} \) are smooth square matrix manifolds. Let \( X \in \mathcal{M} \) be such that \( f(X) \) is defined. We start with recalling the unstructured condition number of \( f(X) \)

\[
\text{cond}(f, X) = \lim_{\epsilon \to 0} \sup_{\|E\| \leq \epsilon} \frac{\|f(X + E) - f(X)\|}{\epsilon}.
\]

When the Fréchet derivative of \( f \) exists we can write the unstructured condition number in terms of Fréchet derivative, which is given in (1.2)

\[
\text{cond}(f, X) = \max_{E \neq 0} \frac{\|L_f(X, E)\|}{\|E\|} = \|L_f(X)\|.
\]

As shown in equation (1.4), by using the fact that the Fréchet derivative is linear in \( E \) the problem of computing the unstructured condition number reduces to finding the 2-norm of \( K_f(X) \).

\[
\text{cond}(f, X) = \max_{E \neq 0} \frac{\|\text{vec}(L_f(X, E))\|_2}{\|\text{vec}(E)\|_2} = \max_{E \neq 0} \frac{\|K_f(X)\text{vec}(E)\|_2}{\|\text{vec}(E)\|_2} = \|K_f(X)\|_2.
\]  

(4.1)

In the definition of the unstructured condition number we assume that \( f \) is complex Fréchet differentiable. There are two issues here: (1) Complex differentiability is an
unaturally strong condition, (2) Some matrix maps of practical interest are only real 
Fréchet differentiable such as the unitary factor of the polar decomposition or other 
generalized matrix functions [66, Sec. 3].

The difference between the real Fréchet differentiability and the complex Fréchet 
differentiability is the linearity. If the complex Fréchet derivative (resp. real Fréchet 
derivative) exists then it is complex linear (resp. real linear). Assume that the complex 
valued function \( f : \mathbb{C}^{n \times n} \to \mathbb{C}^{n \times n} \) is only real Fréchet differentiable then the Fréchet 
derivative of \( f \) is not complex linear in \( E \) and equation (1.3) does not hold so we need 
to define \( \rho \) function. Let

\[
\rho : \mathbb{C}^{n \times n} \to \mathbb{R}^{n \times 2n} : A + iC \mapsto \begin{bmatrix} A & C \end{bmatrix}, \quad A, C \in \mathbb{R}^{n \times n}.
\]

(4.2)

Note that for any \( X \in \mathbb{C}^{n \times n} \) it holds

\[
\vec(\rho(X)) = \begin{bmatrix} \vec(\text{Re}(X)) \\ \vec(\text{Im}(X)) \end{bmatrix},
\]

and that \( \rho^{-1} \) corresponds to right multiplication times \( \begin{bmatrix} I_n \\ iI_n \end{bmatrix} \). With this notation, the 
Fréchet derivative of a real differentiable map \( f : \mathbb{C}^{n \times n} \to \mathbb{C}^{n \times n} \) can be represented as 
a matrix \( K_f^{(R)} \in \mathbb{R}^{2n^2 \times 2n^2} \) such that

\[
\vec(\rho(L_f(X, E))) = K_f^{(R)} \vec(\rho(E)).
\]

(4.3)

Again, \( K_f^{(R)} \) can be explicitly formed computing its \( 2n^2 \) columns as \( \vec(\rho(L_f(X, e_ie_j^T))) \) 
and \( \vec(\rho(L_f(X, ie_i e_j^T))) \), and

\[
\text{cond}(f, X) = \| K_f^{(R)} \|_2.
\]

(4.4)

If \( f \) is real differentiable, but not complex Fréchet differentiable, then the only Fréchet 
derivative that exists and can be computed is \( K_f^{(R)} \). We give an example for the case 
where \( f(x) \) is real Fréchet differentiable, but not complex Fréchet differentiable.

**Example 4.1.** Let \( A \in \mathbb{C}^{n \times n} \) be nonsingular and let \( f : A \mapsto U \) be the map that 
associates \( A \) with the unitary factor \( U \) of its polar decomposition. Then \( f \) is real 
differentiable but not complex differentiable [66]. For example, for \( n = 1 \), then \( f(z) = \)

for \( z \neq 0 \). Writing \( z = x + iy \) with \( x, y \in \mathbb{R} \) and a complex perturbation \( e = e_x + ie_y \) with \( e_x, e_y \in \mathbb{R} \), we have that

\[
f(x, y) = (f_1(x, y), f_2(x, y)) = \left( \frac{x}{(x^2 + y^2)^{1/2}}, \frac{y}{(x^2 + y^2)^{1/2}} \right).
\]

Then the Kronecker form of the real Fréchet derivative can be written as

\[
K_f^{(R)} = \begin{bmatrix}
\frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\
\frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y}
\end{bmatrix} = |z|^{-3} \begin{bmatrix}
y^2 & -xy \\
-xy & x^2
\end{bmatrix},
\]

which implies \([\text{Re}(L_f(z, e)) \quad \text{Im}(L_f(z, e))]^T = K_f^{(R)} [e_x \quad e_y]^T\), and hence

\[
L_f(z, e) = \frac{(y - ix)(ye_x - xe_y)}{|z|^3}.
\]

As we see from the Kronecker form of the Fréchet derivative, the Cauchy-Riemann equations are not satisfied. Therefore, there is no complex Fréchet derivative.

If \( f \) is also complex differentiable then \( K_f^{(R)} \) will have the form

\[
K_f^{(R)} = \begin{bmatrix}
R_f & -P_f \\
P_f & R_f
\end{bmatrix}, \quad R_f, P_f \in \mathbb{R}^{n \times n},
\]

(4.5)

implying that, after applying \( \text{vec} \circ \rho^{-1} \circ \text{unvec} \) to (4.3),

\[
\text{vec}(L_f(X, E)) = (R_f + iP_f)\text{vec}(E) =: K_f^{(C)}.
\]

Matrices of the form (4.5) have a particular singular value decomposition.

**Lemma 4.2.** Let \( A, B \in \mathbb{R}^{m \times n} \). Then the singular values of \( A+iB \) are \( \sigma_1, \sigma_2, \ldots, \sigma_{\min(m,n)} \) if and only if the singular values of

\[
\begin{bmatrix}
A & B \\
-B & A
\end{bmatrix}
\]

are \( \sigma_1, \sigma_1, \sigma_2, \sigma_2, \ldots, \sigma_{\min(m,n)}, \sigma_{\min(m,n)} \).

In particular, it holds

\[
\text{rank}_\mathbb{R} \begin{bmatrix}
A & B \\
-B & A
\end{bmatrix} = 2\text{rank}_\mathbb{C}(A + iB).
\]

**Proof.** Observe that

\[
\begin{bmatrix}
I_m/\sqrt{2} & -iI_m/\sqrt{2} \\
-iI_m/\sqrt{2} & I_m/\sqrt{2}
\end{bmatrix} \begin{bmatrix}
A & B \\
-B & A
\end{bmatrix} \begin{bmatrix}
I_n/\sqrt{2} & iI_n/\sqrt{2} \\
iI_n/\sqrt{2} & I_n/\sqrt{2}
\end{bmatrix} = \begin{bmatrix}
A + iB & 0 \\
0 & A - iB
\end{bmatrix},
\]

which implies the statement as the singular values of \((A - iB)\) and \((A + iB)\) are the same. \( \square \)
4.2. STRUCTURED CONDITION NUMBER

As a corollary, when \( f \) is complex Fréchet differentiable one has \( \|K_f^{(R)}\|_2 = \|K_f^{(C)}\|_2 \). This shows that the computation of the unstructured condition number is independent on whether the real or complex Kronecker forms of the Fréchet derivatives are considered.

Using basic notions of differential geometry we show that the differential \( df \) plays the same role as the Fréchet derivative for the case of unstructured condition number. To define the differential of \( f \) we need to define the tangent spaces to the related smooth manifolds as stated in Definition 1.16, where \( \gamma : \mathbb{F} \to \mathcal{M} \) is given such that \( \gamma(0) = X, \gamma'(0) = E \). The set

\[
T_X \mathcal{M} := \{ \gamma'(0) = E \mid \gamma : \mathbb{F} \to \mathcal{M} \text{ is smooth}, \gamma(0) = X \}
\]

(4.6)
of tangent vectors of \( \mathcal{M} \) at \( X \) is called the tangent space of \( \mathcal{M} \) at \( X \). It is worth pointing out that \( T_X \mathcal{M} \) is a linear subspace of \( \mathbb{F} \). We distinguish between the field in which the entries of the matrices are allowed to lie, denoted by \( \mathbb{K} \) as in the previous chapters, and the field that the tangent space is built on, that we instead denote by \( \mathbb{F} \). The choice of the domain space \( \mathbb{F} \) of \( \gamma(t) \) depends on the choice of the scalar product defining \( \mathcal{M} \). If the scalar product is bilinear then we choose \( \mathbb{K} = \mathbb{F} \) but if it is sesquilinear then we choose \( \mathbb{F} = \mathbb{R} \neq \mathbb{K} \) since for sequilinear forms, \( \mathbb{K} = \mathbb{C} \).

**Definition 4.3** (Differential of smooth maps). Let \( \mathcal{M}, \mathcal{N} \in \mathbb{K}^{n \times n} \) be smooth square matrix manifolds and suppose there is a differentiable map \( f : U \to V \), where \( U \subset \mathcal{M} \) and \( V \subset \mathcal{N} \) are open sets. For a given \( X \in U \), consider a smooth curve \( \gamma(t) : \mathbb{F} \to U \) such that \( \gamma(0) = X \). The differential of the map \( f \) at \( X \) is the map

\[
df_X : T_X \mathcal{M} \to T_{f(X)} \mathcal{N}, \quad df_X(\gamma'(0)) = (f \circ \gamma)'(0).
\]

(4.7)

**Proposition 4.4.** The differential map \( df_X : T_X \mathcal{M} \to T_{f(X)} \mathcal{N} \) is linear.

**Proof.** Let \( \gamma_i, i = 0, 1, 2 \) be smooth curves such that \( \gamma_0(t) = X + t(\alpha E + F) + \mathcal{O}(t^2) \),
\( \gamma_1(t) = X + tE + \mathcal{O}(t^2), \gamma_2(t) = X + tF + \mathcal{O}(t^2) \). Then
\[
\alpha df_X(E) + df_X(F) = \alpha \frac{d}{dt} f(\gamma_1(t)) \bigg|_{t=0} + \frac{d}{dt} f(\gamma_2(t)) \bigg|_{t=0}
= \alpha f'(X)E + f'(X)F
= f'(X)(\alpha E + F)
= \frac{d}{dt} f(\gamma_0(t)) \bigg|_{t=0} = df_X(\alpha E + F).
\]

The differential map \( df_X \) can be seen as a generalization of the Jacobian of a map between Euclidean spaces, \([60]\) i.e., \( df_X \) is the best local (at \( X \)) linear approximation to the map. The linear operator \( df_X \) comes equipped with a naturally induced norm
\[
\|df_X\| := \max_{E \in T_XM, E \neq 0} \frac{\|df_X(E)\|}{\|E\|}.
\]
Here, \( \|df_X(E)\| \) and \( \|E\| \) are any suitable norms on, respectively, \( T_{f(X)}N \) and \( T_XM \). When \( M \) and \( N \) are smooth square matrix manifolds, \( T_XM \) and \( T_{f(X)}N \) are both \( \mathbb{F} \)-linear subspaces, so we can embed them in the larger space and take any usual matrix norm. Before giving the relation between the differential of a map \( f(x) \) and the structured condition number of \( f(x) \) we define an absolute structured condition number.

**Definition 4.5** (Structured condition number of a matrix function between smooth square matrix manifolds). Let \( M, N \) be smooth square matrix manifolds and let \( X \in M \) be such that \( f(X) \in N \) is defined. Then the absolute structured condition number of \( f(X) \) is
\[
\text{cond}_{\text{struc}}(f, X) = \lim_{\epsilon \to 0} \sup_{Y \in M, \|Y - X\| \leq \epsilon} \frac{\|f(Y) - f(X)\|}{\epsilon}.
\]

As Figure 4.1 shows, the extra condition \( Y \in M \) restricts the choice of the perturbation \( Y - X \) to a smaller set, and hence, by the definition of supremum, we have the obvious fact
\[
\text{cond}_{\text{struc}}(f, X) \leq \text{cond}(f, X),
\]
where \( \text{cond}(f, X) \) is the unstructured condition number defined in equation (1.1). Whether equality holds is unclear a priori and might depend on both \( f \) and \( X \). If it does
not hold, and particularly when the ratio between the structured and the unstructured condition numbers is much smaller than 1, we get a clear indication that using a structure preserving numerical method to compute \( f(X) \) would be advantageous.

We assume that the matrix function \( f \) is a smooth map between manifolds and, using basic notions of differential geometry, show that, the differential \( df \) plays the same role as the Fréchet derivative for the case of unstructured condition number, which is also used in [18]. The Fréchet derivative is defined between smooth matrix manifolds so using the differential \( df \) restricts the domain set to the tangent space to the smooth manifolds.

**Theorem 4.6.** Let \( f : U \to V \) be a smooth map between two open subsets \( U \subset \mathcal{M}, V \subset \mathcal{N} \), where \( \mathcal{M}, \mathcal{N} \) are smooth square matrix manifolds. Then for the structured condition number of \( f \) at \( X \in U \) it holds that

\[
\text{cond}_{\text{struc}}(f, X) = \|df_X\|.
\]

**Proof.** Suppose \( X, Y \in U \) and let \( \gamma : \mathbb{F} \to U \) be any curve such that \( \gamma(0) = X, \gamma(\epsilon) = Y \). Letting \( E = \gamma'(0) \), we get \( \|Y - X - \epsilon E\| = O(\epsilon^2) \) by using the definition of derivative along a curve and by using the definition of differential we get \( \|f(Y) - f(X) - \epsilon df_X(E)\| = O(\epsilon^2) \). The definition of structured condition number in (4.8) is equivalent to

\[
\text{cond}_{\text{struc}}(f, X) = \lim_{\|X - Y\| \to 0 \atop Y \in \mathcal{M}} \sup_{Y \in \mathcal{M}} \frac{\|f(Y) - f(X)\|}{\|Y - X\|},
\]

and hence, we obtain

\[
\text{cond}_{\text{struc}}(f, X) = \lim_{\epsilon \to 0} \sup_{\epsilon \|E\| + O(\epsilon^2)} \frac{\epsilon \|df_X(E)\| + O(\epsilon^2)}{\epsilon \|E\| + O(\epsilon^2)} = \|df_X\|. \quad \square
\]
When the differential and the Fréchet derivative exist they are unique. The uniqueness of the differential and the uniqueness of the Fréchet derivative imply that for any \( E \in T_X\mathcal{M} \),

\[
df_X(E) = L_f(X, E).
\]

(4.9)

Since \( T_X\mathcal{M} \) is a \( \mathbb{F} \)-linear subspace of \( \mathbb{K}^{n \times n} \), every element of the tangent space \( E \in T_X\mathcal{M} \) is then encoded by

\[
\text{vec}(E) = By,
\]

(4.10)

where \( B \in \mathbb{K}^{n^2 \times p} \) is a matrix of full rank giving a basis for \( T_X\mathcal{M} \) and \( y \in \mathbb{F}^p \) with \( p = \dim T_X\mathcal{M} \leq n^2 \) is a vector of parameters. For example, if \( T_X\mathcal{M} \) is the real submanifold of \( \mathbb{C}^{2 \times 2} \) defined as the set of the \( 2 \times 2 \) Hermitian matrices with \( a, b, c, d \in \mathbb{R} \), then

\[
E = \begin{bmatrix} a & b + ic \\ b - ic & d \end{bmatrix}, \quad \text{vec}(E) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -i & 0 \\ 0 & 1 & i & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}.
\]

We consider in particular three classes of smooth square matrix manifolds:

- Real submanifolds of the \( n^2 \)-dimensional real vector space \( \mathbb{R}^{n \times n} \),

- Complex submanifolds of the \( n^2 \)-dimensional complex vector space \( \mathbb{C}^{n \times n} \),

- Real submanifolds of the \( 2n^2 \)-dimensional \textit{real} vector space \( \mathbb{C}^{n \times n} \).

These objects have been thoroughly studied in differential geometry [51, Chap. 1], but they commonly arise in matrix theory and in numerical linear algebra as well. They are often associated with, respectively, real bilinear forms, complex bilinear forms, and sesquilinear forms.

We analyse two cases: (1) The field \( \mathbb{K} \) in which the entries of matrices are allowed to lie and the field \( \mathbb{F} \) that the tangent vector space is built on are the same, that is \( \mathbb{K} = \mathbb{F} = \mathbb{R} \) or \( \mathbb{K} = \mathbb{F} = \mathbb{C} \), (2) The fields are different, \( \mathbb{K} = \mathbb{C} \) and \( \mathbb{F} = \mathbb{R} \).
4.2. STRUCTURED CONDITION NUMBER

The case where $\mathbb{K} = \mathbb{F}$

Suppose that $\mathbb{K} = \mathbb{F}$ and the map $f$ is $\mathbb{F}$-Fréchet differentiable. Then the Fréchet derivative can be represented by an $n^2 \times n^2$ matrix $K_f(X)$. Applying the vec operator to \((4.9)\) and using \((1.3)\) yields

\[
\vec(df_X(E)) = \vec(L_f(X, E)) = K_f(X)\vec(E) = K_f(X)By. \tag{4.11}
\]

From Theorem 4.6 and on using the Frobenius norm, we find that

\[
\text{cond}_{\text{struc}}(f, X) = \max_{E \in T_XM} \frac{\|L_f(X, E)\|_F}{\|E\|_F} = \max_{y \in \mathbb{K}^p} \frac{\|K_f(X)By\|_2}{\|By\|_2}. \tag{4.12}
\]

Observe now that $\|B^+\|^{-1}\|y\|_2 \leq \|By\|_2 \leq \|B\|_2\|y\|_2$, where $B^+$ denotes the Moore-Penrose pseudoinverse of $B$, which is given in Definition 1.22. Hence,

\[
\|K_f(X)B\|_2\|B\|_2^{-1} \leq \text{cond}_{\text{struc}}(f, X) \leq \|K_f(X)B\|_2\|B^+\|_2. \tag{4.12}
\]

Note that if $B$ is chosen to have orthonormal columns, i.e., $B^*B = I_p$, then the lower and upper bounds in \((4.12)\) are equal so that

\[
\text{cond}_{\text{struc}}(f, X) = \|K_f(X)B\|_2. \tag{4.13}
\]

One way to compute $\text{cond}_{\text{struc}}(f, X)$ is to get an explicit characterization of $T_XM$, find its dimension $p = \dim T_XM$, and then construct a matrix $B \in \mathbb{F}^{n^2 \times p}$ with orthonormal columns such that for any $E \in T_XM$, $\vec(E) = By$ for some $y \in \mathbb{K}^p$. Then if we assume that we have a numerical method to compute $f(X)$ and $L_f(X, E)$ for a given $X \in M$ and $E \in T_XM$, then we can compute the columns of $K_f(X)B$ using

\[
K_f(X)Be_i = \vec(L_f(X, \text{unvec}(Be_i))), \quad i = 1: p, \tag{4.14}
\]

where unvec is the inverse of the vec operator which in this case is defined from $\mathbb{K}^{n^2}$ to $\mathbb{K}^{n \times n}$.

The case where $\mathbb{K} \neq \mathbb{F}$

In the second case we consider the perturbations restricted on a real submanifold of the $2n^2$-dimensional real vector space $\mathbb{C}^{n \times n}$. In the computation of structured condition number there are two drawbacks:
The function \( f \) may not be complex differentiable. We can write \( \text{vec}(E) = By \) for some \( B \in \mathbb{C}^{n^2 \times p} \) and \( y \in \mathbb{R}^p \), where \( p = \dim(T_X \mathcal{M}) \). By using the map \( \rho \) defined in (4.2) we have that \( \text{vec}(\rho(E)) = B_R y \) for

\[
B_R = \begin{bmatrix}
\text{Re} B \\
\text{Im} B
\end{bmatrix} = \rho(B)^T,
\]

\( (^T \) denotes block transposition). The analysis (when \( \mathbb{K} = \mathbb{F} \)) still holds provided that we consider the real Fréchet derivative \( K_f^{(R)} \) and replace \( B \) with \( B_R \). Weyl’s inequalities on singular values [80], together with Lemma 4.2, imply that

\[
\|B_R\|_2 \leq \|B\|_2 \leq \sqrt{2}\|B_R\|_2, \quad \|B^+\|_2 \leq \|B^+_R\|_2 \leq \sqrt{2}\|B^+\|_2.
\]

Even if \( f \) is complex differentiable and the complex Fréchet derivative \( K_f^{(C)} \) exists, \( T_X \mathcal{M} \) may not be a complex linear subspace of \( \mathbb{C}^{n \times n} \) so that (4.13) and (4.14) may not hold. Then,

\[
\|K_f^{(R)} B_R\|_2 = \max_{0 \neq y \in \mathbb{R}^p} \frac{\|K_f^{(R)} B_R y\|_2}{\|y\|_2} = \max_{0 \neq y \in \mathbb{R}^p} \frac{\|K_f^{(C)} B y\|_2}{\|y\|_2} \leq \|K_f^{(C)} B\|_2.
\]

There is no guarantee that the dominant singular vector of \( K_f^{(C)} B \) can be taken to be real. Therefore, for real submanifolds the structured condition number is inherently associated with the real Fréchet derivative. In the notation, (4.12) becomes

\[
\|K_f^{(R)}(X) B_R\|_2 \|B_R\|_2^{-1} \leq \text{cond}_{\text{struc}}(f, X) \leq \|K_f^{(R)}(X) B_R\|_2 \|B^+_R\|_2, \quad (4.15)
\]

and if \( B_R \) has orthonormal columns,

\[
\text{cond}_{\text{struc}}(f, X) = \|K_f^{(R)}(X) B_R\|_2. \quad (4.16)
\]

We will show in the next example that even if \( f \) is complex differentiable while the bounds in (4.12) do not hold, the bounds in (4.15) are satisfied.

**Example 4.7.** Let \( \mathcal{M} \subset \mathbb{C}^{2 \times 2} \) be the real submanifold of symplectic matrices, and take \( X = \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \in \mathcal{M} \). A basis for \( T_X \mathcal{M} \) is obtained by using (4.29)
4.2. **STRUCTURED CONDITION NUMBER**

(for sesquilinear form)

\[
B = \begin{bmatrix}
0 & 0 & 4 & 4i \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & -1 & i
\end{bmatrix}.
\]

Take \( f(X) = X^2 \). Since \( f \) is complex differentiable, one can compute

\[
K_f^{(C)}(X) = \begin{bmatrix}
4 & 0 & 0 & 0 \\
0 & 5/2 & 0 & 0 \\
0 & 0 & 5/2 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix} \Rightarrow K_f^{(C)}(X)B = \begin{bmatrix}
0 & 0 & 16 & 16i \\
5/2 & 0 & 0 & 0 \\
0 & 5/2 & 0 & 0 \\
0 & 0 & -1 & i
\end{bmatrix},
\]

where \( K_f^{(R)} = \text{diag}(K_f^{(C)}, K_f^{(C)}) \in \mathbb{R}^{8 \times 8} \). Computing \( B_R = \begin{bmatrix} \text{Re}B \\ \text{Im}B \end{bmatrix} \), and orthonormalizing its columns, we find the basis

\[
Z_R = \begin{bmatrix}
0 & 0 & 4/\sqrt{17} & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & -1/\sqrt{17} & 0 \\
0 & 0 & 0 & 4/\sqrt{17} \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/\sqrt{17}
\end{bmatrix},
\]

and hence, \( \text{cond}_{\text{struct}}(f, X) = \|K_f^{(R)}Z_R\|_2 = \sqrt{\frac{257}{17}} \). However, \( \|K_f^{(C)}B\|_2 = 16\sqrt{2} \) and \( \|B\|_2 = 4\sqrt{2} \), and hence, the lower bound in (4.12) fails since \( 4 > \sqrt{\frac{257}{17}} \), demonstrating that indeed (4.12) only holds if \( \mathbb{F} = \mathbb{K} \). On the contrary, \( \|B_R\|_2 = \sqrt{17} \), and \( \|K_f^{(R)}B_R\|_2 = \sqrt{257} \), so that the lower bound in (4.15) holds, and it is in fact attained for this example.

We give the algorithm computing the structured condition number for the cases where \( \mathbb{K} = \mathbb{F} \) or \( \mathbb{K} \neq \mathbb{F} \).

**Algorithm 4.8.** Given any algorithm to compute the Fréchet derivative of \( f : \mathcal{M} \to \mathcal{N} \), where \( \mathcal{M}, \mathcal{N} \) are smooth square manifolds of \( \mathbb{K}^{n \times n} \), and given \( X \in \mathcal{M} \) and \( B \in \mathbb{K}^{n \times r} \),
such that for any \( E \in T_XM \), \( \text{vec}(E) = B y \) for some \( y \in \mathbb{F}^p \), this algorithm computes \( \kappa = \|K_f(X)B\|_2 \) if \( \mathbb{F} = \mathbb{K} \) or \( \|K_f^{(R)}B_{\mathbb{R}}\|_2 \) otherwise. If \( B^*B = I_p \) (for \( \mathbb{F} = \mathbb{K} \)) or \( B_{\mathbb{R}}^TB_{\mathbb{R}} = I_p \) (otherwise), then \( \kappa = \text{cond}_{\text{struc}}(f, X) \).

1. if \( \mathbb{F} = \mathbb{K} \)
2. \( K = 0 \in \mathbb{K}^{n^2 \times p} \)
3. for \( i = 1:p \)
4. Compute \( F = L_f(X, E) \), where \( \text{vec}(E) = Be_i \)
5. \( Ke_i = \text{vec}(F) \)
6. end
7. else
8. \( K = 0 \in \mathbb{K}^{2n^2 \times p} \)
9. for \( i = 1:p \)
10. Compute \( F = L_f(X, E) \), where \( \text{vec}(\rho(E)) = B_{\mathbb{R}}e_i \)
11. \( Ke_i = \text{vec}(\rho(F)) \)
12. end
13. end
14. \( \kappa = \|K\|_2 \)

Note that the construction of \( K_f(X)B \) (or \( B = B_{\mathbb{R}} \)) in this way can be cheaper than that of \( K_f(X) \in \mathbb{K}^{n^2 \times n^2} \) using (1.5) since \( p \leq n^2 \). When \( X \in M = \mathbb{K}^{n \times n} \), \( T_XM = \mathbb{K}^{n \times n} \) and \( B = I_{n^2} \) in (4.10) so that (4.14) reduces to (1.5).

If \( \mathbb{F} = \mathbb{K} \) then the construction of \( K = K_f(X)B \) in Algorithm 4.8 costs \( O(pn^3) \) operations assuming that \( L_f(X, E) \) can be computed in \( O(n^3) \) operations, and the cost of computing the 2-norm of \( K \) in step 14 is \( O(p^2n^2) \) so Algorithm 4.8 costs \( O(pn^3 + p^2n^2) \) operations, which is high in particular when \( p \approx n^2 \). The situation is similar for \( \mathbb{F} \neq \mathbb{K} \), except that the factor 2 in the sizes of \( K_f^{(R)} \) and \( B_{\mathbb{R}} \) leads to higher constants in front of the asymptotic complexities.

Once \( B \) (or \( B = B_{\mathbb{R}} \)) is known, we can use the power method to obtain a lower bound for \( \|K_f(X)B\|_2 \) (resp. \( \|K_f^{(R)}B_{\mathbb{R}}\|_2 \)), which corresponds to a lower bound for \( \text{cond}_{\text{struc}}(f, X) \) when \( B \) (resp. \( B_{\mathbb{R}} \)) has orthonormal columns.
4.3. Computing \text{cond}_{\text{struc}}(F, X) for a given smooth manifold $S_M$

Algorithm 4.9. Given any algorithm to compute the Fréchet derivative of $f : \mathcal{M} \to \mathcal{N}$, where $\mathcal{M}, \mathcal{N}$ are smooth square manifolds of $\mathbb{K}^{n \times n}$, and given $X \in \mathcal{M}$ and $B \in \mathbb{K}^{n^2 \times p}$ such that for any $E \in T_X \mathcal{M}$, $\text{vec}(E) = By$ for some $y \in \mathbb{F}^p$, this algorithm uses the power method to compute $\gamma$ such that $\gamma \leq \|K_f(X)B\|_2$, for $\mathcal{F} = \mathbb{K}$, or $\gamma \leq \|K_f^{(R)}(B_R)\|_2$ otherwise.

1. Choose a nonzero starting vector $z_0 \in \mathbb{F}^p$.
2. for $k = 0$: $\infty$
3. if $\mathcal{F} = \mathbb{K}$, $\text{vec}(E_k) = Bz_k$, else, $\text{vec}(\rho(E_k)) = B_R z_k$, end
4. $W_{k+1} = L_f(X, E_k)$
5. $Y_{k+1} = L_f^*(X, W_{k+1})$
6. if $\mathcal{F} = \mathbb{K}$, $z_{k+1} = B^*\text{vec}(Y_{k+1})$, else, $z_{k+1} = B^T_R \text{vec}(\rho(Y_{k+1}))$, end
7. $\gamma_{k+1} = \|z_{k+1}\|_2/\|W_{k+1}\|_F$
8. if converged, $\gamma = \gamma_{k+1}$, quit, end
9. end

Unless $B$ (or $B = B_R$) has a special structure that can be exploited in step 3 and step 6, the cost of Algorithm 4.9 is $O(n^2 pk)$ operations, where $k$ is the number of iterations and we assume that $L_f(X, E_k)$ and $L_f^*(X, W_{k+1})$ can be computed in $O(n^3)$ operations. The convergence test for the power method is given by

\[ k > \max_{\text{it}}, \quad \text{or} \quad |\gamma_{k+1} - \gamma_k| \leq \text{tol} \cdot \gamma_{k+1}, \]

where $\max_{\text{it}}$ is a prescribed maximal number of iterations and $\text{tol}$ is a fixed tolerance.

We are only interested in computing the correct order of magnitude, so in practice we can take $\text{tol} = 10^{-1}$ or $10^{-2}$.

4.3 Computing \text{cond}_{\text{struc}}(f, X) for a given smooth manifold $S_M$

As a smooth matrix manifold we take $\mathcal{M} = S_M$. To compute or approximate $\text{cond}_{\text{struc}}(f, X)$ for $X \in S_M \in \{J_M, L_M, G_M\}$, we need to
1. characterize the tangent space $T_X S_M$ and its dimension $p$ over $\mathbb{F}$, and

2. construct a matrix $B$ with orthonormal columns, if possible, such that for $E \in T_X S_M$, $\text{vec}(E) = By$ for some vector $y \in \mathbb{F}^p$.

To characterize the tangent space we need to use the vec-permutation matrix $P$ given in Lemma 1.21. It follows that $\text{vec}(e_i e_j^T) = P \text{vec}(e_j e_i^T)$ so

$$Pe_{(i-1)n+j} = e_{(j-1)n+i}, \quad i, j = 1, \ldots, n.$$ 

Hence

$$(P - I)e_{(i-1)n+i} = 0, \quad i = 1, \ldots, n, \quad (4.17)$$

$$(P - \sigma I)(e_{(i-1)n+j} + \sigma e_{(j-1)n+i}) = 0, \quad \sigma = \pm 1, \quad 1 \leq i < j \leq n, \quad (4.18)$$

from which it follows that $P$ has eigenvalue 1 with multiplicity $n(n + 1)/2$ and eigenvalue $-1$ with multiplicity $n(n - 1)/2$, [29].

### 4.3.1 When $S_M$ is a Lie or a Jordan algebra

The tangent space of any vector subspace is the vector subspace itself, which greatly simplifies the theory. Hence, for any $X \in S_M$ with $S_M \in \{J_M, L_M\}$,

$$T_X S_M = S_M.$$ 

To construct a basis for $T_X S_M$, we start with bilinear forms first and then use the results to construct a basis when the scalar product is a sesquilinear form. We need to assume that $M = \mu M^T \in \mathbb{R}^{n \times n}$ with $\mu = \pm 1$ but we do not need to assume that $M$ is orthogonal as in [17].

**Bilinear forms**

$$E \in S_M \iff E^* = sE \iff E^T M - sM E = 0,$$ 

where $s$ is given as

$$s = \begin{cases} 
1 & \text{if } S_M = J_M, \\
-1 & \text{if } S_M = L_M.
\end{cases} \quad (4.20)$$
4.3. COMPUTING $\text{COND}_{\text{STRUC}}(F, X)$ FOR A GIVEN SMOOTH MANIFOLD $S_M$

Applying the vec operator and on using the property 4 from Theorem 1.20 and Lemma 1.21 we find that

$$((M^T \otimes I_n)P - s(I_n \otimes M))\text{vec}(E) = 0. \quad (4.21)$$

If we write $\text{vec}(E) = B_{S_M}y$ for some matrix $B_{S_M}$ of full column rank $p$ and $y \in \mathbb{R}^p$, then (4.21) shows that $\text{range}(B_{S_M}) \subseteq \text{null}((M^T \otimes I_n)P - s(I_n \otimes M))$. Equality holds since we assume $M = \mu M^T$ with $\mu \in \{+1, -1\}$. Indeed, in this case, (4.19) shows that $ME = s\mu(ME)^T$, i.e., $ME$ is either (complex) symmetric or (complex) skew-symmetric so that $ME$ and therefore $E$ since $M$ is nonsingular, depend on $p = n(n + s\mu)/2 = \text{rank}(B_{S_M}) = \dim_K(T_XS_M)$ parameters that are real if $K = \mathbb{R}$ and complex otherwise. The property $P(A \otimes C) = (C \otimes A)P$ combined with $M = \mu M^T$ implies that

$$(M^T \otimes I_n)P - s(I_n \otimes M) = (\mu P - sI_{n^2})(I_n \otimes M).$$

Since $P$ has eigenvalues 1 with multiplicity $n(n + 1)/2$ and $-1$ with multiplicity $n(n - 1)/2$, and $M$ is nonsingular, we have that $\text{rank}((M^T \otimes I_n)P - s(I_n \otimes M)) = \text{rank}((\mu P - sI_{n^2})(I_n \otimes M)) = n(n - \mu s)/2$, so that the dimension of the null space of $(\mu P - sI_{n^2})(I_n \otimes M)$ is $n(n + \mu s)/2$. Hence,

$$\text{range}(B_{S_M}) = \text{null}((P - \mu sI_{n^2})(I_n \otimes M)).$$

If we define $D_{S_M} \in \mathbb{R}^{n^2 \times p}$ by

$${D_{S_M}} = \begin{cases} [\tilde{D}_{\mu s} \hat{I}] & \text{if } \mu s = 1, \\ \tilde{D}_{\mu s} & \text{if } \mu s = -1, \end{cases} \quad (4.22)$$

where $\tilde{D}_{\mu s} \in \mathbb{R}^{n^2 \times n(n-1)/2}$ has for columns the $n(n - 1)/2$ unit vectors $(e_{(i-1)n+j} + \mu s e_{(j-1)n+i})/\sqrt{2}, \quad 1 \leq i < j \leq n,$

and $\hat{I} \in \mathbb{R}^{n^2 \times n}$ has for columns the vectors $e_{(i-1)n+j}, \ i = 1, \ldots n$, then $D_{S_M}$ has full rank and orthonormal columns, and from (4.17)–(4.18) we have that $(P - \mu sI_{n^2})D_{S_M} = 0$. We can then construct $B_{S_M}$ as

$$B_{S_M} = (I_n \otimes M^{-1})D_{S_M}, \quad (4.23)$$
and this construction can be done in at most $O(n^3)$ operations.

Note that if $M$ is orthogonal then $B_{SM} = \mu(I_n \otimes M)D_{SM}$ has orthonormal columns and its construction is essentially computation free. In this case,

$$\text{cond}_{\text{struc}}(f, X) = \|K_f(X)(I_n \otimes M)D_{SM}\|_2,$$

(4.24)

which can be computed by using Algorithm 4.8. Davies showed in [17, Thm. 3.3] that,

$$\text{cond}_{\text{struc}}(f, X) = \frac{1}{2}\|K_f(X)P(I_n \otimes M)(P - \mu s I_n^2)\|_2,$$

which is a less compact expression than in (4.24) since $D_{SM}$ has $p = n(n + \mu s)/2 < n^2$ columns.

When $M$ is not orthogonal then unless it has a special structure, orthonormalizing the columns of $B_{SM}$ costs $O(n^6)$ operations. Instead of orthonormalizing $B_{SM}$, we can use (4.12) to obtain the lower and upper bounds for $\text{cond}_{\text{struc}}(f, X)$ that are cheaper to compute than $\text{cond}_{\text{struc}}(f, X)$. We note that $\|B_{SM}\|_2 \leq \|M^{-1}\|_2$. The Moore-Penrose inverse of a full column rank matrix $A$ is the minimal left inverse, i.e., for any left inverse $A^L$ of $A$, $\|A^L\|_2 \leq \|A^L\|_2$. For an upper bound on $\|B_{SM}^+\|_2$, $D_{SM}^T(I \otimes M)$ is a left inverse for $B_{SM}$ and hence,

$$\|B_{SM}^+\|_2 \leq \|D_{SM}^T(I \otimes M)\|_2 \leq \|M\|_2.$$

Then (4.12) yields the following lower and upper bounds on the structured condition number,

$$\frac{\|K_f(X)B_{SM}\|_2}{\|M^{-1}\|_2} \leq \text{cond}_{\text{struc}}(f, X) \leq \|K_f(X)B_{SM}\|_2\|M\|_2.$$

(4.25)

By exploiting the special structure of $B_{SM}$ in (4.23), approximating $\|K_f(X)B_{SM}\|_2$ using Algorithm 4.9 costs $O(kn^3)$ operations, where $k$ is the number of iterations. So the lower bound in (4.25) and an estimate for the upper bound can be computed in $O(kn^3)$ operations.

**Sesquilinear forms**

When the scalar product is a sesquilinear form, we have that

$$E \in T_X \mathbb{S}_M = \mathbb{S}_M \iff E^* = sE \iff \begin{cases} (\text{Re}(E))^* = s \text{Re}(E), \\ (\text{Im}(E))^* = -s \text{Im}(E), \end{cases}$$
4.3. COMPUTING $\text{COND}_{\text{STRUC}}(F, X)$ FOR A GIVEN SMOOTH MANIFOLD $S_M$ where $s$ is given as in equation (4.20). Hence, if $E \in \mathbb{J}_M$ then $\text{Re}(E) \in \mathbb{J}_M$ and $\text{Im}(E) \in \mathbb{L}_M$ so that
\[
\text{vec}(E) = \text{vec}(\text{Re}(E)) + i\text{vec}(\text{Im}(E)) = B_{\mathbb{J}_M} x + iB_{\mathbb{L}_M} y = [B_{\mathbb{J}_M}\ iB_{\mathbb{L}_M}] \begin{bmatrix} x \\ y \end{bmatrix} =: Bz
\]
for some $z \in \mathbb{R}^{n^2}$. The matrices $B_{\mathbb{J}_M}$ and $B_{\mathbb{L}_M}$ are as in (4.23) so that the $n^2 \times n^2$ matrix
\[
\tilde{B}_{\mathbb{J}_M} := (I_n \otimes M^{-1})[D_{\mathbb{J}_M}\ iD_{\mathbb{L}_M}]
\]
forms a basis over the base field $\mathbb{R}$ for $T_X \mathbb{J}_M$. Similarly, we find that
\[
\tilde{B}_{\mathbb{L}_M} := (I_n \otimes M^{-1})[D_{\mathbb{L}_M}\ iD_{\mathbb{J}_M}]
\]
(4.26)
forms a basis over $\mathbb{R}$ for $T_X \mathbb{L}_M$. Both $\tilde{B}_{\mathbb{J}_M}$ and $\tilde{B}_{\mathbb{L}_M}$ are unitary when $M$ is unitary.

4.3.2 When $S_M$ is an automorphism group

The next lemma provides an explicit characterization of the tangent space $T_X \mathbb{G}_M$ to an automorphism group at $X$.

**Lemma 4.10.** Let $\mathbb{G}_M$ be the automorphism of a scalar product $\langle \cdot, \cdot \rangle_M$ on $\mathbb{K}^n$ and let $X \in \mathbb{G}_M$. Then, the tangent space at $X$ to $\mathbb{G}_M$ is given by
\[
T_X \mathbb{G}_M = \{ E \in \mathbb{K}^{n\times n} \mid E = XF, F \in \mathbb{L}_M \},
\]
where $\mathbb{L}_M$ is the Lie algebra associated with $\langle \cdot, \cdot \rangle_M$.

**Proof.** We only prove the lemma for bilinear forms, the proof for sesquilinear forms is analogous. We start by showing that
\[
T_X \mathbb{G}_M = \{ E \in \mathbb{K}^{n\times n} \mid E^T MX + X^T ME = 0 \}.
\]

By definition, $E \in T_X \mathbb{G}_M$ is equivalent to the existence of a smooth curve $\gamma(t) \in \mathbb{G}_M$ satisfying $\gamma(0) = X, \gamma'(0) = E$. Now $\gamma(t) \in \mathbb{G}_M$ is equivalent to $\gamma(t)^T M \gamma(t) = M$. By differentiating the latter equation and evaluating at $t = 0$, we obtain
\[
\gamma'(0)^T M \gamma(0) + \gamma(0)^T M \gamma'(0) = 0.
\]
Substituting $\gamma(0) = X$ and $\gamma'(0) = E$ gives

$$E^T MX + X^T ME = 0. \tag{4.28}$$

Conversely, suppose that $E$ satisfies (4.28). Since $X \in \mathbb{G}_M$, it satisfies $X^T MX = M$ so that

$$(X + tE)^T M(X + tE) = M + \mathcal{O}(t^2).$$

We deduce that there exists a smooth curve $\gamma(t) = X + tE + \sum_{i=2}^{\infty} G_i t^i \in \mathbb{G}_M$ satisfying $\gamma(0) = X$ and $\gamma'(0) = E$ so that $E \in T_X \mathbb{G}_M$.

Defining $F := X^{-1} E$ (note that $X \in \mathbb{G}_M$ implies that $X$ is nonsingular) we can rewrite (4.28) as $F^T M + MF = 0$, which shows that $-F = M^{-1} F^T M = F^*$, i.e., $F \in \mathbb{L}_M$.

It follows from Lemma 4.10 that any $E \in T_X \mathbb{G}_M$ can be written as $E = XF$ with $F \in \mathbb{L}_M$. If $M = \mu M^T$ with $\mu = \pm 1$ then

$$\text{vec}(E) = \text{vec}(XF) = (I_n \otimes X) \text{vec}(F) = (I_n \otimes X) Q y,$$

where

- for bilinear forms, $Q = B_{L,M}$ is as in (4.23) with $S_M = \mathbb{L}_M$ and $y \in \mathbb{K}^p$ with $p = n(n - \mu)/2$,

- for sesquilinear forms, $Q = \tilde{B}_{L,M}$ is as in (4.26) and $y \in \mathbb{R}^p$ with $p = n^2$.

If we write $\text{vec}(E) = B_{G,M} y$ then

$$B_{G,M} = \begin{cases} 
(I_n \otimes XM^{-1}) D_{L,M}, & \text{for bilinear forms,} \\
(I_n \otimes XM^{-1}) [D_{L,M} i D_{L,M}], & \text{for sesquilinear forms.} 
\end{cases} \tag{4.29}$$

Unless $M$ and $X$ are orthogonal or unitary, $B_{G,M}$ does not have orthonormal columns and orthonormalizing these columns can cost as much as $O(n^6)$ operations. As for the Lie and Jordan algebra case, we can use (4.12) to obtain lower and upper bounds for $\text{cond}_{\text{struc}}(f, X)$ that are cheaper to compute than $\text{cond}_{\text{struc}}(f, X)$. Note that $\|B_{G,M}\|_2 \leq \|M^{-1}\|_2 \|X\|_2$ and since $D_{L,M}^T (I \otimes X^T M)$ is a left inverse for $B_{G,M}$,

$$\|B_{G,M}^+\|_2 \leq \|D_{L,M}^T (I \otimes X^T M)\|_2 \leq \|X\|_2 \|M\|_2.$$
4.3. COMPUTING $\text{COND}_{\text{STRUC}}(F, X)$ FOR A GIVEN SMOOTH MANIFOLD $S_M$

Then (4.12) yields the following lower and upper bounds on the structured condition number,

$$
\frac{\|K_f(X)B_{G_M}\|_2}{\|M^{-1}\|_2\|X\|_2} \leq \text{cond}_{\text{struc}}(f, X) \leq \|K_f(X)B_{G_M}\|_2\|X\|_2\|M\|_2.
$$

(4.30)

As for the bounds in (4.25), the cost of computing the lower and upper bounds in (4.30) is $O(kn^3)$ operations, where $k$ is the number of iterations performed by Algorithm 4.9.

If $M$ is orthogonal then $\|X^{-1}\|_F = \|X\|_F$ since $X^{-1} = X^* = M^{-1}X^*M$. Hence, $\|X\|_F = \kappa_F(X)^{1/2} := \|X\|_F\|X^{-1}\|_F$ so that the quality of the bounds in (4.30) is likely to be influenced by the condition number $\kappa_F(X)$ of $X$. In particular, if $X$ is well-conditioned, i.e., $\kappa_F(X) \approx 1$ then $\|K_f(X)B_{G_M}\|_2/\|f(X)\|_F$ offers a good estimate of the relative structured condition number since

$$
\frac{\|K_f(X)B_{G_M}\|_2}{\|f(X)\|_F} \leq \text{cond}_{\text{struc}}(f, X) \leq \kappa_F(X) \frac{\|K_f(X)B_{G_M}\|_2}{\|f(X)\|_F}.
$$

4.3.3 When $S_M$ is doubly structured

This part is given for real submanifolds of the $n^2$-dimensional real vector space $\mathbb{R}^{n \times n}$ and we assume that $f$ is real differentiable. The set of symplectic-orthogonal matrices $\widetilde{G} = G_I \cap G_J$ (We take $M = I$ or $M = J$) is given by

$$
\widetilde{G} = \{X \in \mathbb{R}^{n \times n} \mid X^TX = I, \; X^TJX = J\}.
$$

(4.31)

The characterization of the tangent space to the set of symplectic-orthogonal matrices is given as follows.

**Lemma 4.11.** Let $\widetilde{G}$ be the set of symplectic-orthogonal matrices and let $X \in \widetilde{G}$.

Then the tangent space to $\widetilde{G}$ at $X$ is given by

$$
T_X\widetilde{G} = \{E \in \mathbb{R}^{n \times n} \mid E = XH, \; H \in L_I \cap L_J\}.
$$

**Proof.** By using Lemma 4.10 we can write $H \in L_I \cap L_J$. From the definition of tangent spaces if $X \in \widetilde{G}$ we have that

$$
(X + tE)^T(X + tE) = I + O(t^2), \quad (X + tE)^TJ(X + tE) = J + O(t^2),
$$

where $O(t^2)$ represents higher order terms in $t$.
which yields $T_X \tilde{G} \subseteq T_X (G_I) \cap T_X (G_J)$. For the proof of other side we deduce that there exists a smooth curve $\gamma(t) = X + tE + \sum_{i=2}^{\infty} G_i t^i \in \tilde{G}$ satisfying $\gamma(0) = X$ and $\gamma'(0) = E$ so that $E \in T_X \tilde{G}$. Hence

\[
T_X \tilde{G} = \{ E \in \mathbb{R}^{n \times n} \mid E = XH, \ H \in \mathbb{L}_I \cap \mathbb{L}_J \}. \quad \square
\]

Lemma 4.11 indicates that any $E \in T_X \tilde{G}$ can be written as $E = XH$ with $H \in \mathbb{L}_I \cap \mathbb{L}_J$ and then

\[
\text{vec}(E) = \text{vec}(XH) = (I_n \otimes X)\text{vec}(H) = (I_n \otimes X)\tilde{D}y =: B_{\tilde{G}}y,
\]

where

\[
B_{\tilde{G}} = \begin{cases} (I_n \otimes X)\tilde{D}_{l_{IJ}}, & \text{for bilinear forms}, \\ (I_n \otimes X)[\tilde{D}_{l_{IJ}} i\tilde{D}_{\tilde{t}_{IJ}}], & \text{for sesquilinear forms}. \end{cases} (4.32)
\]

For $\tilde{D}_{l_{IJ}}$ the matrix $H$ is skew-symmetric and Hamiltonian, which has the block form

\[
H = \begin{bmatrix} F & G \\ -G & F \end{bmatrix}, \quad F = -F^T, \ G = G^T, \ F, G \in \mathbb{R}^{m \times m}, \ 2m = n
\]

so that $E$ depends on $p = m(m - 1)/2 + m(m + 1)/2 = m^2$ parameters. For $\tilde{D}_{\tilde{t}_{IJ}}$ $H$ is symmetric and skew-Hamiltonian, which has the same block form as (4.33) with $F = F^T$ and $G = -G^T$ for $F, G \in \mathbb{R}^{m \times m}$.

The pattern matrix $\tilde{D}$ can be constructed as follows. If we let

\[
H = \begin{bmatrix} F & G \\ -G & F \end{bmatrix}, \ \text{vec}(H) = y,
\]

then

\[
\tilde{D}(i, :) = \begin{cases} 0 & \text{if } y(i) = 0, \\ \text{sign}(y(i))e_i^T & \text{otherwise}, \end{cases} \quad i = 1, \ldots, n^2,
\]

where $e_k$ is the $k$th column of the $m^2 \times m^2$ identity matrix and

\[
\tilde{D} = \begin{cases} \tilde{D}_{l_{IJ}} & \text{if } F = -F^T \text{ and } G = G^T, \\ \tilde{D}_{\tilde{t}_{IJ}} & \text{if } F = F^T \text{ and } G = -G^T. \end{cases}
\]

If we normalize the columns of $\tilde{D}$ then $B_{\tilde{G}} = (I_n \otimes X)\tilde{D}$ has orthonormal columns since $X$ is orthogonal. Note that matrix-vector multiplication with $B_{\tilde{G}}$ can be done in $O(n^2)$ operations by exploiting the structure of $B_{\tilde{G}}$. 

4.4 Examples

The examples here are given for the real submanifolds of the \( n^2 \)-dimensional real
vector space \( \mathbb{R}^{n \times n} \). Let the \( 2 \times 2 \) diagonal symplectic matrix \( X \) be given by (with
\( M = J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \)).

\[
X = \begin{bmatrix} e^a & 0 \\ 0 & e^{-a} \end{bmatrix}, \quad a > 0. \tag{4.34}
\]

The method will be applied for the matrix logarithm and the square root.

**Principal logarithm**

For \( X \in \mathbb{G}_M, \log X \in \mathbb{L}_M \). Indeed

\[
Y := \log X = \begin{bmatrix} a & 0 \\ 0 & -a \end{bmatrix} \tag{4.35}
\]

is Hamiltonian. To compute the unstructured condition number we construct \( K_{\log}(X) \),
one column at a time using \( K_{\log}(X)e_{i+2j-2} = \text{vec}(L_{\log}(X,e_ie_j^T)) \), \( i, j = 1, 2 \) as in (1.5).

Since \( X \) and \( e_ie_j^T \) commute by using [35, Prob. 3.8]

\[
L_{\log}(X,e_ie_j^T) = X^{-1}e_ie_j^T, \quad i, j = 1, 2.
\]

For \( i \neq j \), \( X + e_ie_j^T \) is triangular and we can use [35, p. 84] to compute \( L_{\log}(X,e_ie_j^T) \),
\( i \neq j \). We find that

\[
K_{\log}(X) = \text{diag} \left( e^{-a}, \frac{a}{\sinh a}, \frac{a}{\sinh a}, e^a \right).
\]

Then absolute and relative condition numbers are given by

\[
\text{cond}(\log, X) = \|K_{\log}(X)\|_2 = e^a, \quad \text{cond}_{\text{rel}}(\log, X) = \frac{e^{2a}}{a},
\]

showing that the unstructured absolute and relative condition numbers increase rapidly
with \( a \).

In our example \( s = \mu = -1 \) so \( B_G \) in (4.29) and \( \widetilde{B}_G \), with orthonormal columns
are given as

\[
B_G = \begin{bmatrix}
\frac{\sqrt{2}}{2} e^a & 0 & 0 \\
0 & e^{-a} & 0 \\
0 & 0 & -e^a \\
-\frac{\sqrt{2}}{2} e^{-a} & 0 & 0
\end{bmatrix}, \quad \widetilde{B}_G = \begin{bmatrix}
\frac{e^{2a}}{\sqrt{1+e^{4a}}} & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
-\frac{1}{\sqrt{1+e^{4a}}} & 0 & 0
\end{bmatrix}.
\]
Hence

\[ \text{cond}_{\text{struc}}(\log, X) = \| K_{\log}(X) \tilde{B}_{\text{GJ}} \|_2 = \frac{a}{\sinh a} < 1 \]

and the ratio between structured and unstructured condition number

\[ \text{cond}_{\text{struc}}(\log, X)/\text{cond}(\log, X) = \frac{a}{e^a \sinh a} \]

exponentially decays as \( a \to \infty \). The lower and upper bounds in (4.30) yield

\[ \frac{a}{\sinh a} \leq \text{cond}_{\text{struc}}(\log, X) \leq \frac{ae^{2a}}{\sinh a} \]

showing that for this particular matrix and function \( f \), the lower bound is attained, whereas the upper bound is larger than \( \text{cond}(\log, X) \).

Applying the exponential to \( Y \) gives the symplectic matrix \( X \). For the structured condition number we follow the similar steps and find that

\[ K_{\exp}(Y) = \text{diag} \left( e^a, \frac{\sinh a}{a}, \frac{\sinh a}{a}, e^{-a} \right) = (K_{\log}(X))^{-1}. \]

To compute the structured condition number \( B_{\text{LJ}} = -(I_2 \otimes J) D_{\text{LJ}} \) so that

\[ \text{cond}_{\text{struc}}(\exp, Y) = \| K_{\exp}(Y) B_{\text{LJ}} \|_2 = \sqrt{\cosh 2a}. \]

Hence the ratio between the structured and unstructured condition number becomes

\[ \text{cond}_{\text{struc}}(\exp, Y)/\text{cond}(\exp, Y) = \frac{\sqrt{2}}{2} \left( 1 + e^{-4a} \right)^{1/2} \]

showing that \( \text{cond}_{\text{struc}}(\exp, Y) \approx \text{cond}(\exp, Y) \) for large values of \( a \).

**Principal square root**

For the symplectic matrix \( X \in \mathbb{G}_M, X^{1/2} \in \mathbb{G}_M \) [38], which is given as

\[ X^{1/2} = \begin{bmatrix} e^{a/2} & 0 \\ 0 & e^{-a/2} \end{bmatrix}, \quad a > 0. \]

The Kronecker form of the Fréchet derivative is obtained as

\[ K_{\text{sqrt}}(X) = \text{diag} \left( \frac{e^{-a/2}}{2}, \frac{1}{e^{a/2} + e^{-a/2}}, \frac{1}{e^{a/2} + e^{-a/2}}, \frac{e^{a/2}}{2} \right). \]

Then the unstructured condition number is

\[ \text{cond}(\text{sqrt}, X) = \| K_{\text{sqrt}}(X) \|_2 = \frac{e^{a/2}}{2}. \]
4.5 Numerical experiments

By using $\tilde{B}_{G,J}$, we compute the structured condition number with

$$\text{cond}_{\text{struc}}(\sqrt{X}) = \|K_{\sqrt{X}}\tilde{B}_{G,J}\|_2 = \frac{1}{e^{a/2} + e^{-a/2}},$$

which yields the ratio

$$\text{cond}_{\text{struc}}(\sqrt{X})/\text{cond}(\sqrt{X}) = \frac{2}{e^{a} + 1}.$$  

The ratio decays exponentially as $a \to \infty$.

For the inverse of the square root let the symplectic matrix $X^{1/2} = A$ be given. Then by using $L_{A^2}(A,E) = AE + EA$ [35, eq. 3.10] we obtain

$$K_{A^2}(A) = \text{diag} \left(2e^{a/2}, e^{a/2} + e^{-a/2}, e^{a/2} + e^{-a/2}, 2e^{-a/2}\right).$$

Then we get $\|K_{A^2}(A)\| = 2e^{a/2}$. For the structured condition number

$$\tilde{B}_{G,J} = \begin{bmatrix}
\frac{e^{a}}{\sqrt{1 + e^{2a}}} & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & -1 \\
-\frac{1}{\sqrt{1 + e^{2a}}} & 0 & 0
\end{bmatrix},$$

yields

$$\|K_{A^2}(A)\tilde{B}_{G,J}\|_2 = \left(4\left(\frac{e^{4a}+1}{e^{a}(e^{2a}+1)}\right)\right)^{1/2}.$$  

The ratio between the structured and unstructured condition number

$$\text{cond}_{\text{struc}}(A^2, A)/\text{cond}(A^2, A) = \left(\frac{e^{4a}+1}{e^{2a}(e^{2a}+1)}\right)^{1/2}$$

implies that $\text{cond}_{\text{struc}}(A^2, A) \approx \text{cond}(A^2, A)$ for large values of $a$.

4.5 Numerical experiments

We compare the unstructured and structured condition numbers numerically and illustrate the lower and the upper bounds. Another aim of this section is to compare the relative error of the matrix function with the unstructured and the structured upper bounds on the relative error. The experiments have been carried out in double precision arithmetic, running MATLAB R2014b with the unit roundoff $u \approx 1.1 \times 10^{-16}$.
on a MacBook Pro Retina with an Intel Core i7 2.6 GHz processor. We need to compute the Fréchet derivatives in Algorithm 4.8 and 4.9. For the matrix logarithm we use `logm_frechet_pade` from Higham’s Matrix Function Toolbox [30] to compute \( L_{\log}(X,E) \). For the matrix square root, \( L_{X^{1/2}}(X,E) \) is the solution to the Sylvester equation \( XL_{X^{1/2}} + L_{X^{1/2}}X = E \) [35, p.134], which can be computed with the function `sylvsol` from [30]. The unstructured and structured condition numbers are computed by Algorithm 4.8 with \( B = I_n \) and \( B = B_{GM} \), respectively.

As automorphism groups we consider the real symplectic group \((M = J, K = \mathbb{R})\), the real perplectic group \((M = R, K = \mathbb{R})\) and the real pseudo-orthogonal group \((M = \Sigma_{p,q}, K = \mathbb{R})\), where \( J, R, \Sigma_{p,q} \) are defined in Table 1.1 and use Jagger’s MATLAB Toolbox for Classical Matrix Groups [41] to generate random matrices with specified condition number in these groups. We check that the generated matrices \( X \) have no eigenvalues on the negative real line so that their principal logarithm \( \log X \) and principal square root \( X^{1/2} \) exist. For each matrix \( X \in GM \), we construct the factored form of the pattern matrix \( B_{GM} \) in (4.29) with \( D_{SM} \) as in (4.22) with \( \mu = s = -1 \) if \( X \) is symplectic, and \( \mu = 1, s = -1 \) if \( X \) is perplectic or pseudo-orthogonal.

First, we study the structured and unstructured condition numbers for the matrix logarithm and the square root by analysing sample sets of 1000 symplectic matrices. Each set is characterized by having a certain fixed condition number computed by \( \kappa_2(X) = \|X\|_2\|X^{-1}\|_2 \). In Figure 4.2, 4.3 and 4.4, data \( \rho_f(X) \) gives the following

\[
\rho_f(X) = \frac{\text{cond}_{\text{struc}}(f(X), X)}{\text{cond}(f(X), X)},
\]

which produces the distribution of the structured condition number by using `proplot` in MATLAB.

In Figures 4.5 and 4.6 the lower and upper bounds on the relative structured condition number are illustrated with the computed structured and unstructured condition numbers for the matrix logarithm and the square root. The lower and upper bounds are estimated by using Algorithm 4.9, in which the maximum number of iteration of the power method is taken as 10. The results are ordered according to the increase of \( \kappa_2(X) \) in Figures 4.5 and 4.6. The plots show that the upper bound is not sharp in general especially when \( \kappa_2(X) \) is large as expected. Lower bounds offer a good
4.5. NUMERICAL EXPERIMENTS

(a) Symplectic matrices with $\kappa_2(X) = 10^2$

(b) Symplectic matrices with $\kappa_2(X) = 10^7$

Figure 4.2: Cumulative distribution function for $\rho_{\log}$, empirically estimated for random $10 \times 10$ symplectic matrices.

estimation for the structured condition numbers. For pseudo-orthogonal matrices the lower bound tends to be sharper when the difference $p - q$ is small.

The unstructured upper bound on the relative error is given by

$$\frac{\|f(X + E) - f(X)\|_F}{\|f(X)\|_F} \lesssim \text{cond}_{\text{rel}}(f(X), X) \frac{\|E\|_F}{\|X\|_F}.$$  

By substituting $\text{cond}_{\text{struc}} \frac{\|X\|_F}{\|f(X)\|_F}$ for $\text{cond}_{\text{rel}}(f(X), X)$ we obtain the structured upper bound on the relative error. In Figure 4.7 and 4.8 we obtain approx bound line by estimating the structured condition number instead of computing exactly. In the relative error “exact” value of $f(X)$ is computed by using the Symbolic Math Toolbox. The matrix $X$ is diagonalized in 100 digit precision as $X = VDV^{-1}$ and $f(X)$ is computed by $f(X) = VDV^{-1}$. For the computed matrix logarithm we use Algorithm 2.17 and to obtain the computed square root, we use the structure preserving and cubically converging iteration

$$Y_{k+1} = \frac{1}{3} Y_k [I + 8 (I + 3Z_k Y_k)^{-1}], \quad Y_0 = X, \quad (4.36a)$$

$$Z_{k+1} = \frac{1}{3} [I + 8 (I + 3Z_k Y_k)^{-1}] Z_k, \quad Z_0 = I, \quad (4.36b)$$

where $Y_k, Z_k \in G_M$ and $Y_k \rightarrow X^{1/2}$ ([38, Sec.6]).

As we see from Figures 4.7 and 4.8, in which the results are ordered according to the relative error the structured relative error bound is much closer to the relative error comparing to the unstructured bound especially for the ill conditioned matrices.
In conclusion, in the experiments the distribution of the structured condition number shows that the structured condition number is more likely less than the unstructured condition number for the ill conditioned matrices. We also present the lower and upper bounds of the structured condition number, which are cheaper to compute. We see that the lower bound gives a good estimation for the structured condition number. Finally we show the relation between the lower bound of the structured condition number and the structured relative error, which motivates the usage of the structure preserving algorithms.
(a) Logarithm of symplectic-orthogonal matrices (b) Square root of symplectic-orthogonal matrices

Figure 4.4: Cumulative distribution function for $\rho_{\log}$ (on the left) and $\rho_{1/2}$ (on the right), empirically estimated for random $10 \times 10$ symplectic–orthogonal matrices ($\kappa_2(X) = 1$).
Figure 4.5: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal logarithm of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$.
4.5. NUMERICAL EXPERIMENTS

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$. 

Figure 4.6: Structured and unstructured relative condition numbers, and lower and upper bounds on the structured condition number for the principal square root of $10 \times 10$ randomly generated matrices in an automorphism group with increasing condition number $\kappa_2(X)$. 

(a) Symplectic matrices.  
(b) Perplectic matrices.  
(c) Pseudo-orthogonal matrices with $p = 6, q = 4$.  
(d) Pseudo-orthogonal matrices with $p = 9, q = 1$.
4. STRUCTURED CONDITION NUMBER

Figure 4.7: Error bounds on the relative error $\|f(X + E) - f(X)\|_F/\|f(X)\|_F$ for the logarithm (left plot) and the square root (right plot) of 50 randomly generated $12 \times 12$ symplectic matrices $X$ with $\kappa_2(X) = 10^2$.

Figure 4.8: Error bounds on the relative error $\|f(X + E) - f(X)\|_F/\|f(X)\|_F$ for the logarithm (left plot) and the square root (right plot) of 50 randomly generated $12 \times 12$ symplectic matrices $X$ with $\kappa_2(X) = 10^9$. 
Structured Condition Number for Matrix Factorizations

5.1 Introduction

This chapter aims to apply the theory of structured condition number to matrix factorizations. We are only interested in real manifolds of the $n^2$-dimensional real vector space $\mathbb{R}^{n \times n}$ for the generalized polar decomposition (GPD), structured polar decomposition and structured sign factorization of matrices in $\mathbb{G}_M$, $\mathbb{L}_M$ and $\mathbb{J}_M$, where $\mathbb{G}_M$, $\mathbb{L}_M$ and $\mathbb{J}_M$ are the automorphism group, Lie and Jordan algebras of a scalar product $\langle \cdot, \cdot \rangle_M$ as defined in Chapter 1. The contribution of this work is significant in a sense that it compares the structured and unstructured condition number for the matrix factorizations and evaluates whether we need a structure preserving numerical methods to compute the factors.

The chapter is organized as follows. In Section 5.2 we review the Fréchet derivative of the functions, which give the factors of the decomposition. Section 5.3 gives the algorithm using the power method to estimate the lower and upper bounds of the structured condition number. We present numerical tests comparing the structured and unstructured condition numbers of these matrix factorizations in Section 5.4. The findings of this study suggest that we should use structure preserving algorithms to compute some of the factors in the structured matrix factorizations.
Let $A \in \mathbb{R}^{n \times n}$ have a matrix factorization $A = A_1A_2$, where $A_1 \in \mathbb{R}^{n \times n}$ and $A_2 \in \mathbb{R}^{n \times n}$. Depending on the factorization type, existence and uniqueness conditions vary. Let the functions $f_1 : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ and $f_2 : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ define the mappings

$$f_1(A) = A_1, \quad f_2(A) = A_2.$$ 

In the rest of the work we assume that $f_1$ and $f_2$ are uniquely determined. In the computation of matrix factors, $A$ is subject to perturbations. Our aim is to understand the sensitivity of matrix factors to perturbations, which can be measured by a condition number. We can treat $f_1$ and $f_2$ as matrix functions and by using the definition of the Fréchet derivative of matrix functions (1.2) we can state the following lemma.

**Lemma 5.1.** Let $A \in \mathbb{R}^{n \times n}$ be factored as $A = A_1A_2$ and let $f_1(A) = A_1$ and $f_2(A) = A_2$. Then the Fréchet derivative $L_{f_j}(A, \Delta A)$ of $f_j$, $j = 1, 2$ satisfies the following

$$\|\Delta A_j - L_{f_j}(A, \Delta A)\| = o(\|\Delta A\|), \quad j = 1, 2.$$

In the sensitivity analysis of matrix factorizations there is no restriction on the structure of $A$ and the perturbed matrix $A + \Delta A$. For structured sensitivity analysis of matrix factorizations we restrict the perturbation $\Delta A$ to a smaller set by forcing the perturbed matrix $A + \Delta A$ to have the same structure as $A$.

### 5.2 Structured matrix factorization

Since we treat $f_1$ and $f_2$ as matrix functions we can use equation (4.8) to define the structured condition number. We state the computation of the structured and unstructured condition numbers for the generalized polar decomposition, the structured polar and sign factorizations.

#### 5.2.1 Generalized polar decomposition

**Definition 5.2.** Suppose $\mathbb{C}^n$ is a scalar product space with automorphism group $\mathbb{G}_M$ and Jordan algebra $\mathbb{J}_M$. Then for a matrix $A \in \mathbb{C}^{n \times n}$, a generalized polar decomposition with respect to this scalar product is a factorization $A = WS$, where $W \in \mathbb{G}_M$, $S \in \mathbb{J}_M$, and $\text{sign}(S) = I$. 
If $A$ has a generalized polar decomposition as stated in Definition 5.2 then the conditions for the uniqueness are given in the following theorem, where $A^*$ is defined in Definition 1.5.

**Theorem 5.3.** Let $C^n$ be an arbitrary scalar product space. Then a matrix $A \in C^{n \times n}$ has a GPD with respect to this scalar product if and only if $(A^*)^* = A$ and $A^*A$ has no eigenvalues on $\mathbb{R}^- := \{x \in \mathbb{R} : x \leq 0\}$. When such a factorization exists it is unique.

For more details about the generalized polar decomposition see [63]. We analyse the sensitivity of the generalized polar decomposition to perturbations by computing the condition number of the two factors. Similar analysis is done for the canonical generalized polar decomposition in [81].

Let $A = WS$ be the generalized polar decomposition of $A$ and let

$$f_1(A) = W, \quad f_2(A) = S.$$ 

Let $A + \Delta A \in \mathbb{R}^{n \times n}$ have the generalized polar decomposition

$$A + \Delta A = (W + \Delta W)(S + \Delta S). \quad (5.1)$$

We start with the perturbation analysis of the factor $S$. By using the property $A^*A = S^2$ we have that

$$M^{-1}(A + \Delta A)^T M (A + \Delta A) = (S + \Delta S)(S + \Delta S).$$

Reorganizing the equation gives

$$S\Delta S + \Delta SS = M^{-1}A^T M \Delta A + M^{-1}\Delta A^T M A + o(\|\Delta A\|).$$

Then $L_{f_2} := L_{f_2}(A, \Delta A)$ satisfies the following.

$$SL_{f_2} + L_{f_2}S = M^{-1}A^T M \Delta A + M^{-1}\Delta A^T M A.$$ 

Applying the vec operator yields

$$\text{vec}(L_{f_2}) = C_{SS}^{-1}C_{AM} \text{vec}(\Delta A),$$
where \( C_S = I \otimes S + S^T \otimes I \) and \( C_{AM} = (I \otimes M^{-1} A^T M) + (A^T M^T \otimes M^{-1})P \) with vec\( (A^T) = P \text{vec}(A) \). According to (1.3) the Kronecker form of \( L_{f_2} \) is given by \( K_{f_2} = C_S^{-1} C_{AM} \).

For the perturbation analysis of \( W \), it follows from (5.1) that
\[
\Delta W = \Delta AS^{-1} - W \Delta SS^{-1} + o(\|\Delta A\|).
\]
Then the Fréchet derivative \( L_{f_1} := L_{f_1}(A, \Delta A) \) of \( f_1 \) gives the following equality,
\[
L_{f_1} = \Delta AS^{-1} - W L_{f_2} S^{-1}.
\]
Applying the vec operator and substituting vec\( (L_{f_2}) \) into equation yield
\[
\text{vec}(L_{f_1}) = [(S^{-T} \otimes I) - (S^{-T} \otimes W) C_S^{-1} C_{AM} ] \text{vec}(\Delta A),
\]
which implies that the Kronecker form of \( L_{f_1} \) is given by
\[
K_{f_1} = [(S^{-T} \otimes I) - (S^{-T} \otimes W) C_S^{-1} C_{AM} ].
\]

**Lemma 5.4.** Let \( A \in \mathbb{C}^{n \times n} \) have generalized polar decomposition \( A = WS \) and let the functions \( f_1 \) and \( f_2 \) give the factors as
\[
f_1(A) = W, \quad f_2(A) = S. \tag{5.2}
\]
Then the Fréchet derivative of \( f_1 \) and \( f_2 \) satisfy the followings.
\[
L_{f_1}(A, \Delta A) = \Delta AS^{-1} - W L_{f_2}(A, \Delta A) S^{-1},
\]
\[
SL_{f_2}(A, \Delta A) + L_{f_2}(A, \Delta A) S = M^{-1} A^T M \Delta A + M^{-1} \Delta A^T M A.
\]

By using equation (4.13) the structured condition number of \( f_1 \) and \( f_2 \) are given in the next lemma.

**Lemma 5.5.** Let \( A \in S_M \in \{G_M, L_M, J_M\} \) have a generalized polar decomposition \( A = WS \) and let \( f_1, f_2 \) be as in (5.2). Let the orthonormal columns of \( B \) span the tangent space to \( S_M \) at \( A \). Then from (4.13) the structured condition numbers of \( f_1 \) and \( f_2 \) are respectively given by
\[
\text{cond}_{\text{struc}}(f_1, A) = \|((S^{-T} \otimes I) - (S^{-T} \otimes W) C_S^{-1} C_{AM}) B\|_2,
\]
\[
\text{cond}_{\text{struc}}(f_2, A) = \|C_S^{-1} C A B\|_2,
\]
where \( C_S = I \otimes S + S^T \otimes I \), \( C_{AM} = (I \otimes M^{-1} A^T M) + (A^T M^T \otimes M^{-1})P \).

**Proof.** The equalities come from equations (4.12) and (4.13). \qed
5.2. STRUCTURED MATRIX FACTORIZATION

5.2.2 Structured polar decomposition

For a square matrix, the structured polar decomposition is a special case of the generalized polar decomposition where the scalar product is the standard inner product $\langle x, y \rangle = x^* y$ on $\mathbb{C}^n$. We first define the unstructured polar decomposition.

**Theorem 5.6** (Polar decomposition, [27]). Let $A \in \mathbb{C}^{m \times n}$ with $m \geq n$. There exists a matrix $U \in \mathbb{C}^{m \times n}$ with orthonormal columns and a unique Hermitian positive semidefinite matrix $H \in \mathbb{C}^{n \times n}$ such that $A = UH$.

The importance of the polar decomposition stems from the fact that the unitary factor $U$ is the nearest unitary matrix to $A$ in any unitarily invariant norm [22]. The Hermitian factor is always unique, whereas the unitary factor is unique only if $A$ is nonsingular [40, Thm. 7.3.1]. By taking the nonsingular matrices in the automorphism group, Lie algebras and Jordan algebras we investigate the sensitivity of the factors when the perturbed matrix has the same structure as $A$.

Let $A \in \mathbb{R}^{n \times n}$ be factorised as $A = UH$ and

$$f_1(A) = U, \quad f_2(A) = H.$$ 

We will give the structured polar decomposition for the matrices in $S_M \in \{G_M, J_M, L_M\}$.

**Theorem 5.7.** [63, Thm. 5.1, 5.2] Let $S_M \in \{G_M, L_M, J_M\}$ and let $A = UH$ be the polar factorization of nonsingular $A$. Then

$$A \in S_M \in \{L_M, J_M\} \implies U \in S_M,$$

and

$$A \in G_M \implies U, H \in G_M.$$ 

Let $A + \Delta A \in \mathbb{R}^{n \times n}$ have the polar decomposition

$$A + \Delta A = (U + \Delta U)(H + \Delta H). \tag{5.3}$$

First we investigate the perturbation analysis of $H$. Since the polar factor $H$ satisfies $A^T A = H^2$ we get

$$\Delta H H + H \Delta H = A^T \Delta A + \Delta A^T A + o(\|\Delta A\|).$$
Then the Fréchet derivative $L_{f_2} := L_{f_2}(A, \Delta A)$ of $f_2$ satisfies
\[
L_{f_2}H + HL_{f_2} = A^T \Delta A + \Delta A^T A.
\]
Applying the vec operator gives
\[
\text{vec}(L_{f_2}) = C^{-1}_H C_A \text{vec}(\Delta A),
\]
where $C_H = H^T \otimes I + I \otimes H$ and $C_A = I \otimes A^T + (A^T \otimes I)P$.

For the perturbation analysis of $U$ we obtain
\[
\Delta U = \Delta AH^{-1} - U \Delta HH^{-1} + o(\|\Delta A\|).
\]
Then the Fréchet derivative $L_{f_1} := L_{f_1}(A, \Delta A)$ of $f_1$ satisfies
\[
L_{f_1} = \Delta AH^{-1} - UL_{f_2}H^{-1}.
\]
By applying the vec operator and later substituting $\text{vec}(L_{f_2})$ into the equation we get
\[
\text{vec}(L_{f_1}) = (H^T \otimes I)^{-1}(I_{n^2} - (I \otimes U)C^{-1}_H C_A)\text{vec}(\Delta A).
\]

The Kronecker form of the Fréchet derivatives $L_{f_1}$ and $L_{f_2}$ are given as
\[
K_{f_2}(A) = C^{-1}_H C_A,
\]
\[
K_{f_1}(A) = (H^T \otimes I)^{-1}(I_{n^2} - (I \otimes U)C^{-1}_H C_A).
\]

Lemma 5.8. Let $A \in \mathbb{C}^{n \times n}$ have a polar decomposition $A = UH$ and let the functions $f_1$ and $f_2$ be given as
\[
f_1(A) = U, \quad f_2(A) = H. \quad (5.4)
\]
Then the Fréchet derivatives of $f_1$ and $f_2$ satisfy the followings.
\[
L_{f_1}(A, \Delta A) = (\Delta A - UL_{f_2}(A, \Delta A))H^{-1},
\]
\[
HL_{f_2}(A, \Delta A) + L_{f_2}(A, \Delta A)H = A^T \Delta A + \Delta A^T A.
\]

Lemma 5.9. Let $A \in \mathbb{S}_M \in \{G_M, J_M, L_M\}$ have a structured polar decomposition $A = UH$ and let $f_1, f_2$ be as in (5.4). Let the orthonormal columns of $B$ span the
5.2. STRUCTURED MATRIX FACTORIZATION

The tangent space to $S_M$ at $A$. Then from (4.13) the structured condition numbers of $f_1$ and $f_2$ are respectively given by

$$\text{cond}_{\text{struc}}(f_1, A) = \|(H^T \otimes I)^{-1} (I_{n^2} - (I \otimes U) C^{-1}_H C_A^T) B\|_2,$$

$$\text{cond}_{\text{struc}}(f_2, A) = \|C^{-1}_H C_A B\|_2,$$

where $C_H = H^T \otimes I + I \otimes H$, $C_A = I \otimes A^T + (A^T \otimes I) P$.

Proof. The proof follows from equations (4.12) and (4.13). □

5.2.3 Structured sign decomposition

The matrix sign function was first developed by Roberts [71] to solve Riccati equations. The details can be found in the monograph [35, Chap. 5]. The matrix sign decomposition of $A \in \mathbb{C}^{n \times n}$ is defined for $A \in \mathbb{C}^{n \times n}$ having no pure imaginary eigenvalues as

$$A = SN, \quad S = \text{sign}(A), \quad N = (A^2)^{1/2}.$$ 

The structure of the factors are given in Theorem 5.10 when $A$ is in the automorphism group, Lie or Jordan algebras.

Theorem 5.10. [63, Thm. 4.1] Let $S_M \in \{G_M, L_M, J_M\}$ of an arbitrary scalar product, and suppose $A \in S_M$ has a sign decomposition $A = SN$. Then

$$A \in S_M \implies S \in S_M \text{ and } N \in \begin{cases} S_M & \text{when } S_M = G_M, J_M, \\ J_M & \text{when } S_M = L_M. \end{cases}$$

The sensitivity analysis of $S$ is given in [46] for the unstructured perturbation. Let $A \in \mathbb{R}^{n \times n}$ be factorised as $A = SN$ and

$$f_1(A) = S, \quad f_2(A) = N.$$ 

Let $A + \Delta A$ have a sign decomposition

$$A + \Delta A = (S + \Delta S)(N + \Delta N). \quad (5.5)$$

Since $SA = AS$ we can write

$$(A + \Delta A)(S + \Delta S) = (S + \Delta S)(A + \Delta A),$$
which gives
\[ A\Delta S - \Delta SA = S\Delta A - \Delta AS. \] (5.6)

We also have \((S + \Delta S)^2 = I\). Then we get
\[ S\Delta S + \Delta SS = o(\|\Delta A\|). \] (5.7)

Pre multiplying (5.6) by \(S\) and combining with (5.7) give
\[ N\Delta S + \Delta SN = \Delta A - S\Delta AS + o(\|\Delta A\|). \]

The Fréchet derivative \(L_{f_1} := L_{f_1}(A, \Delta A)\) of \(f_1\) satisfies
\[ NL_{f_1} + L_{f_1}N = \Delta A - S\Delta AS. \]

Applying the vec operator gives
\[ (I \otimes N + N^T \otimes I)\text{vec}(L_{f_1}) = (I_{n^2} - S^T \otimes S)\text{vec}(\Delta A), \]
which yields the Kronecker form of \(L_{f_1}(A, \Delta A)\)
\[ K_{f_1}(A) = C_N^{-1}(I_{n^2} - S^T \otimes S), \]
where \(C_N = I \otimes N + N^T \otimes I\).

The Fréchet derivative of \(f_1\) is stated in [35, Thm. 5.3] but the Fréchet derivative of \(f_2\) is not mentioned. For the perturbation analysis of \(N\), from (5.5) we get
\[ \Delta N = S^{-1}\Delta A - S^{-1}\Delta SN + o(\|\Delta A\|). \]

Then the Fréchet derivative \(L_{f_2} := L_{f_2}(A, \Delta A)\) of \(f_2\) satisfies the following
\[ L_{f_2} = S^{-1}\Delta A - S^{-1}L_{f_1}N. \]

Applying the vec operator and substituting \(\text{vec}(L_{f_1})\) yield
\[ \text{vec}(L_{f_2}) = [(I \otimes S^{-1}) - (N^T \otimes S^{-1})C_N^{-1}(I_{n^2} - S^T \otimes S)]\text{vec}(\Delta A), \]
which provides the Kronecker form of the Fréchet derivative of \(f_2\).
\[ K_{f_2}(A) = [I \otimes S^{-1} - (N^T \otimes S^{-1})C_N^{-1}(I_{n^2} - S^T \otimes S)]. \]
Lemma 5.11. Let $A \in \mathbb{C}^{n \times n}$ have a sign factorization $A = SN$ and let the functions $f_1$ and $f_2$ give the factors as

$$f_1(A) = S, \quad f_2(A) = N. \quad (5.8)$$

Then the Fréchet derivatives of $f_1$ and $f_2$ satisfy the followings.

$$NLf_1(A, \Delta A) + Lf_1(A, \Delta A)N = \Delta A - S\Delta AS, \quad Lf_2(A, \Delta A) = S^{-1}\Delta A - S^{-1}Lf_1(A, \Delta A)N.$$ 

Lemma 5.12. Let $A \in S_M \in \{G_M, L_M, J_M\}$ have a structured sign factorization $A = SN$ and let $f_1$, $f_2$ be as in (5.8). Let the orthonormal columns of $B$ span the tangent space to $S_M$ at $A$. Then from (4.13) the structured condition numbers of $f_1$ and $f_2$ are respectively computed by

$$\text{cond}_{\text{struc}}(f_1, A) = \|C_N^{-1}(I_{n^2} - S^T \otimes S)B\|_2,$$

$$\text{cond}_{\text{struc}}(f_2, A) = \|[I \otimes S^{-1} - (N^T \otimes S^{-1})C_N^{-1}(I_{n^2} - S^T \otimes S)]B\|_2,$$

where $C_N = I \otimes N + N^T \otimes I$.

5.3 Estimating the structured condition number

Since forming the Kronecker form of the Fréchet derivative is expensive instead of computing the exact structured condition number we can use the power method to estimate it. The Fréchet derivative of one of the mappings depends on the other’s Fréchet derivative so by adapting Algorithm 4.9 to the matrix factorization for $K = F = \mathbb{R}$ we give the estimation of the structured condition number in the following algorithms.

Algorithm 5.13. Given $A \in S_M \in \{G_M, L_M, J_M\}$ with the structured matrix factorization $A = A_1A_2$ and $f_1(A) = A_1$, $f_2(A) = A_2$, the algorithm uses the power method to estimate $\gamma \leq \|K_{f_i}(A)B\|_2$, $i = 1, 2$.

1. Start with $z_0$
2. for $k = 0$: maxit
5.5 Numerical experiments

We present numerical results for the structured condition number of matrix factorizations. All the experiments are carried out by using MATLAB R2014b with the unit roundoff $u \approx 1.1 \times 10^{-16}$. Test matrices in an automorphism group are generated by using Jagger’s MATLAB Toolbox [41]. Hamiltonian and skew-Hamiltonian matrices are built by using the structures in Table 2.1. We use Algorithm 4.8 with $K = F = \mathbb{R}$ to compute the structured condition number of $f_1$ and $f_2$ giving matrix factors and the associated $B$ is evaluated either by (4.23) or (4.29). In Algorithm 4.8 $B = I$ gives the unstructured condition number. We evaluate the Fréchet derivatives of matrix factors by using Lemma 5.4, 5.8 and 5.11. The lower and the upper bounds on the structured condition number are evaluated by using equation (4.25) and (4.30), in which $\|K_{fi}(A)B_{S,M}\|_2$ is estimated by using Algorithm 5.13. The number of iteration we have used in the power method is 10.

For a given matrix factorization $A = A_1 A_2$, let the functions $f_1 : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ and $f_2 : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ define the mappings

$$f_1(A) = A_1, \quad f_2(A) = A_2.$$ 

In the plots, while $A_{1\text{struc}}$ and $A_{2\text{struc}}$ represents the structured condition number of the functions $f_1$ and $f_2$, $A_{1\text{unstruc}}$ and $A_{2\text{unstruc}}$ represent the unstructured condition numbers, respectively. The pairs $A_{1\text{low}}, A_{1\text{upp}}, A_{2\text{low}}, A_{2\text{upp}}$ give the lower and upper bounds
of the structured condition number of $f_j, j = 1, 2$. The condition number of $A$ is computed by $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$. The results are ordered according to either the condition number of matrix $\kappa_2(A)$ or the experiment number.

The factors are computed as follows:

1. For the generalized polar decomposition the relations $A^*A = S^2$ and $W = AS^{-1}$ are used.

2. The orthogonal factor $U$ in the structured polar decomposition is computed by using the cubically convergent structure preserving iteration from [37], given by

$$X_{k+1} = \frac{1}{3}X_k[I + 8(I + 3X_k^T X_k)^{-1}], \quad X_0 = A,$$

where $\lim_{k \to \infty} X_k = U$ and $H$ is given by $H = U^T A$.

3. For $A \in \{L_M, J_M\}$, the factor $S$ in the structured sign factorization is computed by the function $\text{signm}$ from the Matrix Function Toolbox [30] using the Schur decomposition. For $A \in G_M$, $S$ is computed by the structure preserving iteration [37].

$$X_{k+1} = X_k(3I + X_k^2)(I + 3X_k^2)^{-1}, \quad X_0 = A,$$

where $\lim_{k \to \infty} X_k = S$. The $N$ factor is obtained by $N = SA$.

### 5.4.1 Numerical tests for the generalized polar decomposition

The structured, unstructured condition numbers and the lower, upper bounds on the structured condition number are presented for the generalized polar decomposition of $A \in S_M$ where $S_M \in \{G_M, L_M, J_M\}$.
Figure 5.1: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $W$ in plot (a) and the factor $S$ in plot (b) of the generalized polar decomposition of randomly generated pseudo-orthogonal matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in \mathbb{G}_M$) with increasing condition number $\kappa_2(A)$.

Figure 5.2: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $W$ in plot (a) and the factor $S$ in plot (b) of the generalized polar decomposition of randomly generated Hamiltonian matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in \mathbb{L}_M$) with $\kappa_2(A)$ changing between $10$ to $10^5$. 
5.4. NUMERICAL EXPERIMENTS

Figure 5.3:Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $W$ in plot (a) and the factor $S$ in plot (b) of the generalized polar decomposition of randomly generated skew-Hamiltonian matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in \mathbb{J}_M$) with $\kappa_2(A)$ changing between 10 to $10^5$.

While in Figure 5.1 and 5.2 the structured and unstructured condition number are almost the same, the gap between them is getting bigger in Figure 5.3 for the factor $S$. We get a good estimation for the lower bound on the structured condition number in Figure 5.2 for both factors.

From the experiments we can conclude that there is not much difference between the structured and unstructured condition numbers. We also observe that although the upper bound on the structured condition number is larger than the unstructured condition number in most of the cases there are counter examples.

5.4.2 Numerical tests for the structured polar decomposition

For the case where $S_M \in \{\mathbb{L}_M, \mathbb{J}_M\}$ only $U$ preserves the structure so we give the comparison between the structured and unstructured condition numbers for the $U$ factor only. We give the structured condition numbers for both factors $U$ and $H$ for $S_M = \mathbb{G}_M$. 

Figure 5.4: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $U$ in plot (a) and the factor $H$ in plot (b) of the structured polar decomposition of randomly generated symplectic matrices $A \in \mathbb{R}^{10\times 10} \:(A \in \mathbb{G}_M)$ with increasing condition number $\kappa_2(A)$.

Figure 5.5: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $U$ in plot (a) and the factor $H$ in plot (b) of the structured polar decomposition of randomly generated pseudo-orthogonal matrices $A \in \mathbb{R}^{10\times 10} \:(A \in \mathbb{G}_M)$ with increasing condition number $\kappa_2(A)$. 

5.4. NUMERICAL EXPERIMENTS

(a) Skew-symmetric $A \in \mathbb{R}^{10 \times 10}$ with $\kappa_2(A)$ changing between 10 to $10^3$
(b) Hamiltonian $A \in \mathbb{R}^{10 \times 10}$ with $\kappa_2(A)$ changing between 10 to $10^5$

Figure 5.6: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the structured polar decomposition of randomly generated skew-symmetric and Hamiltonian matrices $A \in \mathbb{L}_M$, representing the condition number of the factor $U$ in plot (a) and (b).

(a) Symmetric $A \in \mathbb{R}^{10 \times 10}$ with $\kappa_2(A)$ changing between 10 to $10^4$
(b) Skew-Hamiltonian $A \in \mathbb{R}^{10 \times 10}$ with $\kappa_2(A)$ changing between 10 to $10^4$

Figure 5.7: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the structured polar decomposition of randomly generated symmetric and skew-Hamiltonian matrices $A \in \mathbb{J}_M$, representing the condition number of the factor $U$ in plot (a) and (b).

Figure 5.4 and 5.5 suggest us to use structure preserving algorithms to compute
the factor $U$ for the matrices in $\mathbb{G}_M$. On the other hand, we observe no significant difference between the structured and unstructured condition numbers of the factor $U$ for the matrices in $\mathbb{L}_M$ and $\mathbb{J}_M$. We also observe that the unstructured condition number is smaller than the upper bound of the structured condition number for the matrices in $\mathbb{G}_M$, $\mathbb{L}_M$ and $\mathbb{J}_M$.

5.4.3 Numerical tests for the structured sign decomposition

![Figure 5.8](image)

Figure 5.8: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $S$ in plot (a) and the factor $N$ in plot (b) of the sign factorization of randomly generated symplectic matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in \mathbb{G}_M$) with increasing condition number $\kappa_2(A)$. 


5.4. NUMERICAL EXPERIMENTS

Figure 5.9: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $S$ in plot (a) and the factor $N$ in plot (b) of the sign factorization of randomly generated perplectic matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in \mathbb{G}_M$) with increasing condition number $\kappa_2(A)$.

Figure 5.10: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $S$ in plot (a) and the factor $N$ in plot (b) of the structured sign factorization of randomly generated Hamiltonian matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in \mathbb{L}_M$) with $\kappa_2(A)$ changing between $10$ to $10^5$. 
Figure 5.11: Structured and unstructured condition numbers and the lower and upper bounds on the structured condition number for the factor $S$ in plot (a) and the factor $N$ in plot (b) of the structured sign factorization of randomly generated skew-Hamiltonian matrices $A \in \mathbb{R}^{10 \times 10}$ ($A \in \mathcal{J}_M$) with $\kappa_2(A)$ changing between $10$ to $10^5$.

The experiments carried out for the sign factorization point out that we need structure preserving algorithms especially to compute the sign factor $S$. As seen from Figures 5.8 and 5.9 the unstructured and structured condition numbers differ by several orders of magnitude for the sign factor $S$ for the matrices in $\mathcal{G}_M$. On the other hand, as seen from Figures 5.8, 5.9, 5.10 and 5.11 the unstructured and structured condition numbers are almost the same for the factor $N$ for the matrices in $\mathcal{S}_M \in \{\mathcal{G}_M, \mathcal{L}_M, \mathcal{J}_M\}$ and for the factor $S$ for the matrices in $\mathcal{S}_M \in \{\mathcal{L}_M, \mathcal{J}_M\}$. We obtain a good estimation for the lower bound on the structured condition number for the matrices in $\mathcal{G}_M$. 
6

Perturbation Analysis for the Canonical Generalized Polar Decomposition

6.1 Introduction

In the literature there is much research on the perturbation analysis of the polar decomposition. Computation of the ordinary polar decomposition with the perturbation analysis is given in [31] by using the idea that the polar factors are differentiable. Strict bounds with no order terms are also derived in [5] by first differentiating the polar factors and deriving the upper bounds of the derivatives. Then the integrals are applied to these upper bounds to obtain the bounds on the perturbations. The results in [5] are valid for the Frobenius norm. They are strengthen by Mathias [64] and given for all the unitarily invariant norms. The perturbation bound for the unitary polar factor of full rank square matrices in [64] is improved by Li [52] for any unitarily invariant norm. The analysis of the perturbations for unitary and isometric matrices of the (generalized) polar decomposition under the Frobenius norm is given in [57]. Different bounds are obtained for the real and complex cases. The bound given in the previous paper is improved by [58] for many practical perturbation problems. The change in the unitary polar factor for the generalized polar decomposition is considered in [14].
Both multiplicative and additive perturbations are studied, showing neither type of bounds is generally uniformly better than the other. Absolute perturbation bounds of unitary polar factors in polar decompositions are given for rectangular complex matrices [59]. On the other hand, very little is known about the perturbation analysis of the canonical generalized polar decomposition (CGPD). This factorization is studied in [48] with its applications and numerical computation. It is also analyzed in [39] with the derivation of key properties in its theory and the iteration to compute the factors. The recent paper [81] introduces the normwise and componentwise perturbation bounds for the factors of the canonical generalized polar decomposition for real matrices.

In this work, we are interested in additive and multiplicative perturbation analysis of the canonical generalized polar decomposition of $A \in \mathbb{K}^{m \times n}$ ($\mathbb{K} = \mathbb{C}$ or $\mathbb{K} = \mathbb{R}$). We also bound the norm of the changes in the polar factors by using the standard technique as used in the ordinary polar decomposition [31]. Whereas we revise the additive perturbation analysis of the canonical generalized polar decomposition [81], we try to provide a new understanding of the perturbation analysis by using new methods. This work makes several contributions to the current literature. It serves (1) to give a theoretical aspect of the perturbation analysis of the canonical generalized polar decomposition by using the multiplicative perturbation analysis and differentiability, (2) to analyze the methods that we propose in detail and (3) to determine the best method giving the sharper bounds.

The structure of the chapter is as follows. In Section 6.2 we give some preliminaries for the canonical generalized polar decomposition. Additive perturbation analysis for real and complex matrices is provided in Section 6.3. Next, multiplicative perturbation analysis is presented in Section 6.4. The last technique for the perturbation analysis is discussed in Section 6.5. Finally, we give the numerical experiments and compare all these bounds in terms of sharpness in Section 6.6.
6.2 The canonical generalized polar decomposition

We first give some fundamental information for the canonical generalized polar decomposition [39].

**Definition 6.1.** A matrix $U \in \mathbb{K}^{m \times n}$ is a partial isometry if $\|Ux\|_2 = \|x\|_2$ for all $x \in \text{range}(U^*)$ or equivalently if $U^+ = U^*$, where $U^+$ denotes the Moore-Penrose inverse given in Definition 1.22.

**Definition 6.2.** Let $\mathbb{K}^m$ and $\mathbb{K}^n$ be equipped with scalar products induced by nonsingular matrices $M \in \mathbb{K}^{m \times m}$ and $N \in \mathbb{K}^{n \times n}$, respectively. For a matrix $A \in \mathbb{K}^{m \times n}$ the $(M,N)$-adjoint of $A$ is given by $A^{*,M,N}$ satisfying the identity

$$\langle Ax, y \rangle_M = \langle x, A^{*,M,N} y \rangle_N$$

for all $x \in \mathbb{K}^n$ and all $y \in \mathbb{K}^m$. We have the explicit expression for $A^{*,M,N}$

$$A^{*,M,N} = \begin{cases} N^{-1}A^T M, & \text{for bilinear forms}, \\ N^{-1}A^* M, & \text{for sesquilinear forms}. \end{cases}$$

We use the orthosymmetric scalar products, which is given in Definition 1.6 so under this assumption the conditions for the existence of the canonical generalized polar decomposition can be identified.

**Definition 6.3.** The nonsingular matrices $M \in \mathbb{K}^{m \times m}$ and $N \in \mathbb{K}^{n \times n}$ form an orthosymmetric pair if

$$M^T = \alpha M, \quad N^T = \alpha N, \quad \alpha = \pm 1, \quad \text{for bilinear forms},$$

$$M^* = \beta M, \quad N^* = \beta N, \quad \beta \in \mathbb{C}, \quad |\beta| = 1, \quad \text{for sesquilinear forms}.$$ 

The partial isometry related to an orthosymmetric pair $M$ and $N$ is given in the following definition.

**Definition 6.4.** $W \in \mathbb{K}^{m \times n}$ is a partial $(M,N)$-isometry if $WW^{*,M,N}W = W$.

In the rest of the work, we also refer to an $(M,N)$-Moore-Penrose pseudoinverse which is defined as follows.
Definition 6.5. For a matrix \( A \in \mathbb{K}^{m \times n} \) an \((M, N)\)-Moore-Penrose pseudoinverse of \( A \) is a matrix \( X \in \mathbb{K}^{n \times m} \) satisfying the conditions

i) \( AXA = A \),

ii) \( XAX = X \),

iii) \( AX = (AX)^* \),

iv) \( XA = (XA)^* \).

An \((M, N)\)-Moore-Penrose pseudoinverse of \( A \) will be denoted by \( A^\dagger \).

Theorem 6.6. [39] Let \( M \in \mathbb{K}^{m \times m} \) and \( N \in \mathbb{K}^{n \times n} \) form an orthosymmetric pair. \( W \in \mathbb{K}^{m \times n} \) is a partial \((M, N)\)-isometry if and only if \( W^* M, N = W^\dagger \).

We also need the next definition for the self-adjoint factor of the canonical generalized polar decomposition.

Definition 6.7. A matrix \( S \in \mathbb{K}^{n \times n} \) is called an \( N \)-selfadjoint matrix if \( S^* N = N^{-1} S^* N = S \).

The generalized polar decomposition and the canonical generalized polar decomposition are extensions of the classical polar decomposition and the canonical polar decomposition to the general scalar products, respectively.

Definition 6.8. [39, Canonical generalized polar decomposition] Let \( M \in \mathbb{K}^{m \times m} \) and \( N \in \mathbb{K}^{n \times n} \) form an orthosymmetric pair. A canonical generalized polar decomposition of \( A \in \mathbb{K}^{m \times n} \) is a decomposition \( A = WS \), where \( W \in \mathbb{K}^{m \times n} \) is a partial \((M, N)\)-isometry, \( S \in \mathbb{K}^{n \times n} \) is an \( N \)-selfadjoint matrix whose eigenvalues are contained in the open right half plane, and \( \text{range}(W^* M, N) = \text{range}(S) \).

The existence and the uniqueness of the canonical generalized polar decomposition are stated in the next theorem.

Theorem 6.9. [39, Thm. 3.9] Let \( M \in \mathbb{K}^{m \times m} \) and \( N \in \mathbb{K}^{n \times n} \) form an orthosymmetric pair. Then \( A \in \mathbb{K}^{m \times n} \) has a unique canonical generalized polar decomposition if and only if
\[ A^{*_{M,N}}A \text{ has no negative real eigenvalues;} \]
\[ \text{if zero is an eigenvalue of } A^{*_{M,N}}A \text{ then it is semisimple; and} \]
\[ \text{null}(A^{*_{M,N}}A) = \text{null}(A). \]

The relation between the canonical generalized polar decomposition and the principal square root is given in the following statement.

**Lemma 6.10.** [39, Lemma 3.7] Let \( M \in \mathbb{K}^{m \times m} \) and \( N \in \mathbb{K}^{n \times n} \) form an orthosymmetric pair. If \( A \) has a canonical generalized polar decomposition \( A = WS \) then \( A^{*_{M,N}}A \) has no negative real eigenvalues, \( S \) is a square root of \( A^{*_{M,N}}A \), and

\[ W^{*_{M,N}}WS = S. \]

We are concerned with the changes in the factors \( W \) and \( S \) caused by a perturbation in \( A \). We will analyse how small changes in \( A \) affect the factors by applying perturbation analysis; that is we try to bound the values \( \|\Delta W\| \) and \( \|\Delta S\| \) in terms of \( \|\Delta A\| \). The only perturbation bound for the canonical generalized polar decomposition of real matrices in the literature is given in [81]. Apart from extending the theory to complex matrices we obtain new perturbation bounds by applying new techniques.

### 6.3 Additive perturbation analysis

First we summarise the method given for the real case [81] under additive perturbation analysis and then show the changes for the complex case.

#### 6.3.1 Real case

Let \( A \in \mathbb{R}^{m \times n} \) have a unique CGPD, \( A = WS \). Then a perturbed matrix \( A + \Delta A \in \mathbb{R}^{m \times n} \) has the canonical generalized polar decomposition as

\[ A + \Delta A = (W + \Delta W)(S + \Delta S). \quad (6.1) \]

By using the property in Lemma 6.10 we can write

\[ (A + \Delta A)^{*_{M,N}}(A + \Delta A) = (S + \Delta S)(S + \Delta S). \quad (6.2) \]
Then we get
\[ N^{-1} (A + \Delta A)^T M (A + \Delta A) = (S + \Delta S)(S + \Delta S). \] (6.3)

We rearrange it as
\[ S\Delta S + \Delta SS = N^{-1}A^T M \Delta A + N^{-1} \Delta A^T MA + O(\|\Delta A\|^2). \]

Applying the vec operator gives
\[(I \otimes S + S^T \otimes I) \text{vec}(\Delta S) = (I \otimes N^{-1} A^T M + (A^T M^T \otimes N^{-1}) P) \text{vec}(\Delta A) + O(\|\Delta A\|^2), \]
where \( P \) is the vec permutation matrix. Then we get
\[ \text{vec}(\Delta S) = T_{S^{-1}} T_{AMN} \text{vec}(\Delta A) + O(\|\Delta A\|^2). \] (6.4)

where \( T_S = I \otimes S + S^T \otimes I \) and \( T_{AMN} = I \otimes N^{-1} A^T M + a (A^T M \otimes N^{-1}) P \). By using
the property \( \|\Delta S\|_F = \|\text{vec}(\Delta S)\|_2 \) we can bound \( \|\Delta S\|_F \) by
\[ \|\Delta S\|_F \leq \|T_{S^{-1}} T_{AMN}\|_2 \|\Delta A\|_F + O(\|\Delta A\|_F^2). \] (6.5)

For the change in the polar factor \( W \) it follows simply from (6.1) that
\[ \Delta A = W \Delta S + \Delta WS + O(\|\Delta A\|^2). \]

Since \( S \) is nonsingular we can write
\[ \Delta W = \Delta AS^{-1} - W \Delta SS^{-1} + O(\|\Delta A\|^2). \] (6.6)

Applying the vec operator gives
\[ \text{vec}(\Delta W) = (S^{-T} \otimes I_m) \text{vec}(\Delta A) - (S^{-T} \otimes W) \text{vec}(\Delta S) + O(\|\Delta A\|^2). \] (6.7)

After combining (6.4) and (6.7) we get
\[ \text{vec}(\Delta W) = (S^{-T} \otimes I_m - (S^{-T} \otimes W) T_{S^{-1}} T_{AMN}) \text{vec}(\Delta A) + O(\|\Delta A\|^2). \]

By using the fact that \( \|\Delta W\|_F = \|\text{vec}(\Delta W)\|_2 \) we bound \( \|\Delta W\|_F \).
\[ \|\Delta W\|_F \leq \|(S^{-T} \otimes I_m - (S^{-T} \otimes W) T_{S^{-1}} T_{AMN})\|_2 \|\Delta A\|_F + O(\|\Delta A\|_F^2). \]

So we can write the perturbation bounds for the factors \( W \) and \( S \).
Theorem 6.11. [81] Let $M \in \mathbb{R}^{m \times m}$ and $N \in \mathbb{R}^{n \times n}$ form an orthosymmetric pair. Suppose that $A^* M N$ and $(A + \Delta A)^* M N (A + \Delta A)$ have a unique CGPD. For the perturbation bounds of the factors $S$ and $W$, where $A = WS$ we can write

$$
\frac{\|\Delta S\|_F}{\|S\|_F} \leq \frac{\|S^{-1}\|_2}{\|S\|_F} \left\{ \|N^{-1}A^T M\|_2 + \|MA\|_2 \right\} \|\Delta A\|_F + O(\|\Delta A\|_F^2),
$$

$$
\frac{\|\Delta W\|_F}{\|W\|_F} \leq \frac{\|S^{-1}\|_2}{\|W\|_F} \left\{ 1 + 2\|W\|_2 \right\} \|N^{-1}A^T M\|_2 + \|MA\|_2 \|N^{-1}\|_2 \} \|\Delta A\|_F + O(\|\Delta A\|_F^2).
$$

where $T_S = I \otimes S + S^T \otimes I$ and $T_{AMN} = I \otimes N^{-1}A^T M + \alpha (A^T M \otimes N^{-1}) P$. Since evaluating $T_S$ and $T_{AMN}$ requires many operations the upper bounds are given in the next theorem.

Lemma 6.12. [81] With the same assumptions in Theorem 6.11 the perturbation bounds for $S$ and $W$ can be given as follows:

$$
\frac{\|\Delta S\|_F}{\|S\|_F} \leq 2 \frac{\|S^{-1}\|_2}{\|S\|_F} \left\{ \|N^{-1}A^T M\|_2 + \|MA\|_2 \right\} \|\Delta A\|_F + O(\|\Delta A\|_F^2),
$$

$$
\frac{\|\Delta W\|_F}{\|W\|_F} \leq \frac{\|S^{-1}\|_2}{\|W\|_F} \left\{ 1 + 2\|W\|_2 \right\} \|N^{-1}A^T M\|_2 + \|MA\|_2 \|N^{-1}\|_2 \} \|\Delta A\|_F + O(\|\Delta A\|_F^2).
$$

If we take $M = I_m$ and $N = I_n$ we then get

$$
\frac{\|\Delta S\|_F}{\|S\|_F} \leq 4 \frac{\|S^{-1}\|_2}{\|S\|_F} \|A\|_2 \|\Delta A\|_F + O(\|\Delta A\|_F^2),
$$

$$
\frac{\|\Delta W\|_F}{\|W\|_F} \leq \frac{\|S^{-1}\|_2}{\|W\|_F} \left\{ 1 + 4\|S^{-1}\|_2\|W\|_2 \right\} \|\Delta A\|_F + O(\|\Delta A\|_F^2).
$$

6.3.2 Complex case

In this part we look at the changes for the complex case. Let $A \in \mathbb{C}^{m \times n}$ have a unique CGPD $A = WS$. Then the perturbed matrix satisfies

$$
N^{-1}(A + \Delta A)^* M (A + \Delta A) = (S + \Delta S)(S + \Delta S).
$$

After omitting the second order terms we get

$$
N^{-1}A^* M \Delta A + N^{-1} \Delta A^* M A \approx S \Delta S + \Delta SS.
$$

Applying the vec operator yields

$$
(I \otimes N^{-1}A^* M) \text{vec}(\Delta A) + (A^* M^* \otimes N^{-1}) P \text{vec}(\Delta A) \approx (I \otimes S + S^* \otimes I) \text{vec}(\Delta S),
$$
where $\text{vec}(\Delta A^*) = P\text{vec}(\overline{\Delta A})$ and $\overline{\Delta A}$ denotes the complex conjugate of $\Delta A$. By separating the real and the imaginary parts denoted as

$$\text{vec}(\text{real}(\Delta A)) = \Delta a_r, \quad \text{vec}(\text{real}(\Delta S)) = \Delta s_r,$$
$$\text{vec}(\text{imag}(\Delta A)) = \Delta a_{im}, \quad \text{vec}(\text{imag}(\Delta S)) = \Delta s_{im},$$

we can write

$$\begin{bmatrix}
\Delta s_r \\
\Delta s_{im}
\end{bmatrix} =
\begin{bmatrix}
T_{S}^{-1}T_{A} & 0 \\
0 & T_{S}^{-1}T_{B}
\end{bmatrix}
\begin{bmatrix}
\Delta a_r \\
\Delta a_{im}
\end{bmatrix} + O(\|\Delta A\|^{2}),$$  \quad (6.8)

where

$$T_{S} = I \otimes S + S^{*} \otimes I,$$
$$T_{A} = I \otimes N^{-1}A^{*}M + (A^{*}M^{*} \otimes N^{-1})P,$$
$$T_{B} = I \otimes N^{-1}A^{*}M - (A^{*}M^{*} \otimes N^{-1})P.$$  \quad (6.9a,b,c)

Then we bound the norm of $\Delta S$ as

$$\|\Delta S\|_{F} \leq \|K\|_{2}\|\Delta A\|_{F} + O(\|\Delta A\|^{2}_{F}),$$

where $K = \text{diag}(T_{S}^{-1}T_{A}, T_{S}^{-1}T_{B})$.

For the perturbation analysis of $W$ we apply the $\text{vec}$ operator to equation (6.6) and obtain

$$\text{vec}(\Delta W) = (S^{*} \otimes I_{m})\text{vec}(\Delta A) - (S^{*} \otimes W)\text{vec}(\Delta S) + O(\|\Delta A\|^{2}).$$

By separating the real and the imaginary parts of $W$ denoted as

$$\text{vec}(\text{real}(\Delta W)) = \Delta w_{r}, \quad \text{vec}(\text{imag}(\Delta W)) = \Delta w_{im},$$

we get

$$\Delta w_{r} \approx (S^{*} \otimes I_{m})\Delta a_{r} - (S^{*} \otimes W)\Delta s_{r},$$
$$\Delta w_{im} \approx (S^{*} \otimes I_{m})\Delta a_{im} - (S^{*} \otimes W)\Delta s_{im}.$$
We can rewrite this equation in the matrix form

\[
\begin{bmatrix}
\Delta w_r \\
\Delta w_{im}
\end{bmatrix} = \begin{bmatrix}
S^{-s} \otimes I_m & 0 \\
0 & S^{-s} \otimes I_m
\end{bmatrix} \begin{bmatrix}
\Delta a_r \\
\Delta a_{im}
\end{bmatrix} + \begin{bmatrix}
S^{-s} \otimes W & 0 \\
0 & S^{-s} \otimes W
\end{bmatrix} \begin{bmatrix}
\Delta s_r \\
\Delta s_{im}
\end{bmatrix}
\]

+ \text{O}(\|\Delta A\|^2). \quad (6.10)

After substituting (6.8) into (6.10) we get

\[
\begin{bmatrix}
\Delta w_r \\
\Delta w_{im}
\end{bmatrix} = \begin{bmatrix}
U_1 & 0 \\
0 & U_2
\end{bmatrix} \begin{bmatrix}
\Delta a_r \\
\Delta a_{im}
\end{bmatrix} + \text{O}(\|\Delta A\|^2),
\]

where \(U_1 = S^{-s} \otimes I_m\) and \(U_2 = S^{-s} \otimes W\). This implies

\[
\|\Delta W\|_F \leq \|T_{SI} + T_{SW} K\|_2 \|\Delta A\|_F + \text{O}(\|\Delta A\|^2_F),
\]

with \(T_{SI} = \text{diag}(U_1, U_1)\) and \(T_{SW} = \text{diag}(U_2, U_2)\).

**Lemma 6.13.** Let \(M \in \mathbb{C}^{m \times m}\) and \(N \in \mathbb{C}^{n \times n}\) form an orthosymmetric pair. Suppose that \(A^{*,m,n} A\) and \((A + \Delta A)^{*,m,n}(A + \Delta A)\) have a unique CGPD. The perturbation bounds of the factors \(S\) and \(W\) are, respectively given by

\[
\frac{\|\Delta S\|_F}{\|S\|_F} \leq \frac{\|K\|_2}{\|S\|_F} \|\Delta A\|_F + \text{O}(\|\Delta A\|^2_F),
\]

\[
\frac{\|\Delta W\|_F}{\|W\|_F} \leq \frac{\|T_{SI} + T_{SW} K\|_2}{\|W\|_F} \|\Delta A\|_F + \text{O}(\|\Delta A\|^2_F),
\]

where \(K = \text{diag}(T_{S}^{-1} T_{A}, T_{S}^{-1} T_{B})\) with \(T_{S}, T_{A}\) and \(T_{B}\) given in (6.9), \(T_{SI} = \text{diag}(U_1, U_1)\) and \(T_{SW} = \text{diag}(U_2, U_2)\) with \(U_1 = S^{-s} \otimes I_m\) and \(U_2 = S^{-s} \otimes W\).

Since building \(T_{SI}\) and \(T_{SW}\) is expensive we can bound the perturbation bounds.

**Lemma 6.14.** With the same assumptions in Lemma 6.13, upper bounds of the perturbation bounds for the canonical generalized polar factors \(S\) and \(W\) are given as

\[
\frac{\|\Delta S\|_F}{\|S\|_F} \leq \frac{\|T_{S}^{-1}\|_2}{\|S\|_F} \max\{\|T_{A}\|_2, \|T_{B}\|_2\} \|\Delta A\|_F + \text{O}(\|\Delta A\|^2_F),
\]

\[
\frac{\|\Delta W\|_F}{\|W\|_F} \leq \frac{\|S^{-1}\|_2}{\|W\|_F} (1 + \|W\|_2 \|K\|_2) \|\Delta A\|_F + \text{O}(\|\Delta A\|^2_F),
\]

where \(K = \text{diag}(T_{S}^{-1} T_{A}, T_{S}^{-1} T_{B})\) with \(T_{S}^{-1}, T_{A}\) and \(T_{B}\) given in (6.9).
Proof. For the perturbation bound of the factor $S$ we write

$$\|K\|_2 = \max\{\|T_S^{-1}T_A\|_2, \|T_S^{-1}T_B\|_2\}.$$ 

For the perturbation bound of the factor $W$ we get

$$\|T_SI + T_SW K\|_2 \leq \|T_SI\|_2 + \|T_SW K\|_2.$$ 

### 6.4 Multiplicative perturbation analysis

There are many applications of multiplicative perturbation analysis of certain problems in numerical linear algebra such as $QR$ and polar factorizations [13, 53] and singular value decomposition [54, 55, 56]. In the literature there is only additive perturbation analysis of the canonical generalized polar decomposition. Here we analyse the cases where $A \in \mathbb{C}^{m \times n}$ is subject to right and left multiplicative perturbation. That is, $A$ is perturbed to $\tilde{A} = D_L A$ or $\tilde{A} = AD_R$ with $D_L$ and $D_R$ near the identities. Multiplicative perturbations arise from matrix scaling, which might occur in iterative methods for computing the canonical generalized polar decomposition.

#### 6.4.1 Left multiplicative perturbation

Let the perturbed matrix $\tilde{A} \equiv D_L A$ have the unique canonical generalized polar decomposition, where $D_L = I_m + E$ with small $E$.

$$\tilde{A} = (W + \Delta W)(S + \Delta S).$$  \hfill (6.11)

By using the property $A^{*M,N}A = S^2$ we can write

$$(S + \Delta S)(S + \Delta S) = N^{-1}((I_m + E)A)^*M(I_m + E)A.$$ 

We rearrange the equation as

$$S\Delta S + \Delta SS = N^{-1}A^*MEA + N^{-1}A^*E^*MA + O(\|\Delta A\|^2).$$
Since the matrices on the right hand side and $S$ have a finite norm \[7\] the solution can be given by

$$
\Delta S = \int_0^\infty e^{-tS} (N^{-1}A^*MEA + N^{-1}A^*E^*MA)e^{-tS} dt.
$$

Then the bound for $\|\Delta S\|$ yields

$$
\|\Delta S\|_F \leq \int_0^\infty \|e^{-tS} (N^{-1}A^*MEA + N^{-1}A^*E^*MA)e^{-tS}\|_F dt + O(\|EA\|^2).
$$

By using the inequality which is a corollary of the theorem in \[40, \text{Thm. 7.4.8.4}\]

$$
\|ABC\|_F \leq \|A\|_2 \|B\|_F \|C\|_2
$$

we obtain

$$
\|\Delta S\|_F \leq 2\gamma(s)\|N^{-1}\|_2\|A\|_2\|M\|_2\|\Delta A\|_F + O(\|\Delta A\|_F^2). \tag{6.13}
$$

where $\gamma(S) = \int_0^\infty \|e^{-tS}\|_2^2 dt$ and $\Delta A = EA$.

For the perturbation analysis of $W$ we rewrite \eqref{eq:6.11} as

$$(I_m + E)A = WS + W\Delta S + \Delta WS + O(\|EA\|^2).$$

Since $S$ is nonsingular we get

$$
\Delta W = EAS^{-1} - W\Delta SS^{-1} + O(\|EA\|^2).
$$

We can bound $\|\Delta W\|$ by using \eqref{eq:6.12}

$$
\|\Delta W\|_F \leq \|S^{-1}\|_2\|EA\|_F + \|W\|_2\|\Delta S\|_F + O(\|EA\|^2),
$$

where $\Delta A = EA$. Substituting the bound of $\|\Delta S\|_F$ gives

$$
\|\Delta W\|_F \leq \|S^{-1}\|_2 \left(1 + 2\gamma(s)\|W\|N^{-1}\|A\|_2\|M\|_2\right)\|\Delta A\|_F + O(\|\Delta A\|_F^2). \tag{6.14}
$$

To bound $\gamma(S)$ we will follow the same framework in \[6, \text{X. 3}\]. Assume that $S$ is diagonalizable $S = XDX^{-1}$, where $D = \text{diag}(\lambda_i)$ and $\lambda_i$ are the eigenvalues of $S$. Then we have

$$
\|e^{-tS}\|_2 \leq \kappa_2(X)\|e^{-tD}\|_2 = \kappa_2(X)e^{-t\alpha},
$$

where $\kappa_2(X) = \|X\|_2\|X^{-1}\|_2$ and $\alpha = |\min(\text{re}(\lambda_i))|$. The bound of $\gamma(S)$ is given as

$$
\gamma(S) \leq \int_0^\infty \kappa_2^2(X)e^{-2t\alpha} dt = \frac{1}{2\alpha} \kappa_2^2(X). \tag{6.15}
$$

As seen the bound depends on the real parts of the eigenvalues and the condition number of $X$. 
6.4.2 Right multiplicative perturbation

Let the perturbed matrix $\tilde{A} \equiv AD_R$, where $D_R = I_n + K$ with small $K$ have the canonical generalized polar decomposition.

$$\tilde{A} = (W + \Delta W)(S + \Delta S).$$  \hfill (6.16)

By using the property $A^*M, N A = S^2$ we can write

$$S^2 + S \Delta S + \Delta SS = N^{-1}(A(I_n + K))^* MA(I_n + K) + O(\|AK\|^2),$$

which yields

$$S \Delta S + \Delta SS \approx N^{-1}A^* MAK + N^{-1}K^* A^* MA + O(\|AK\|^2).$$

Since the matrices on the right hand side and $S$ have a finite norm we can write the solution [7] as

$$\Delta S \approx \int_0^\infty e^{-tS} (N^{-1}A^* MA + N^{-1}K^* A^* MA) e^{-tS} dt,$$

where $\Delta A = AK$. By taking the norm of both sides and using (6.12) we get

$$\|\Delta S\|_F \leq 2\gamma(S)\|N^{-1}\|_2\|A\|_2\|M\|_2\|\Delta A\|_F + O(\|\Delta A\|^2_F),$$  \hfill (6.17)

where $\gamma(S) = \int_0^\infty \|e^{-tS}\|^2 dt$.

For the perturbation analysis of $W$ we write

$$A(I_n + K) = WS + W \Delta S + \Delta WS + O(\|AK\|^2).$$

Since $S$ is nonsingular the equation yields

$$\Delta W = AKS^{-1} - W \Delta SS^{-1} + O(\|\Delta A\|^2).$$

Then we bound the norm of $\Delta W$ as

$$\|\Delta W\|_F \leq \|S^{-1}\|_2\|\Delta A\|_F + \|WS^{-1}\|_2\|\Delta S\|_F + O(\|\Delta A\|^2_F).$$

Substituting the bound of $\|\Delta S\|_F$ gives

$$\|\Delta W\|_F \leq \|S^{-1}\|_2 (1 + 2\gamma(S)\|W\|_2\|N^{-1}\|_2\|A\|_2\|M\|_2) \|\Delta A\|_F + O(\|\Delta A\|^2_F).$$  \hfill (6.18)

It is shown that $\gamma(s)$ is bounded by (6.15). As seen from the pairs (6.13), (6.17) and (6.14), (6.18) we obtain the same coefficients in the first order perturbation bounds for the left and right perturbation analysis.
Lemma 6.15. Let the nonsingular matrices $M \in \mathbb{C}^{m \times m}$ and $N \in \mathbb{C}^{n \times n}$ form an orthosymmetric pair. Suppose that $A \in \mathbb{C}^{m \times n}$ and $A + \Delta A$ have a unique CGPD. Assume that $S$ is diagonalizable $S = XDX^{-1}$, where $D = \text{diag} (\lambda_i)$ and $\alpha = |\min\text{re}(\lambda_i)|$.

When $A$ is subject to the left multiplicative perturbation $A + \Delta A = D_L A$, where $D_L = I_m + E$ with small $E$ or the right multiplicative perturbation $A + \Delta A = AD_R$, where $D_R = I_n + K$ with small $K$ the perturbation bounds of the factors $S$ and $W$ are, respectively given by

$$\|\Delta S\|_F \leq \frac{2\gamma(s)}{\|S\|_F} \|N^{-1}\|_2 \|A\|_2 \|M\|_2 \|\Delta A\|_F + O(\|\Delta A\|_F^2), \quad \Delta A = EA$$

$$\|\Delta W\|_F \leq \frac{\|S^{-1}\|_2^2}{\|W\|_F} (1 + 2\gamma(s) \|A\|_2 \|N^{-1}\|_2 \|A\|_2 \|M\|_2) \|\Delta A\|_F + O(\|\Delta A\|_F^2), \quad \Delta A = AK.$$

where $\gamma(S) \leq \frac{1}{2\alpha} \kappa_2^2(X)$.

For $M = I_m$ and $N = I_n$ the upper bounds become

$$\|\Delta S\|_F \leq \frac{2\gamma(s)}{\|S\|_F} \|A\|_2 \|\Delta A\|_F + O(\|\Delta A\|_F^2),$$

$$\|\Delta W\|_F \leq \frac{\|S^{-1}\|_2^2}{\|W\|_F} (1 + 2\gamma(s) \|A\|_2) \|\Delta A\|_F + O(\|\Delta A\|_F^2).$$

### 6.5 Using standard techniques

We will use the technique given for the polar decomposition in [31, Thm. 2.5] to decide the bounds for the perturbations of the CGPD factors. Let $E \in \mathbb{C}^{m \times n}$ and $t$ be small enough so that $A(t) = A + tE$ has the canonical generalized polar decomposition as stated in Theorem 6.9

$$A(t) = W(t)S(t), \quad 0 \leq t \leq \epsilon,$$  \hspace{1cm} (6.19)

where $W(t)$ is a partial $(M, N)$-isometry and $S(t)$ is an $N$-selfadjoint matrix. By using the property $A^{*_{M, N}} A = S^2$ we can write

$$A^{*_{M, N}}(t) A(t) = S^2(t).$$  \hspace{1cm} (6.20)

We assume that $W(t)$ and $S(t)$ are twice continuously differentiable. By differentiating equation (6.20) with respect to $t$ we get

$$A^{*_{M, N}}(t) A(t) + A^{*_{M, N}}(t) \dot{A}(t) = \dot{S}(t) S(t) + S(t) \dot{S}(t),$$  \hspace{1cm} (6.21)
where \( \dot{A}(t) \) denotes the derivative of \( A(t) \) with respect to \( t \). Taking \( t = 0 \) and \( A(0) = WS \) gives
\[
E^{*,N}WS + S^{*,N}W^{*,M,N}E = \dot{S}(0)S + S\dot{S}(0),
\]
where \( \dot{A}^{*,M,N}(0) = E^{*,M,N} \), \( \dot{A}(0) = E \) and \( S(0) = S \). Assume that \( S \) is diagonalizable \( S = XDX^{-1} \), where \( D = \text{diag}(\lambda_i) \). So we have
\[
E^{*,N}WXDX^{-1} + XDX^{-1}W^{*,M,N}E = \dot{S}(0)XDX^{-1} + XDX^{-1}\dot{S}(0).
\]
We can rearrange this equation as
\[
YD + DY = PD + DR,
\]
where \( Y = X^{-1}\dot{S}(0)X \), \( P = X^{-1}E^{*,M,N}WX \) and \( R = X^{-1}W^{*,M,N}EX \). The solution to this equation is given as
\[
y_{ij} = \frac{p_{ij}\lambda_j + \lambda_j r_{ij}}{\lambda_i + \lambda_j}, \quad 1 \leq i, j \leq n.
\]
By using the Cauchy-Schwarz inequality [79] and \( \lambda_i^2 + \lambda_j^2 \leq (\lambda_j + \lambda_i)^2 \) we can write
\[
|y_{ij}|^2 \leq |p_{ij}|^2 + |r_{ij}|^2.
\]
This implies
\[
\|Y\|_F^2 \leq \|P\|_F^2 + \|R\|_F^2.
\]
By substituting \( P \) and \( R \) in the bound and using (6.12) we get
\[
\|P\|_F^2 + \|R\|_F^2 \leq \|X^{-1}E^{*,M,N}WX\|_F^2 + \|X^{-1}W^{*,M,N}EX\|_F^2 \leq 2\kappa_2^2(X)\|W\|_2^2\|N^{-1}\|_2^2\|M\|_2^2\|E\|_F^2,
\]
where \( \kappa_2(X) = \|X\|_2\|X^{-1}\|_2 \). Then we get
\[
\|Y\|_F \leq \sqrt{2}\kappa_2(X)\|W\|_2\|N^{-1}\|_2\|M\|_2\|E\|_F,
\]
The bound for \( \|\dot{S}(0)\|_F \) is given as
\[
\|\dot{S}(0)\|_F \leq \|XYX^{-1}\|_F \leq \sqrt{2}\kappa_2^2(X)\|W\|_2\|N^{-1}\|_2\|M\|_2\|E\|_F. \quad (6.22)
\]
The Taylor series gives
\[
S + \Delta S \equiv S(t) = S(0) + \dot{S}(0)(A(t) - A) + O(t^2).
\]
Then we get
\[ \| \Delta S \|_F = \| S(t) - S(0) \|_F \leq \| \dot{S}(0) \|_F \| A(t) - A \|_F + O(t^2), \]
which yields
\[ \| S(t) - S \|_F \leq \sqrt{2} \kappa_2^2(X) \| W \|_2 \| N^{-1} \|_2 \| M \|_2 \| A(t) - A \|_F + O(t^2). \]
Assuming \( S(t) \) is nonsingular for all \( t \) we rewrite (6.19) as
\[ W(t) = A(t)S^{-1}(t). \quad (6.23) \]
Differentiating with respect to \( t \) yields
\[ \dot{W}(t) = \dot{A}(t)S^{-1}(t) - A(t)S^{-1}(t)\dot{S}(S^{-1}(t)). \]
Evaluating this at \( t = 0 \) gives
\[ \dot{W}(0) = ES^{-1} - AS^{-1}\dot{S}(0)S^{-1}, \]
where \( E = \dot{A}(0) \) and \( S^{-1}(0) = S^{-1} \). Then by using (6.12) we show that
\[ \| \dot{W}(0) \|_F \leq \| E \|_F \| S^{-1} \|_2 + \kappa_2(X) \| W \|_2 \| S^{-1} \|_2 \| \dot{S}(0) \|_F. \]
By substituting (6.22) we get
\[ \| \dot{W}(0) \|_F \leq \| E \|_F \| S^{-1} \|_2 + \sqrt{2} \kappa_2^2(X) \| W \|_2^2 \| S^{-1} \|_2 \| N^{-1} \|_2 \| M \|_2 \| E \|_F. \quad (6.24) \]
Then in a similar way Taylor expansion gives
\[ \| \Delta W \|_F = \| W(t) - W(0) \|_F \leq \| \dot{W}(0) \|_F \| A(t) - A \|_F + O(t^2) \]
which yields
\[ \| W(t) - W \|_F \leq \| S^{-1} \|_2 \left( 1 + \sqrt{2} \kappa_2^2(X) \| W \|_2^2 \| N^{-1} \|_2 \| M \|_2 \right) \| A(t) - A \|_F + O(t^2). \]

Lemma 6.16. Let \( A \in \mathbb{C}^{m \times n} \) have a unique canonical generalized polar decomposition.
Let \( t \) be small enough so that \( A(t) = A + tE \) has a unique canonical generalized polar decomposition
\[ A(t) = W(t)S(t), \quad 0 \leq t \leq \epsilon. \quad (6.25) \]
Then perturbation bounds are given by
\[
\frac{\|S(t) - S\|_F}{\|S\|_F} \leq \sqrt{2} \kappa_2^2(X) \frac{\|W\|_2 \|N^{-1}\|_2 \|M\|_2 \|A(t) - A\|_F + O(t^2)}{\|S\|_F},
\]
\[
\frac{\|W(t) - W\|_F}{\|W\|_F} \leq \frac{\|S^{-1}\|_2}{\|W\|_F} \left(1 + \sqrt{2} \kappa_2^2(X) \frac{\|W\|_2^2 \|N^{-1}\|_2 \|M\|_2}{\|W\|_F} \right) \|A(t) - A\|_F + O(t^2).
\]

To compare the bounds given in Lemma 6.16 with the previous results we take \(\epsilon = 1\). If we take \(M = I_m\) and \(N = I_n\) for the standard canonical polar decomposition then the upper bounds reduce to
\[
\frac{\|S(t) - S\|_F}{\|S\|_F} \leq \sqrt{2} \kappa_2^2(X) \|W\|_2 \|E\|_F + O(t^2),
\]
\[
\frac{\|W(t) - W\|_F}{\|W\|_F} \leq \frac{\|S^{-1}\|_2}{\|W\|_F} \left(1 + \sqrt{2} \kappa_2^2(X) \right) \|E\|_F + O(t^2).
\]

### 6.6 Numerical experiments

We present some numerical experiments to give further insight into the perturbation analysis. We will compare the coefficients of the first order perturbation bounds given in Lemma 6.11, 6.15, and 6.16. The experiments are carried out in MATLAB R2014b, in which the unit round off is \(u \approx 1.1 \times 10^{-16}\). The factors of the canonical generalized polar decomposition are computed by using the properties
\[
S = (A^{*M,N}A)^{1/2}, \quad W = AS^{-1}.
\]

- Each experiment is provided for 1000 matrices that are built by using either Jagger’s MATLAB Toolbox [41] or \texttt{rand} from MATLAB.

- In the figures, we present the coefficients of the first order perturbation bounds given in Lemma 6.11, 6.15, and 6.16.

- The nonsingular matrices \(M \in \mathbb{K}^{m \times m}\) and \(N \in \mathbb{R}^{n \times n}\) are built such that they form an orthosymmetric pair as given in Definition 6.3.

- In the legends of figures \texttt{add}, \texttt{stand} and \texttt{multip} stand for the additive, standard and multiplicative perturbation analysis, respectively.
• **Data** gives the coefficients of the first order perturbation bounds of the factors in a probability plot showing the distribution of **Data** obtained by using `probplot` from MATLAB.

**Experiment 1**

In the first experiment we compute the coefficient of the first order perturbation bounds for rectangular matrices with the orthosymmetric pair $M \in \mathbb{R}^{12 \times 12}$ and $N \in \mathbb{R}^{8 \times 8}$, that is $M^T = M$ and $N^T = N$. We build $M$ and $N$ by using `rand` function in MATLAB. The upper bounds for 1000 random $A \in \mathbb{R}^{12 \times 8}$ test matrices are provided in the distribution figures.

![Figure 6.1:](image)

(a) Upper bounds for $S$  
(b) Upper bounds for $W$

Figure 6.1: The bounds are given for random $A \in \mathbb{R}^{12 \times 8}$ matrices with $\kappa_2(A) = 10^6$.

Figure 6.1 shows the probabilities of getting the perturbation bounds after analysing sample sets of 1000 matrices. **Data** represents the coefficients of the first order perturbation bounds given in Lemma 6.11, 6.15, and 6.16. As seen from the figure standard technique gives sharper bounds especially for the factor $S$.

**Experiment 2**

In the second experiment we repeat the first experiment for nonsingular matrices $A \in \mathbb{R}^{10 \times 8}$ with $\kappa_2(A) = 10^2$, $M = I_{10}$ and $N = I_8$, where $I_n$ is an identity matrix.
When we look at Figure 6.2 the coefficient of the perturbation bounds for $W$ and $S$ given in Lemma 6.16 are sharper comparing to the other bounds. For the perturbation bound $W$, the additive and multiplicative perturbation bounds give the same results.

Experiment 3

Here the upper bounds are compared for a random nonsingular matrix $A \in \mathbb{R}^{12 \times 8}$ with the orthosymmetric pair $M = R_{12} \in \mathbb{R}^{12 \times 12}$ and $N = R_8 \in \mathbb{R}^{8 \times 8}$, where $R_n$ is given as

$$R_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$
Figure 6.3: The bounds are given for $A \in \mathbb{R}^{12 \times 8}$ with $\kappa_2(A) = 10^6$, $M = R_{12} \in \mathbb{R}^{12 \times 12}$ and $N = R_8 \in \mathbb{R}^{8 \times 8}$.

Figure 6.3 gives the distribution of the coefficients of the first order perturbation bounds of $W$ and $S$ for 1000 experiments. Depending on the choice of $A$ the upper bound represented by red diamonds gives sharper bounds in many cases, especially for the $S$ factor.

Experiment 4

In the final experiments $M \in \mathbb{R}^{12 \times 12}$ and $N \in \mathbb{R}^{8 \times 8}$ are chosen as

$$M = \Sigma_{8,4} = \begin{bmatrix} I_8 & 0 \\ 0 & -I_4 \end{bmatrix}, \quad N = \Sigma_{6,2} = \begin{bmatrix} I_6 & 0 \\ 0 & -I_2 \end{bmatrix}$$

with nonsingular $A \in \mathbb{R}^{12 \times 8}$. 
Figure 6.4: The bounds are given for $A \in \mathbb{R}^{12 \times 8}$ with $\kappa_2(A) = 10^5$, $M = \Sigma_{8,4}$ and $N = \Sigma_{6,2}$.

We observe in Figure 6.4 that in most of the cases standard technique gives sharper bounds for the factor $S$ and $W$.

The results show that the sharpness of the upper bounds depends on not only the choice of $A$ but also the choice of the orthosymmetric pair $M$ and $N$. The results also enable us to identify the best method to measure the sensitivity of factors to the small perturbations. Especially for the ill-conditioned matrices the standard technique gives better bounds comparing to other methods.
Conclusions

Functions of structured matrices have many applications hence it is important to exploit the structure to gain more meaningful results for the physical problems behind them. That structure exploitation can help us to develop more accurate numerical algorithms in the computation of matrix functions. This thesis has discussed functions of structured matrices from both theoretical and numerical aspects. This final chapter summarises the research findings and gives concluding remarks.

Chapter 2 focuses on computing the logarithm of symplectic matrices. The study has gone some way towards enhancing our understanding of the structure of the logarithm of symplectic matrices. By using the Schur-like form of the symplectic matrices we have improved the current algorithms using the inverse scaling and squaring method with the forward error analysis. It is not always possible to obtain the structured Schur form but the numerical experiments show that the algorithm using the Schur-like form when possible is structure preserving in the sense that it produces Hamiltonian matrix. We also observe the best accurate results with the usage of Schur-like form of a symplectic matrix instead of using the Schur decomposition or the full form of a matrix. In the future more condensed structured Schur forms obtained by non-orthogonal but symplectic similarities can be used in the inverse scaling and squaring method but there is no guarantee for the numerical robustness.

It is crucial to answer “How far is the function of a nearly structured matrix to the function of the nearest structured matrix?” Here “far” refers to the measurement of
distance in a matrix norm. The purpose of Chapter 3 is to determine the distance of the function of the nearly structured matrix to the function of the nearest structured matrix. One of the advantages of computing the function of the nearest structured matrix instead of nearly structured one is to explore the structure during the computation. It also lets us use the structure preserving algorithms which are faster and more accurate with physically meaningful solutions. The analysis is given for the matrix logarithm, square root, exponential and trigonometric functions. The distance measure is provided for nearly unitary, nearly Hermitian and nearly Hermitian positive semidefinite matrices. The findings will serve as a basis for future studies in the functions of nearly structured matrices by pointing out the cases where the forward error bound approaches zero or infinity.

In the computation of matrix functions the matrix is subject to perturbations that can be rounding errors, truncating errors or uncertainties. We measure the sensitivity of matrix functions to small perturbations by condition numbers. In the general theory of condition numbers there is no restriction on the structure of a starting matrix or perturbed matrix. However, it is important to preserve the structure to respect the physical problem behind the computation. In Chapter 4 we have derived the structured condition number for the matrix functions defined between smooth square matrix manifolds by using the differential of function between tangent spaces to the related smooth matrix manifolds. We have compared the structured and unstructured condition numbers for specific matrix functions. The findings of this study make several contributions to the theory of functions of structured matrices. One of the more significant findings to emerge from the numerical experiments is that a structured condition number can be much smaller than an unstructured one, which emphasizes the importance of using structure preserving algorithms in the computation of matrix function. It is also observed that the gap between the structured and unstructured condition number is getting bigger for ill-conditioned matrices. We have considered decreasing the cost of the algorithm computing the structured condition number by using the power method. We have shown the estimation of the lower and the upper bounds on the structured condition number with even less computational cost than the cost of estimating structured condition number. Another finding is the structured
relative error bound obtained via structured condition number. The structured and unstructured relative error bounds are presented for the matrix logarithm and matrix square root.

Perturbation theory also plays an important role in matrix factorizations. We can have an understanding of the structured perturbations in the structured matrix factorizations by computing the structured condition number. In Chapter 5 we have extended the theory of structured condition number of matrix functions to the matrix factorizations. By following a similar framework to that of Chapter 4 we have derived the structured condition number of $A_1$ and $A_2$ for a given structured matrix factorization $A = A_1A_2$. To reduce the cost we have also derived algorithms using the power method to estimate the structured condition number. The theory has been applied to the structured polar decomposition, structured sign factorization and the generalised polar decomposition.

We can state the canonical generalized polar decomposition as one of the essential matrix factorizations since the concept of canonical polar decomposition can be generalized to general orthosymmetric scalar products and rectangular matrices. The canonical generalized polar decomposition has also applications in the solution of Procrustes problems. The work presented in Chapter 6 gives an unstructured perturbation analysis of the canonical generalized polar decomposition by using different methods: additive, multiplicative and analytic. We have described the upper bounds for the changes in the matrix factors and discussed the sharpness of the bounds obtained via these methods. It is shown that the new method used for the perturbation analysis gives sharper bounds in most of the situations.

In future investigations, it might be possible to use other structured reduced forms of the structured matrices in the inverse scaling and squaring method. Stable algorithms for obtaining the Schur-like form of symplectic matrices can improve the computation of the logarithm of symplectic matrices. The research of the structured condition number can serve as a basis for future studies in the structure preserving algorithms. The analysis can also be applied to other matrix functions defined between smooth manifolds, and matrix factorizations such as sparse matrix factorization.
Bibliography


