A Galerkin Method for Large-scale Autonomous Differential Riccati Equations based on the Loewner Partial Order

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von **M. Sc. Maximilian Behr**

geb. am **21.03.1990** in Schwedt/Oder

genehmigt durch die Fakultät für Mathematik

der Otto-von-Guericke-Universität Magdeburg

Gutachter: **Jun.-Prof. Dr. Jan Heiland**

Dr. Tony Stillfjord

eingereicht am: **27.10.2021**

Verteidigung am: **17.12.2021**

PUBLICATIONS

Publications related to this thesis:

[14] M. Behr, P. Benner, and J. Heiland, *On an Invariance Principle for the Solution Space of the Differential Riccati Equation*, Proc. Appl. Math. Mech., 18 (2018), p. e201800031, https://doi.org/10.1002/pamm.201800031.

[15] M. Behr, P. Benner, and J. Heiland, *Solution Formulas for Differential Sylvester and Lyapunov Equations*, Calcolo, 56 (2019), p. 51, https://doi.org/10.1007/s10092-019-0348-x.

[16] M. Behr, P. Benner, and J. Heiland, *Galerkin trial spaces and Davison–Maki methods for the numerical solution of differential Riccati equations*, Appl. Math. Comp., 410 (2021), p. 126401, https://doi.org/10.1016/j.amc.2021.126401.

ABSTRACT

This thesis deals with the numerical solution approximation of large-scale (autonomous) differential Riccati equations. The first part of the thesis focuses on the differential Lyapunov equation. We recapitulate well-known explicit solution formulas and use them to motivate a Galerkin approach for the numerical solution approximation. For the trial space of the Galerkin method, we propose to use a system of orthonormal eigenvectors of the solution of the algebraic Lyapunov equation. We motivate our choice by estimating the projection error on the trial space using the Loewner partial order. Then, the Galerkin condition yields a system of a smaller order, which can be treated numerically more efficiently. Finally, we compare the proposed Galerkin approach with the BDF-ADI method in terms of accuracy and computational time in several numerical experiments.

In the second part, we extend the proposed Galerkin method to the differential Riccati equation. First, we review the essential analytical properties of the solution of the differential Riccati equation. Then, we estimate the projection error of the solution of the differential Riccati equation using the Loewner partial order and, therefore, motivating a Galerkin approach based on a system of orthonormal eigenvectors of the solution of the algebraic Riccati equation. The Galerkin condition yields a small-scale differential Riccati equation. We recapitulate the Davison–Maki and the modified Davison–Maki method for the numerical solution of the small-scale differential Riccati equation. We compare the proposed Galerkin approach with different splitting methods in terms of accuracy and computing time in several numerical experiments. Furthermore, we discuss a possible extension of the Galerkin method to the case of non-zero initial conditions.

ZUSAMMENFASSUNG

Diese Arbeit befasst sich mit der numerischen Lösungsapproximation von großskaligen (autonomen) Riccati–Differentialgleichungen. Der erste Teil der Arbeit konzentriert sich auf die Lyapunov–Differentialgleichung. Wir rekapitulieren bekannte explizite Lösungsformeln und leiten anhand dessen einen Galerkinansatz zur numerischen Lösungsapproximation her. Für den Ansatzraum des Galerkinverfahrens schlagen wir vor ein System orthonormalen Eigenvektoren der Lösung der algebraisch Lyapunov–Gleichung zu verwenden. Wir motivieren unsere Wahl durch die Abschätzung des Projektionsfehlers auf den Ansatzraum durch Nutzung der Loewner–Halbordnung. Die Galerkinbedingung liefert dann ein System kleinerer Ordnung, welches sich numerisch effizienter behandeln lässt. Schließlich vergleichen wir das vorgeschlagene Galerkinverfahren mit dem BDF-ADI Verfahren hinsichtlich der Genauigkeit und Rechenzeit in mehreren numerischen Experimenten.

Im zweiten Teil erweitern wir das vorgeschlagene Galerkinverfahren auf die Riccati–Differentialgleichung. Zunächst wiederholen wir wichtige analytische Eigenschaften der Lösung der Riccati–Differentialgleichung. Wir geben eine Abschätzung des Projektionsfehlers der Lösung der Riccati–Differentialgleichung unter Ausnutzung der Loewner–Halbordnung an und motivieren dadurch einen Galerkinansatz basierend auf einem System von orthonormalen Eigenvektoren der Lösung der algebraischen Riccati–Gleichung zur numerischen Approximation zu verwenden. Die Galerkinbedingung führt dann auf eine kleinskalige Riccati–Differentialgleichung. Zur numerischen Lösung der kleinskaligen Riccati–Differentialgleichung rekapitulieren wir das Davison–Maki und das modifizierte Davison–Maki Verfahren. Wir vergleichen das vorgeschlagene Galerkinverfahren mit verschiedenen Splitting Verfahren hinsichtlich der Genauigkeit und Rechenzeit in mehreren numerischen Experimenten. Des Weiteren diskutieren wir eine mögliche Erweiterung des Galerkinverfahrens auf den Fall von Nichtnull–Anfangsbedingungen.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF ACRONYMS AND ABBREVIATIONS

Algorithms

ADI	alternating directions implicit
RADI	a low-rank ADI-type algorithm for large-scale algebraic Riccati equations
BDF	backward differentiation formula

Hardware and Software

MATLAB®	software by The MathWorks Inc.
Xeon®	processor series by Intel®

Linear Algebra

hpsd	Hermitian positive semidefinite
spsd	symmetric positive semidefinite
SVD	singular value decomposition

Matrix Equations

ALE	algebraic Lyapunov equation
DLE	differential Lyapunov equation
ASE	algebraic Sylvester equation
DSE	differential Sylvester equation
ABE	algebraic Bernoulli equation
ARE	algebraic Riccati equation
DRE	differential Riccati equation
SDRE	symmetric differential Riccati equation
NDRE	nonsymmetric differential Riccati equation

Others

IVP	initial value problem
ODE	ordinary differential equation

LIST OF SYMBOLS

Numbers

$\delta_{i,j}$	the Kronecker delta, 1 if $i = j$ otherwise 0
$\mathbb{N}$, $\mathbb{N}_0$	the natural numbers (including 0)
$\mathbb{R}$, $\mathbb{C}$	the fields of real and complex numbers
$\mathbb{C}_-$, $\mathbb{C}_+$	the open right/open left complex half-plane
$\mathbb{R}^n$, $\mathbb{C}^n$	the linear space of real/complex n-tuples
$\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$	the real/complex $m \times n$ matrices
$\lvert\xi\rvert$	the absolute value of a real or complex scalar
$\boldsymbol{\imath}$	the imaginary unit ($\boldsymbol{\imath}^2 = -1$)
$\varepsilon_{\texttt{mach}}$	$:= 2^{-52} \approx 2.2 \cdot 10^{-16}$, the machine epsilon

Sets and Distances

∂D	the boundary of a set D
$\operatorname{dist}(x, D)$	$:= \inf\{d(x, y) \mid y \in D\}$, the distance from a point $x \in X$ to a set $D \subseteq X$ in a metric space (X, d)
$X_1 \pm X_2$	$:= \{x_1 \pm x_2 \mid x_1 \in X_1, x_2 \in X_2\}$, the sum and difference of the sets X_1 and X_2

Special Functions

e^x, $\exp(x)$	the exponential function
$\ln(x)$	the natural logarithm
$\operatorname{sgn}(x)$	the sign of real number x, $\frac{x}{\lvert x\rvert}$ if $x \neq 0$ otherwise 0
id_X	the identity function on the set X

Matrices and Vectors

$A_{i,j}$	the (i,j)-th entry of A
e_i	the i-th standard unit vector
$0, 0_n$	the zero matrix (of size $n \times n$)
I	the identity matrix of generic size
I_n	the identity matrix of size $n \times n$
$\Re(A)$, $\Im(A)$	the real and imaginary part of a complex matrix $A = \Re(A) + \boldsymbol{\imath}\Im(A)$
$\overline{A}$	$:= \Re(A) - \boldsymbol{\imath}\Im(A)$, the complex conjugate of A
A^{T}	the transpose of A
A^{H}	$:= (\overline{A})^{\mathsf{T}}$, the complex conjugate transpose
A^{-1}	the inverse of a nonsingular matrix A
$A^{-\mathsf{T}}$, $A^{-\mathsf{H}}$	the inverse of A^{T}, A^{H}
$A^{1/2}$	the square root of a real symmetric positive semidefinite matrix A (Definition/Theorem 2.7)
$\ker(A)$	the kernel or null space of A
$\operatorname{im}(A)$	the image, column space, or range of A
$\operatorname{rank}(A)$	the rank of A
$\det(A)$	the determinant of A
$A \preceq B$	the Loewner partial order (Definition 2.3)
$A \odot B$	the Hadamard product of A and B (Definition 2.34)
$A \otimes B$	the Kronecker product of A and B (Definition 2.35)
$\operatorname{vec}(A)$	the vectorization operator of A (Definition 2.36)
χ_A	the characteristic polynomial of a square matrix A

Special Matrices

$\operatorname{diag}(a_1, \ldots, a_n)$	the diagonal matrix with diagonal entries $a_1, \ldots, a_n$
$\operatorname{tridiag}(\alpha, \beta, \gamma)$	the tridiagonal matrix with α, β, γ on the subdiagonal, diagonal, and superdiagonal, respectively
$\mathfrak{J}$	see Definition 2.40

$\mathcal{K}(A, B)$	the Krylov matrix generated by $A \in \mathbb{C}^{m \times m}$ and $B \in \mathbb{C}^{m \times n}$ (Definition 2.45)

Special Linear Subspaces

$\mathcal{G}(X)$	the graph subspace of $X \in \mathbb{C}^{n \times n}$ (Equation (5.4))
$E^s(A), E^c(A), E^u(A)$	the stable, center, and unstable subspace of $A \in \mathbb{R}^{n \times n}$ (Definition 2.46)
$U_1 \oplus U_2$	the direct sum of two linear subspaces
$U^\perp$	$:= \{v \in V \mid \langle v, w \rangle = 0 \text{ for all } w \in U\}$, the orthogonal complement of $U \subseteq V$ in an inner product space $(V, \langle \cdot, \cdot \rangle)$

Norms and Inner Products

$\|x\|_p$	$:= \left(\sum_{i=1}^{n} \lvert x_i \rvert^p \right)^{1/p}$ for $x \in \mathbb{C}^n$ and $1 \leq p < \infty$
$\|x\|_\infty$	$:= \max_{i=1,\ldots,n} \lvert x_i \rvert$, the maximum norm of x
$\|x\|$	the Euclidean vector norm $\|x\|_2$
$\|A\|_p$	$:= \sup\{\|Au\|_p \mid \|u\|_p \leq 1\}$, the subordinate matrix p-norm, $1 \leq p \leq \infty$
$\operatorname{tr}(A)$	$:= \sum_{i=1}^{n} A_{i,i}$, the trace of $A \in \mathbb{C}^{n \times n}$
$\langle A, B \rangle_{\mathrm{F}}$	$:= \operatorname{tr}\big(B^{\mathsf{H}} A\big)$, the Frobenius inner product of $A, B \in \mathbb{C}^{m \times n}$
$\|A\|_{\mathrm{F}}$	$:= \sqrt{\langle A, A \rangle_{\mathrm{F}}}$, the Frobenius norm of $A \in \mathbb{C}^{m \times n}$
$\|A\|$	$:= \|A\|_2$, the 2-norm $A \in \mathbb{C}^{m \times n}$
$\kappa_2(A)$	$:= \|A\|_2 \, \|A^{-1}\|_2$, the condition number of $A \in \mathbb{C}^{n \times n}$ with respect to $\|\cdot\|_2$

Eigenvalues, Singular Values, and Numerical Range

$\Lambda(A)$	the spectrum of A
$\lambda_k^{\downarrow}(A)$	the k-th largest eigenvalue of Hermitian A
$\operatorname{spa}(A)$	$:= \max_{\lambda \in \Lambda(A)} \Re(\lambda)$, the spectral abscissa of A
$\sigma_k(A)$	the k-th largest singular value of A
$\mu_2[A]$	the logarithmic norm of A (Definition 2.25)

$\mathrm{W}(A)$	the numerical range of A (Definition 2.27)

Derivatives

$\dot{f}$	$:= \frac{\mathrm{d}f}{\mathrm{d}t}$, the first derivative of f with respect to t
$\ddot{f}$	$:= \frac{\mathrm{d}^2 f}{\mathrm{d}t^2}$, the second derivative of f with respect to t
$\frac{\partial f}{\partial x_i}$	the partial derivative with respect to x_i of f

Operators

T^*	the adjoint of a bounded linear operator $T\colon H \to H$ on a Hilbert space H (Definition/Theorem 2.10)

CHAPTER 1 INTRODUCTION

Contents

1.1 Motivation

This thesis deals with the numerical solution of the large-scale autonomous differential Riccati equation

$$\begin{aligned}\dot{X}(t) &= A^{\mathsf{T}}X(t) + X(t)A - X(t)BB^{\mathsf{T}}X(t) + C^{\mathsf{T}}C,\\ X(0) &= X_0\end{aligned}$$

The differential Riccati equation plays an important role in many fields of applied mathematics like model order reduction, optimal control and differential games; cf. [1, 22, 26, 66, 84]. The equation finds applications in symplectic geometry as well; cf. [11, 85].

As the differential Riccati equation is a nonlinear system of n^2 scalar equations, the numerical approximation of the solution $X(t) \in \mathbb{R}^{n\times n}$ comes with high computational complexity and storage demands if n is large. Consider for example $n = 10^5$ and assume that $X(t_k)$ is approximated by the matrix X_k of size $n \times n$. If X_k is stored in IEEE 754 double-precision, this would require about 80 Gigabytes memory.

Stationary points of the differential Riccati equation are solutions of the algebraic Riccati equation

$$0 = A^{\mathsf{T}}X + XA - XBB^{\mathsf{T}}X + C^{\mathsf{T}}C.$$

The algebraic Riccati equation has similar applications as the differential Riccati equation. Here, of primary interest are symmetric positive semidefinite solutions $X \in \mathbb{R}^{n\times n}$. During the last 20 years, efficient low-rank approaches were developed for the numerical approximation of the solution of the large-scale algebraic Riccati equation; cf. [19, 74, 82, 105, 123]. These approaches circumvent the large memory requirements by approximating the solution as a product of low-rank matrices $ZZ^\mathsf{T} \approx X$, where the numbers of columns of Z is much smaller than n, and are based on the property that the solution X often admits a quick eigenvalue decay; cf. [13, 18, 94, 108].

Example 1.1 (Algebraic Riccati Equation: Eigenvalue Decay, [40, p. 15]): We illustrate the eigenvalue decay of the symmetric positive semidefinite solution of the algebraic Riccati Equation by an example in Figure 1.1. We have chosen the matrices of the `TRIDIAG`(α) benchmark problem with $\alpha \in \{1, 2, 3, 5, 10, 50, 10^4\}$ (Table A.1). We have used the variable-precision arithmetic of MATLAB with 512 significant digits for the numerical solution approximation and eigenvalue computations. Figure 1.1 visualizes the eigenvalues arranged in a non-increasing order.

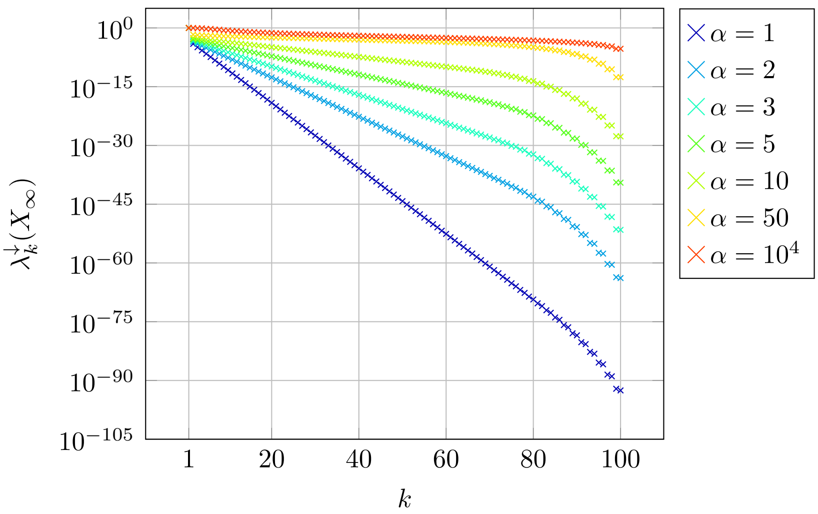


Fig. 1.1: Eigenvalue Decay of the Solution of the Algebraic Riccati Equation.

An implicit time discretization scheme applied to the differential Riccati leads to a series of algebraic Riccati equations with additional indefinite terms of low-rank. Therefore, the achievements in efficiently solving the algebraic Riccati equation have transferred to the solution of the differential Riccati equation. The relevant works analyze the discretization schemes in view of definiteness of the resulting low-rank approximations and propose variants of the low-rank approximation with possible indefinite low-rank approximations; [80, 113].

However, an indefinite low-rank approximation $L_k D_k L_k^\mathsf{T} \approx X(t_k)$ has still to be computed and stored for every grid-point in time. Here, the low-rank factor L_k is still of size $n \times l_k$. As calculated above, if n was 10^5 and the low-rank factors L_k had 100 columns each, then storing the approximation to the differential Riccati equation on 1000 discrete time steps would require approximately 80 Gigabytes of memory. Generally, the memory demand for these approaches scales with nn_t, where n_t is the number of discrete time points.

To decouple the memory and computational requirements in n and n_t, methods that approximate $X(t)$ by $Q\tilde{X}(t)Q^{\mathsf{T}}$ have gained interest recently. Here, usually the matrix $Q \in \mathbb{R}^{n \times k}$ has orthonormal columns and the time-dependent matrix $\tilde{X}(t)$ is of size $k \times k$. Hence, the memory requirements scale down to $kn + k^2 n_t$. These methods then differ in how the basis Q is chosen and how the approximated $\tilde{X}(t)$ is computed; see [49, 65, 70] for Krylov subspace approaches combined with backward differentiation formula time integration. A major result of this thesis adds an algorithm to this class of methods that chooses Q depending on a low-rank approximation of the associated algebraic Riccati equation and computes $\tilde{X}(t)$ by a suitable modification of the Davison–Maki method.

Although it could be seen as a special case of the differential Riccati equation, the differential Lyapunov equation

$$\begin{aligned} \dot{X}(t) &= A^{\mathsf{T}} X(t) + X(t)A + C^{\mathsf{T}}C, \\ X(0) &= X_0 \end{aligned}$$

is a major section of this thesis. The linearity of the differential Lyapunov equation allows for the application of spectral theory which we use to reassemble known solution representations, and that serves as a motivation for numerical approximations.

Similarly to the algorithm for the differential Riccati equation, this thesis proposes to parametrize the solution of the differential Lyapunov equation in the space spanned by the eigenvectors corresponding to the largest eigenvalues of the low-rank approximation of the stationary point and provides an analysis of this approximation.

All proposed numerical methods are thoroughly tested in benchmark examples and compared to state-of-the-art implementations of comparable approaches.

Even with advanced mathematical techniques, the numerical approximation of large-scale matrix valued differential equations easily reach dimensions that require the use of computing clusters. Because of differing individual configurations of compute servers, any implementation of numerical algorithms performs differently in different environments. This makes general assessments of the performance of the algorithms difficult, even if they are reported relative to other approaches and for acknowledged benchmark problems.

Therefore, a focus of this thesis has been laid on the development of a code repository that resorts to low-level routines or to established function implementation of MATLAB. The code and all scripts needed to reproduce the numerical experiments are made publicly available.

1.2 Outline of the Thesis

This thesis is organized as follows. In Chapter 2, necessary mathematical notation, concepts, and fundamental results from the theory of ordinary differential equations are introduced. In Chapters 3 and 4 and Chapters 5 and 6, the algebraic and differential

Code and Data Availability

The data and the source code of the implementations used to compute the presented results is available from:

`doi:10.5281/zenodo.4460618`

Fig. 1.2: Link to Source Code and Data.

Lyapunov and Riccati equations are treated, respectively. We first consider the algebraic version for both problem classes and derive and review solution representations and properties. Then the corresponding differential equations are described, and solution formulas and numerical solution approaches are discussed. Thirdly, the collected theoretical results and considerations are cast into an algorithm for numerical approximation. Where appropriate, generalizations are discussed. The numerical performance is discussed in the individual chapters, whereas the plots of the simulations are appended in Appendix D. The appendix also includes the specification of the hardware and software ecosystem. Furthermore, the appendix explains basic routines used for the approximation and a description of the benchmark examples. The thesis itself is completed by a conclusion in Chapter 7.

CHAPTER 2

MATHEMATICAL PRELIMINARIES

Contents

This chapter collects basic mathematical concepts used throughout this thesis. It is organized as follows. Sections 2.1 and 2.2 deal with *Hermitian positive semidefinite* (hpsd) matrices. The Loewner partial ordering is recalled, and its fundamental properties are summarized. An existence and uniqueness theorem about the hpsd matrix square root of a hpsd matrix is given. Section 2.3 summarizes bounded linear operators, adjoint operators, normal operators, projections, and best approximations. Section 2.4

presents some best approximation problems for matrices. Section 2.5 focuses on the matrix exponential and the exponential of a bounded linear operator. Then, Sections 2.6 and 2.7 consider the logarithmic norm, the numerical range, and states connections to the matrix exponential. Section 2.8 gives facts of the theory of *ordinary differential equations* (ODEs). Section 2.9 recalls the Hadamard product, the Kronecker product, and the vectorization of a matrix. Sections 2.10 and 2.11 review Hamiltonian and Krylov matrices. Finally, Section 2.12 is about the stable, center, and unstable subspace of a matrix.

2.1 Loewner Partial Order

We define the Loewner partial order. Theorem 2.4 presents some relevant properties.

Definition 2.1 ([59, Def. 7.7.1]):
Let $A, B \in \mathbb{C}^{n\times n}$ be Hermitian. We write $A \preccurlyeq B$ if $B - A$ is hpsd. ◊

Lemma 2.2 (Partial Order, [59, Sec. 7.7]):
The relation $\preccurlyeq$ is a partial order on the set of Hermitian matrices. That is, for all Hermitian $A, B, C \in \mathbb{C}^{n\times n}$, it holds:

(i) $A \preccurlyeq A$.

(ii) If $A \preccurlyeq B$ and $B \preccurlyeq A$, then $A = B$.

(iii) If $A \preccurlyeq B$ and $B \preccurlyeq C$, then $A \preccurlyeq C$. ◊

Definition 2.3 (Loewner Partial Order, [59, Sec. 7.7]):
The partial order $\preccurlyeq$ on the set of Hermitian matrices is called *Loewner partial order.* ◊

There are Hermitian matrices such that neither $A \preccurlyeq B$ nor $B \preccurlyeq A$ holds. Therefore, the Loewner partial order is not a total order.

Theorem 2.4 ([59, Thm. 7.7.2, Cor. 7.7.4, Prob. 7.1.P1]):
Let $A, B \in \mathbb{C}^{n\times n}$ be Hermitian, and $S \in \mathbb{C}^{n\times m}$. It holds:

(i) $\lambda_n^{\downarrow}(A)I \preccurlyeq A \preccurlyeq \lambda_1^{\downarrow}(A)I$.

(ii) $-\left\|A\right\|_2 I \preccurlyeq A \preccurlyeq \left\|A\right\|_2 I$.

(iii) If $A \preccurlyeq B$, then $S^{\mathsf{H}}AS \preccurlyeq S^{\mathsf{H}}BS$.

(iv) If $A \preccurlyeq B$, then $\lambda_k^{\downarrow}(A) \leq \lambda_k^{\downarrow}(B)$ for all $k = 1, \dots, n$.

(v) If $0 \preccurlyeq A \preccurlyeq B$, then $\left\|A\right\|_2 \leq \left\|B\right\|_2$.

(vi) If $0 \preccurlyeq A$, then $\left|A_{i,j}\right| \leq \sqrt{A_{i,i}A_{j,j}}$ for all $i, j = 1, \dots, n$. ◊

Theorem 2.4 (iv) is a direct consequence of the Courant–Fischer–Weyl min-max principle; cf. [59, Thm. 4.2.6]. If A is Hermitian, then

$$\|A\|_2 = \max_{\lambda \in \Lambda(A)} |\lambda| .$$

Therefore, (ii) and (v) of Theorem 2.4 follow from (i) and (iv), respectively.

Lemma 2.5 ([59, Obs. 7.1.6]):
Let $A \in \mathbb{C}^{n\times n}$ be hpsd and let $x \in \mathbb{C}^n$. Then

$$x^{\mathsf{H}} A x = 0 \iff Ax = 0. \qquad \Diamond$$

Lemma 2.6 (Image and Loewner Partial Order, [59, Prob. 7.7.P6]):
Let $A, B \in \mathbb{C}^{n\times n}$ be Hermitian. If $0 \preccurlyeq A \preccurlyeq B$, then $\operatorname{im}(A) \subseteq \operatorname{im}(B)$. $\Diamond$

2.2 Matrix Square Root

The principal square root of a number can be generalized to Hermitian matrices. Here, we recall the existence and uniqueness of a hpsd matrix square root.

Definition/Theorem 2.7 (Matrix Square Root, [59, Thm. 7.2.6]):
Let $A \in \mathbb{C}^{n\times n}$ be hpsd. Then there is a unique hpsd matrix $B \in \mathbb{C}^{n\times n}$ such that $B^2 = A$ hold. We define the *matrix square root* of A as $A^{1/2} := B$. If A is real, then $A^{1/2}$ is real. $\Diamond$

Clearly, by considering diagonal matrices, the equation $B^2 = A$ may have multiple solutions. The condition "B is hpsd" ensures uniqueness.

2.3 Linear Functional Analysis

Section 2.3 is divided into Sections 2.3.1, 2.3.2, and 2.3.3. Section 2.3.1 reviews equivalent conditions for the continuity of a linear operator and the operator norm definition. Adjoint and normal operators on a Hilbert space are part of Section 2.3.2. Then, Section 2.3.3 collects the primary results of projections and best approximations. Throughout Section 2.3, we consider all linear spaces to be over the field of real or complex numbers.

2.3.1 Bounded Linear Operators

We start with a characterization of the continuity of linear operators.

Lemma 2.8 (Continuity of Linear Operators, [3, Sec. 5.1]):
Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed linear spaces. $T\colon X \to Y$ linear, and $x_0 \in X$. Then the conditions are equivalent:

(i) T is continuous.

(ii) T is continuous at x_0.

(iii) $\sup\limits_{\|x\|_X \leq 1} \|Tx\|_Y < \infty$.

(iv) There exists a constant C with $\|Tx\|_Y \leq C \|x\|_X$ for all $x \in X$. ◊

Next, we continue with bounded linear operators and their operator norm.

Definition 2.9 (Bounded Linear Operator/Operator Norm, [62, Sec. 3.1]):
Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed linear spaces. If $T\colon X \to Y$ is linear and continuous, then T is called *bounded linear*. The *operator norm* of T is defined as

$$\|T\| := \sup_{\|x\|_X \leq 1} \|Tx\|_Y \,. \qquad \lozenge$$

If X is finite-dimensional, then each linear $T\colon X \to Y$ is bounded linear; cf. [62, Sec. 3.1].

2.3.2 Hilbert Space Adjoint Operator and Normal Operator

Here, we consider bounded linear operators acting on Hilbert spaces. We review the definitions of an adjoint of an operator and a normal operator.

Definition/Theorem 2.10 (Hilbert Space Adjoint, [62, Sec. 7.3, Thm. 7.5]):
Let $(H, \langle \cdot, \cdot \rangle_H)$ and $(G, \langle \cdot, \cdot \rangle_G)$ be Hilbert spaces. If $T\colon H \to G$ is bounded linear, then there exists a unique bounded linear $T^*\colon G \to H$ such that

$$\langle Tx, y \rangle_G = \langle x, T^* y \rangle_H \,,$$

for all $x \in H$ and $y \in G$. The operator T^* is called the *adjoint* of T. ◊

Definition 2.11 (Normal Operator, [62, Sec. 7.3, Def.]):
Let $(H, \langle \cdot, \cdot \rangle_H)$ be a Hilbert space. A bounded linear operator $T\colon H \to H$ is called *normal* if

$$TT^* = T^* T. \qquad \lozenge$$

2.3.3 Projection and Best Approximation

We are concerned with the best approximation problem in a Hilbert space and projections.

Definition/Theorem 2.12 (Best Approximation, [3, Thm. 4.3]):
Let $(H, \langle\cdot,\cdot\rangle_H)$ be a Hilbert space, and let $Y \subseteq X$ be nonempty, closed, and convex. Then there exists a unique function $P\colon X \to X$ such that

$$\|x - P(x)\|_H = \inf_{y\in Y} \|x - y\|_H$$

for all $x \in H$. The map P is called the *(orthogonal) projection onto* Y. ◊

Definition 2.13 (Projection, [3, Def. 9.13]):
Let X be a linear space, and $Y \subseteq X$ be a subspace of X. A linear function $P\colon X \to X$ is called *(linear) projection onto* Y if

$$P^2 = P \text{ and } \operatorname{im}(P) = Y. \qquad ◊$$

The following theorem gives equivalent characterizations of the orthogonal projection.

Theorem 2.14 (Orthogonal Projection, [3, Lem. 9.18]):
Let $(H, \langle\cdot,\cdot\rangle_H)$ be a Hilbert space, and $P\colon H \to H$ be linear. Then the conditions are equivalent:

(i) P is the orthogonal projection onto $\operatorname{im}(P)$, i.e., $\|x - Px\| \le \|x - Py\|$ for all $x, y \in H$.

(ii) $\langle x - Px, Py\rangle_H = 0$ for all $x, y \in H$.

(iii) $P^2 = P$ and $\langle Px, y\rangle_H = \langle x, Py\rangle_H$ for all $x, y \in H$.

(iv) $P^2 = P$ and P is bounded linear, with $\|P\| \le 1$. ◊

2.4 Matrix Nearness Problems

Section 2.4 deals with some best approximation problems for matrices. The term "matrix nearness problems" is also frequently used in the literature; cf. [55]. The problem consists of finding the nearest matrix with some prescribed properties to a given matrix X. The nearest matrix of rank at most k is given in terms of the *singular value decomposition* (SVD) (Theorem 2.15). We also consider the nearest Hermitian matrix (Theorem 2.16) and the nearest hpsd matrix problems (Theorem 2.18).

2.4.1 Best Low-Rank Approximation

Theorem 2.15 (Eckart–Young/Schmidt–Mirsky, [117, Thm. 5.8–5.9]):
Let $X = U\Sigma V^{\mathsf{H}}$ be an SVD of $X \in \mathbb{C}^{n\times m}$ with unitary matrices

$$U = \begin{bmatrix} u_1, \dots, u_n \end{bmatrix} \in \mathbb{C}^{n\times n}, \quad V = \begin{bmatrix} v_1, \dots, v_m \end{bmatrix} \in \mathbb{C}^{m\times m},$$

and diagonal matrix

$$\Sigma = \operatorname{diag}\bigl(\sigma_1(X), \dots, \sigma_{\min\{n,m\}}(X)\bigr) \in \mathbb{R}^{n\times m}.$$

For any number $0 \le k < \min\{n,m\}$ let $X_k = \sum\limits_{i=1}^{k} \sigma_i(X) u_i v_i^{\mathsf{H}}$. Then

$$\begin{aligned}
\|X - X_k\|_2 &= \inf_{\substack{Y\in\mathbb{C}^{n\times m}\\ \operatorname{rank}(Y)\le k}} \|X - Y\|_2 = \sigma_{k+1}(X),\\
\|X - X_k\|_{\mathrm{F}} &= \inf_{\substack{Y\in\mathbb{C}^{n\times m}\\ \operatorname{rank}(Y)\le k}} \|X - Y\|_{\mathrm{F}} = \sqrt{\sigma_{k+1}(X)^2 + \cdots + \sigma_{\min\{n,m\}}(X)^2}.
\end{aligned}$$ ◊

2.4.2 Best Hermitian Approximation

Theorem 2.16 (Fan/Hoffman, [56, Thm. 8.7]):
Let $X_{\mathrm{H}} = \frac{1}{2}\bigl(X + X^{\mathsf{H}}\bigr)$ be the Hermitian part of $X \in \mathbb{C}^{n\times n}$ and let $\|\cdot\|$ be any unitarily invariant norm. Then X_{H} is a best possible Hermitian approximation to X, i.e.,

$$\|X - X_{\mathrm{H}}\| = \inf\{\|X - Y\| \mid Y \textit{ is Hermitian}\}.$$

The solution is unique in the Frobenius norm. ◊

2.4.3 Best Hermitian Positive Semidefinite Approximation

Definition/Theorem 2.17 (Polar Decomposition, [56, Thm. 8.1]):
Let $X \in \mathbb{C}^{m\times n}$ with $m \ge n$. There exists a matrix $U \in \mathbb{C}^{m\times n}$ with orthonormal columns and a unique hpsd matrix $P \in \mathbb{C}^{n\times n}$ such that

$$A = UP.$$

The decomposition is called *polar decomposition.* ◊

Theorem 2.18 (Higham, [56, Thm. 8.8]):
Let $X \in \mathbb{C}^{n\times n}$ be, $X_{\mathrm{H}} = \frac{1}{2}\bigl(X + X^{\mathsf{H}}\bigr)$ be the Hermitian part of X, and let $X_{\mathrm{H}} = UP$ be a polar decomposition. Then $X_{\mathrm{psd}} = \frac{1}{2}(X_{\mathrm{H}} + P)$ is the unique solution to

$$\bigl\|X - X_{\mathrm{psd}}\bigr\|_{\mathrm{F}} = \inf\{\|X - Y\|_{\mathrm{F}} \mid Y \textit{ is hpsd}\}.$$ ◊

Results on best hpsd approximations in the 2–norm and the computational aspects of the problem are studied in [53, 55].

Remark 2.19 (Nearest hpsd Matrix to a Hermitian Matrix):
Let $X \in \mathbb{C}^{n\times n}$ be Hermitian, and $X = QDQ^{\mathsf{H}}$ be its spectral decomposition. Then

$$X = \bigl(QD_UQ^{\mathsf{H}}\bigr)\bigl(QD_PQ^{\mathsf{H}}\bigr) = UP$$

is a polar decomposition of X, where D_U and D_P are diagonal matrices such that

$$(D_U)_{i,i} = \operatorname{sgn}\big(D_{i,i}\big) \quad \text{and} \quad (D_P)_{i,i} = \big|D_{i,i}\big| \,.$$

The best possible hpsd approximation is given by

$$X_{\text{psd}} = \tfrac{1}{2}(X + P) = \tfrac{1}{2}Q(D + D_P)Q^{\mathsf{H}}.$$

Therefore, the negative eigenvalues of X are replaced by 0, and the nonnegative ones remain unchanged. $\Diamond$

2.5 Matrix Exponential

In this section, we review the matrix exponential and the exponential of a bounded linear operator. Theorem 2.21 summarizes essential properties of the matrix exponential. We begin with the definition of the matrix exponential of a square matrix.

Definition 2.20 (Matrix Exponential, [125, §18 II.]):
The *matrix exponential* of $A \in \mathbb{C}^{n\times n}$ is defined to be

$$e^A := \sum_{k=0}^{\infty} \tfrac{1}{k!} A^k. \qquad \Diamond$$

Theorem 2.21 (Matrix Exponential, [125, §18 III., §18 V.]):
For the matrix exponential, it holds:

(i) $\frac{\mathrm{d}}{\mathrm{d}t} e^{tA} = Ae^{tA}$ for all $t \in \mathbb{R}$.

(ii) If $AB = BA$, then $e^{A+B} = e^A e^B$.

(iii) If B is nonsingular, then $e^{B^{-1}AB} = B^{-1}e^A B$.

(iv) $\big(e^A\big)^{-1} = e^{-A}$.

(v) $\big(e^A\big)^{\mathsf{H}} = e^{A^{\mathsf{H}}}$.

(vi) If $A = \operatorname{diag}(a_1, \ldots, a_n)$, then $e^A = \operatorname{diag}(e^{a_1}, \ldots, e^{a_n})$. $\Diamond$

Definition 2.22 (Exponential of an Operator, [3, Ch. 5, Ex. 5.10]):
Let $(X, \|\cdot\|_X)$ be a Banach space over the complex or real numbers. The *exponential of a bounded linear operator* $A\colon X \to X$ is defined to be

$$e^A := \sum_{k=0}^{\infty} \tfrac{1}{k!} A^k. \qquad \Diamond$$

The exponential of a bound linear operator is a bounded linear operator, and the properties Theorem 2.21 (i)–(iv) also apply; cf. [3, Ch. 3 Ex. 3.10].

2.6 Logarithmic Norm

We review bounds on the norm of the matrix exponential. For that, we recall the logarithmic norm introduced in 1958 by Dahlquist and Lozinskii; cf. [107]. We start with the definition of a stable matrix.

Definition 2.23 (Stable Matrix):
If each eigenvalue of $A \in \mathbb{C}^{n\times n}$ has a negative real-part, then A is called *stable*, i.e.,

$$A \text{ is stable} :\Longleftrightarrow \Lambda(A) \subseteq \mathbb{C}_-. \qquad \Diamond$$

Theorem 2.24 (Matrix Exponential Bound, [56, Thm. 10.12], [125, §29 V.]):
Let $A \in \mathbb{C}^{n\times n}$ be and $\alpha \in \mathbb{R}$ with $\operatorname{spa}(A) < \alpha$. Then there is a positive constant γ such that

$$e^{t\operatorname{spa}(A)} \leq \left\|e^{tA}\right\|_2 \leq \gamma e^{\alpha t} \tag{2.1}$$

for all $t \geq 0$. $\Diamond$

If A is stable, then the constant α can be chosen negatively, and the norm of the matrix exponential decays exponentially. Conversely, if there is an eigenvalue with a positive real part, the norm of the matrix exponential grows exponentially. The constant γ depends on α and may, in general, not be known explicitly.

If A is normal, then A can be unitarily diagonalized. The 2-norm is unitarily invariant, and we obtain

$$\left\|e^{tA}\right\|_2 = \left\|e^{tUDU^{\mathsf{H}}}\right\|_2 = \left\|Ue^{tD}U^{\mathsf{H}}\right\|_2 = \left\|e^{tD}\right\|_2 = \sqrt{\max_{\lambda\in\Lambda(A)} e^{t\bar{\lambda}}e^{t\lambda}} = e^{t\operatorname{spa}(A)} \tag{2.2}$$

for all $t \geq 0$. Therefore, the norm of the matrix exponential of a normal matrix can be bounded without an additional constant. By introducing the logarithmic norm, we can extend this to a larger class of matrices.

Definition 2.25 (Logarithmic Norm, [39, Ch. II 8], [107]):
The *logarithmic norm* of $A \in \mathbb{C}^{n\times n}$ is the right-sided derivative of $\|\cdot\|_2$ at I in direction A, i.e.,

$$\mu_2[A] := \lim_{h\searrow 0} \frac{\|I + hA\|_2 - 1}{h}. \tag{2.3}$$

$\Diamond$

Each convex function defined on a linear space has a right-sided derivative. Hence, the limit (2.3) exists; cf. [60, Thm. 3.50].

Theorem 2.26 ([39, Ch. II 8], [60, Thm. 3.51], [107], [118, Thm. 17.4]):
For the logarithmic norm, it holds:

(i) $\mu_2[\alpha A] = \alpha\mu_2[A]$ for all $\alpha \geq 0$.

(ii) $\mu_2[A+B] \leq \mu_2[A] + \mu_2[B]$.

(iii) $\operatorname{spa}(A) \leq \mu_2[A]$.

(iv) $\mu_2[A] \leq \|A\|_2$.

(v) $\mu_2[A] = \lambda_1^{\downarrow}\big(\frac{1}{2}(A + A^{\mathsf{H}})\big)$.

(vi) $\left\|e^{tA}\right\|_2 \leq e^{t\mu_2[A]}$ for all $t \geq 0$.

(vii) The logarithmic norm is the initial slope of the norm of the matrix exponential, i.e.,

$$\lim_{h \searrow 0} \frac{\left\|e^{hA}\right\|_2 - 1}{h} = \mu_2[A]. \qquad \Diamond$$

If the logarithmic norm $\mu_2[A]$ is negative, then Theorem 2.26 (iii) ensures that A is stable. Generally, the converse is not true.

If A is normal, then Theorem 2.26 (iii) and (v) yield

$$\lambda_1^{\downarrow}\big(\tfrac{1}{2}(A + A^{\mathsf{H}})\big) = \operatorname{spa}(A) = \mu_2[A].$$

Consequently, A is normal and stable is equivalent to A is normal and the logarithmic norm $\mu_2[A]$ is negative – and Equation (2.2) can be stated as $\left\|e^{tA}\right\|_2 = e^{t\mu_2[A]}$.

On the other hand, the logarithmic norm of a nonnormal matrix might be nonnegative. We summarize.

$$A \text{ is normal and } \mu_2[A] < 0 \Leftrightarrow A \text{ is normal and stable} \Rightarrow \mu_2[A] < 0 \Rightarrow A \text{ is stable.}$$

Therefore, the logarithmic norm concept gives bounds on the matrix exponential without additional constants under more general assumptions; cf. Theorem 2.26 (vi).

2.7 Numerical Range

We present the definition of the numerical range. Theorem 2.28 collects some properties of the numerical range. Example 2.1 illustrates a few aspects of Theorems 2.26 and 2.28.

Definition 2.27 (Numerical Range, [58, Def. 1.1.1]):
The *numerical range* of $A \in \mathbb{C}^{n\times n}$ is defined by

$$\mathrm{W}(A) := \left\{x^{\mathsf{H}}Ax \mid x \in \mathbb{C}^n, \|x\|_2 = 1\right\}. \qquad \Diamond$$

Theorem 2.28 ([58, Prop. 1.2.1–1.2.6, Prop. 1.2.9, Lem. 1.5.7]):
For the numerical range, it holds:

(i) Toeplitz–Hausdorff Theorem: $\mathrm{W}(A)$ is a compact and convex set in $\mathbb{C}$.

(ii) If A is normal, then $\mathrm{W}(A)$ is the convex hull of $\Lambda(A)$.

(iii) Each eigenvalue of A is contained in $\mathrm{W}(A)$, i.e., $\Lambda(A) \subseteq \mathrm{W}(A)$.

(iv) $\mathrm{W}(A + zI) = \mathrm{W}(A) + \{z\}$ and $\mathrm{W}(zA) = z\,\mathrm{W}(A)$ for all $z \in \mathbb{C}$.

(v) $\mathrm{W}(A + B) \subseteq \mathrm{W}(A) + \mathrm{W}(B)$.

(vi) $\Re(\mathrm{W}(A)) = \mathrm{W}\big(\frac{1}{2}\left(A + A^{\mathsf{H}}\right)\big)$ and $\Im(\mathrm{W}(A)) = \mathrm{W}\big(\frac{1}{2\imath}\left(A - A^{\mathsf{H}}\right)\big)$.

(vii) $\max\{\Re(z) \mid z \in \mathrm{W}(A)\} = \lambda_1^{\downarrow}\big(\frac{1}{2}\left(A + A^{\mathsf{H}}\right)\big)$.

(viii) If the columns of $Q \in \mathbb{C}^{n\times k}$ are orthonormal, then $\mathrm{W}\big(Q^{\mathsf{H}}AQ\big) \subseteq \mathrm{W}(A)$. ◊

Theorem 2.28 (vi) gives information about the projection of the numerical range onto the real and imaginary axis. The matrices

$$\tfrac{1}{2}\left(A + A^{\mathsf{H}}\right) \quad \text{and} \quad \tfrac{1}{2\imath}\left(A - A^{\mathsf{H}}\right)$$

are Hermitian. Therefore, their numerical ranges are the convex hulls of their extremal eigenvalues.

Combining Theorem 2.26 (v) and Theorem 2.28 (vii) yields that the logarithmic norm $\mu_2[A]$ is the rightmost extent of the numerical range of A. Using Theorem 2.28 (viii), we have $\mu_2\big[Q^{\mathsf{H}}AQ\big] \leq \mu_2[A]$. In particular, if the logarithmic norm of A is negative, $Q^{\mathsf{H}}AQ$ is stable.

Example 2.1 (Numerical Range and Logarithmic Norm):
We illustrate the numerical range, the logarithmic norm, and the bound of the matrix exponential by an example in Figure 2.1. We consider the matrices

$$A_1 = \tfrac{1}{2}\begin{bmatrix} -1 & 4 \\ -1 & -1 \end{bmatrix} \quad \text{and} \quad A_2 = \tfrac{1}{2}\begin{bmatrix} -1 & 2 \\ -1 & -1 \end{bmatrix}.$$

Both matrices are stable, and the real part of all eigenvalues of A_1 and A_2 is equal. We have

$$\Lambda(A_1) = \left\{-\tfrac{1}{2} \pm 1\imath\right\} \quad \text{and} \quad \Lambda(A_2) = \left\{-\tfrac{1}{2} \pm \tfrac{1}{\sqrt{2}}\imath\right\}.$$

Moreover, $\mu_2[A_1] = \frac{1}{4}$ and $\mu_2[A_2] = -\frac{1}{4}$. For the numerical ranges, it holds

$$\begin{aligned} \Re(\mathrm{W}(A_1)) &= \left[-\tfrac{5}{4}, \tfrac{1}{4}\right], & \Im(\mathrm{W}(A_1)) &= \left[-\tfrac{5}{4}, \tfrac{5}{4}\right], \\ \Re(\mathrm{W}(A_2)) &= \left[-\tfrac{3}{4}, -\tfrac{1}{4}\right], & \Im(\mathrm{W}(A_2)) &= \left[-\tfrac{3}{4}, \tfrac{3}{4}\right]. \end{aligned}$$ ◊

Figure 2.1 visualizes the numerical range, the spectrum, and the logarithmic norm of A_1 and A_2. Figure 2.2 shows the norm of the matrix exponential and the logarithmic norm bound on the interval $[0, 2]$. The logarithmic norm of A_1 is positive. Therefore, the bound $e^{t\mu_2[A_1]}$ grows exponentially, whereas the norm of the matrix exponential exhibits a transient growth phase and decays exponentially after some time; cf. Figure 2.2a.

In Figure 2.2b, both quantities decay exponentially, as the logarithmic norm of A_2 is negative. Moreover, the functions $\left\|e^{tA_i}\right\|_2$ and $e^{t\mu_2[A_i]}$ have the same initial slope at $t = 0$; cf. Theorem 2.26 (vii).

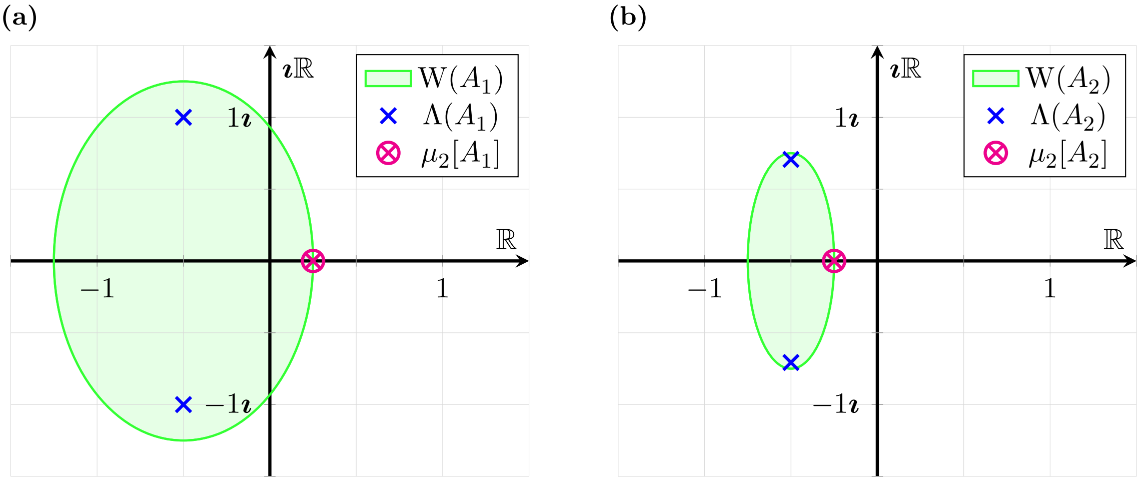


Fig. 2.1: **(a), (b)**: The Numerical Range, the Spectrum, and the Logarithmic Norm of A_1 and A_2.

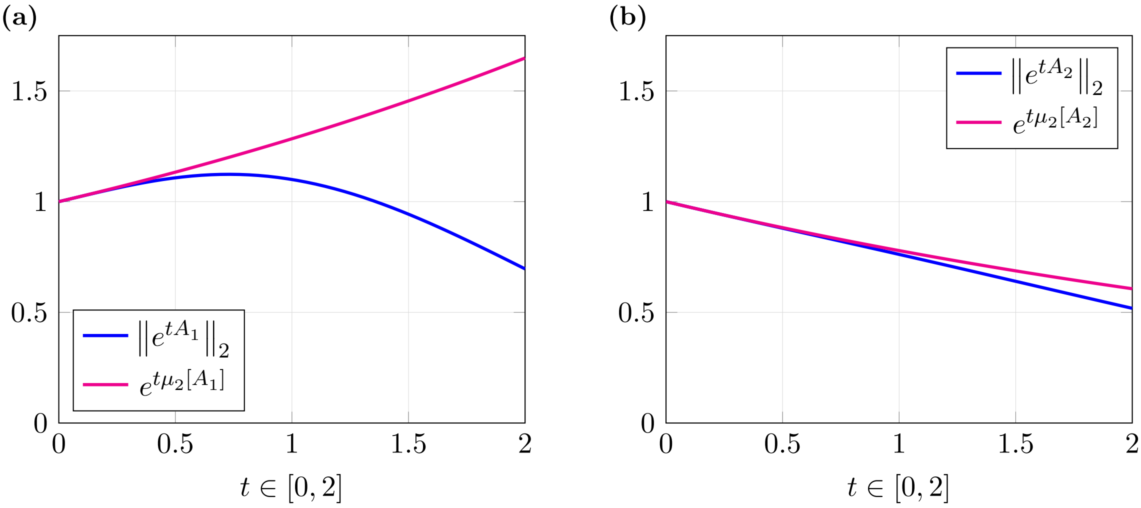


Fig. 2.2: **(a), (b)**: The Norm of the Matrix Exponential and the Logarithmic Norm Bound.

2.8 Ordinary Differential Equations

Section 2.8 recalls the fundamental results of the theory of ODEs. Section 2.8.1 reviews the existence and uniqueness results of solutions of *initial value problems* (IVPs). Finally, Section 2.8.2 focuses on linear IVPs.

2.8.1 Existence and Uniqueness of Solutions

We start with the Peano existence theorem, the Picard-Lindelöf theorem, and a result on maximal solutions.

Theorem 2.29 (Peano, Picard–Lindelöf, and Maximal Solutions): Let $G \subseteq \mathbb{R} \times \mathbb{R}^n$ be an open set with $(t_0, x_0) \in G$. We consider the IVP

$$\dot{x}(t) = f(t, x(t)), \tag{2.4a}$$
$$x(t_0) = x_0, \tag{2.4b}$$

where $f \colon G \to \mathbb{R}^n$. It holds:

- If f is continuous, then the IVP (2.4) has a maximal solution
$$x \colon (t_-, t_+) \to \mathbb{R}^n$$
with $t_0 \in (t_-, t_+)$; cf. [95, Thm. 6.1.1, Thm. 6.2.1].

- If f is continuous and locally Lipschitz continuous with respect to x, then the IVP (2.4) has a unique maximal solution
$$x \colon (t_-, t_+) \to \mathbb{R}^n$$
with $t_0 \in (t_-, t_+)$; cf. [95, Thm. 2.2.2, Thm. 2.3.2].

We have the following alternatives for the right-endpoint t_+:

(i) $t_+ = \infty$.

(ii) $t_+ < \infty$ and $\lim_{r \searrow 0} \inf\{\operatorname{dist}((t, x(t)), \partial G) \mid t_+ - r < t < t_+\} = 0$.

(iii) $t_+ < \infty$ and $\lim_{r \searrow 0} \inf\{\operatorname{dist}((t, x(t)), \partial G) \mid t_+ - r < t < t_+\} > 0$ and $\lim_{t \nearrow t_+} \|x(t)\| = \infty$.

Analogous conditions hold for the left-endpoint t_-; cf. [95, Thm. 2.3.2, Thm. 6.2.1]. ◊

The alternative (ii) of Theorem 2.29 means that the solution comes arbitrarily close to the boundary of G. The case (iii) is commonly referred as *escape in finite-time*. In particular, if $G = \mathbb{R} \times \mathbb{R}^n$, then $\partial G = \varnothing$ and $\operatorname{dist}((t, x(t)), \partial G) = \infty$. Therefore, either $t_+ = \infty$ holds or the solution escapes in finite-time.

Usually, we omit the word "maximal" and write (t_-, t_+) to refer to the *maximal interval of existence*. In general, the maximal interval of existence depends on t_0 and on the initial value x_0.

Lemma 2.30 (Locally Lipschitz Continuous, [95, Prop. 2.1.5], [125, §10 V.]): Let $G \subseteq \mathbb{R} \times \mathbb{R}^n$ be an open set and $f \colon G \to \mathbb{R}^n$ be continuous. If all partial derivatives $\frac{\partial f}{\partial x_i}(t, x)$ are continuous, then f is locally Lipschitz continuous with respect to x. ◊

2.8.2 Linear Initial Value Problems

We recall results about linear IVPs. We begin with the homogeneous linear IVP.

Theorem 2.31 (Wronski, [125, §15 I., §15 III.]):
Let $\mathbb{I} \subseteq \mathbb{R}$ be an open interval with $t_0 \in \mathbb{I}$, $A\colon \mathbb{I} \to \mathbb{R}^{n\times n}$ be a continuous matrix-valued function, and $\Phi_0 \in \mathbb{R}^{n\times n}$. The homogeneous linear IVP

$$\begin{aligned}\dot{\Phi}(t) &= A(t)\Phi(t),\\ \Phi(t_0) &= \Phi_0\end{aligned}$$

has a unique solution $\Phi\colon \mathbb{I} \to \mathbb{R}^{n\times n}$, and the determinant of $\Phi(t)$ fulfills

$$\det(\Phi(t)) = \det(\Phi_0)\exp\left(\int\limits_{t_0}^{t} \operatorname{tr}(A(s))\,\mathrm{d}s\right). \tag{2.5}$$

◊

Equation (2.5) is also known as the Abel–Jacobi–Liouville identity. The determinant $\det(\Phi(t))$ is called *Wronskian determinant* or just *Wronskian*. If Φ_0 is nonsingular, then $\Phi(t)$ is nonsingular for all $t \in \mathbb{I}$, and $\Phi(t)$ is called *fundamental matrix* of the system $\dot{x}(t) = A(t)x(t)$.

Theorem 2.32 (Variation of Constants Formula, [125, §16 III.]):
Let $\mathbb{I} \subseteq \mathbb{R}$ be an open interval with $t_0 \in \mathbb{I}$, $A\colon \mathbb{I} \to \mathbb{R}^{n\times n}$ and $b\colon \mathbb{I} \to \mathbb{R}^n$ be continuous functions, and $x_0 \in \mathbb{R}^n$. The inhomogeneous linear IVP

$$\begin{aligned}\dot{x}(t) &= A(t)x(t) + b(t),\\ x(t_0) &= x_0\end{aligned}$$

has a unique solution $x\colon \mathbb{I} \to \mathbb{R}^n$ given by

$$x(t) = \Phi(t)x_0 + \int\limits_{t_0}^{t} \Phi(t)\Phi(s)^{-1}b(s)\,\mathrm{d}s,$$

where $\Phi(t)$ is the fundamental matrix of the system $\dot{x}(t) = A(t)x(t)$ with $\Phi(t_0) = I$. ◊

Theorem 2.33 (Variation of Constants Formula, [125, §18 VI.]):
Let $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $x_0 \in \mathbb{R}^n$, and $t_0 \in \mathbb{R}$. The autonomous inhomogeneous linear IVP

$$\begin{aligned}\dot{x}(t) &= Ax(t) + b,\\ x(t_0) &= x_0\end{aligned}$$

has a unique solution $x\colon \mathbb{R} \to \mathbb{R}^n$ given by

$$x(t) = e^{(t-t_0)A}x_0 + \int\limits_{t_0}^{t} e^{(t-s)A}b\,\mathrm{d}s.$$

◊

2.9 Hadamard Product, Kronecker Product, and Vectorization

This section recalls the Hadamard product, the Kronecker product, and the vectorization of a matrix.

Definition 2.34 (Hadamard Product, [58, Def. 5.0.1]):
The *Hadamard product* of $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times n}$ is defined by

$$A \odot B := \big(A_{i,j}B_{i,j}\big)_{\substack{i=1,\ldots,m\\ j=1,\ldots,n}}. \qquad \lozenge$$

Definition 2.35 (Kronecker Product, [58, Def. 4.2.1]):
The *Kronecker product* of $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$ is denoted by $A \otimes B$ and is defined to be the block matrix

$$A \otimes B := \begin{bmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & \ddots & \vdots \\ A_{m,1}B & \cdots & A_{m,n}B \end{bmatrix} \in \mathbb{C}^{mp\times nq}. \qquad \lozenge$$

Definition 2.36 (Vectorization, [58, Def. 4.2.9]):
The *vectorization* of $A \in \mathbb{C}^{m\times n}$ is defined to be

$$\operatorname{vec}(A) := \big[A_{1,1}, \ldots, A_{m,1}, A_{1,2}, \ldots, A_{m,2}, \ldots, A_{1,n}, \ldots, A_{m,n}\big]^{\mathsf{T}} \in \mathbb{C}^{mn}. \qquad \lozenge$$

Lemma 2.37 (Kronecker Product and Vectorization, [58, Lem. 4.3.1]):
Let $A \in \mathbb{C}^{m\times n}$, $X \in \mathbb{C}^{n\times p}$, and $B \in \mathbb{C}^{p\times q}$. It holds:

$$\operatorname{vec}(AXB) = \big(B^{\mathsf{T}} \otimes A\big)\operatorname{vec}(X). \qquad \lozenge$$

Theorem 2.38 (Kronecker Products and Eigenvalues, [58, Thm. 4.4.5]):
Let $A \in \mathbb{C}^{m\times m}$ and $B \in \mathbb{C}^{n\times n}$. It holds:

$$\Lambda(I_n \otimes A + B \otimes I_m) = \Lambda(A) + \Lambda(B). \qquad \lozenge$$

Lemma 2.39 (Mixed Product Property, [58, Lem. 4.2.10]):
Let $A \in \mathbb{C}^{m\times n}, B \in \mathbb{C}^{p\times q}, C \in \mathbb{C}^{n\times k}$, and $D \in \mathbb{C}^{q\times r}$. It holds:

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

2.10 Hamiltonian Matrices

We define a Hamiltonian matrix and collect some valuable properties.

Definition 2.40 (Matrix $\mathcal{J}$, [24, Sec. 1.5]):
We define

$$\mathcal{J} := \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \in \mathbb{C}^{2n \times 2n}. \qquad \Diamond$$

Lemma 2.41 (Properties of $\mathcal{J}$, [1, Def. 2.1.1], [24, Sec. 1.5]):
It holds:

$$\mathcal{J}^{-1} = \mathcal{J}^{\mathsf{T}} = -\mathcal{J} \quad \text{and} \quad \mathcal{J}^2 = -I. \qquad \Diamond$$

Definition 2.42 (Hamiltonian Matrix, [1, Def. 2.1.1], [24, Sec. 1.5]):
A matrix $\mathcal{H} \in \mathbb{C}^{2n \times 2n}$ is called *Hamiltonian* if $\mathcal{J}\mathcal{H}$ is Hermitian, i.e.,

$$\mathcal{J}\mathcal{H} = (\mathcal{J}\mathcal{H})^{\mathsf{H}}. \qquad \Diamond$$

Definition 2.43 (Symplectic Matrix, [1, Def. 2.1.1], [24, Sec. 1.5]):
A matrix $M \in \mathbb{C}^{2n \times 2n}$ is called *symplectic* if

$$M^{\mathsf{H}} \mathcal{J} M = \mathcal{J}. \qquad \Diamond$$

Lemma 2.44 (Hamiltonian Matrix, [1, Lem. 2.2.3], [24, Sec. 1.5]):
Let $\mathcal{H} \in \mathbb{C}^{2n \times 2n}$ be a Hamiltonian matrix. It holds:

(i) If $\mathcal{H}$ is nonsingular, then $\mathcal{H}^{-1}$ is Hamiltonian.

(ii) The spectrum of $\mathcal{H}$ is symmetric around the origin, i.e., $\lambda \in \Lambda(\mathcal{H})$ implies that $-\overline{\lambda} \in \Lambda(\mathcal{H})$.

(iii) The matrix exponential $e^{t\mathcal{H}}$ is symplectic for all $t \in \mathbb{R}$.

(iv) Let $S, Q \in \mathbb{C}^{n \times n}$ be Hermitian and $A \in \mathbb{C}^{n \times n}$. The block matrix

$$\begin{bmatrix} A & -S \\ -Q & -A^{\mathsf{H}} \end{bmatrix}$$

is Hamiltonian. Conversely, each Hamiltonian matrix has such a block structure. $\Diamond$

2.11 Krylov Matrix

We define the Krylov matrix and recall some properties of its image.

Definition 2.45 (Krylov Matrix, [25, Def. 1.1.9]):
Let $A \in \mathbb{C}^{n\times n}$ and $B \in \mathbb{C}^{n\times b}$. The *Krylov matrix* generated by A and B is defined to be

$$\mathcal{K}(A,B) := \left[B, AB, \dots, A^{n-1}B\right] \in \mathbb{C}^{n\times nb}. \qquad \Diamond$$

The image of the Krylov matrix $\operatorname{im}(\mathcal{K}(A,B)) \subseteq \mathbb{C}^n$ is an A-invariant subspace, and, generally, its dimension is not n. The orthogonal complement of the image of the Krylov matrix is the kernel of the conjugated transposed Krylov matrix, i.e.,

$$\operatorname{im}(\mathcal{K}(A,B))^{\perp} = \ker\bigl(\mathcal{K}(A,B)^{\mathsf{H}}\bigr).$$

It holds

$$x \in \ker\bigl(\mathcal{K}(A,B)^{\mathsf{H}}\bigr) \iff \begin{bmatrix} B^{\mathsf{H}} \\ B^{\mathsf{H}}A^{\mathsf{H}} \\ \vdots \\ B^{\mathsf{H}}A^{\mathsf{H}^{n-1}} \end{bmatrix} x = 0.$$

Furthermore, the Cayley–Hamilton theorem yields

$$x \in \ker\bigl(\mathcal{K}(A,B)^{\mathsf{H}}\bigr) \iff B^{\mathsf{H}}A^{\mathsf{H}^{k}}x = 0 \text{ for all } k \in \mathbb{N}_0.$$

The linear space $\ker\bigl(\mathcal{K}(A,B)^{\mathsf{H}}\bigr)$ is an A^{H}-invariant subspace.

2.12 Stable, Center, and Unstable Subspace

We consider the stable, center, and unstable subspace of a real square matrix.

Definition 2.46 (Stable, Center, and Unstable Subspace, [100, Def. 18.13]):
Let $A \in \mathbb{R}^{n\times n}$ be and χ_A be its characteristic polynomial. Assume that χ_A is factored into a product of real monic polynomials $\chi_A = \chi_s\chi_c\chi_u$ such that the roots of χ_s, χ_c, and χ_u are subsets of $\mathbb{C}_-$, $\imath\mathbb{R}$, and $\mathbb{C}_+$, respectively. We define

- the *stable subspace* $E^s(A) := \ker(\chi_s(A)) \subseteq \mathbb{R}^n$,
- the *center subspace* $E^c(A) := \ker(\chi_c(A)) \subseteq \mathbb{R}^n$, and
- the *unstable subspace* $E^u(A) := \ker(\chi_u(A)) \subseteq \mathbb{R}^n$. $\Diamond$

The stable, center, and unstable subspace of a matrix can be defined equivalently and more explicitly by its Jordan canonical form; cf. [24, Sec. 1.4.2], [104, Def. 2.4]. Usually, the more explicit description makes the proofs more involved.

Theorem 2.47 ([47, Sec. 13.2, Cor.II], [100, Thm. 18.14]):
It holds:

(i) The stable, center, and unstable subspace are real A-invariant subspaces.

(ii) $\mathbb{R}^n = E^s(A) \oplus E^c(A) \oplus E^u(A)$.

(iii) $E^s(A) = \operatorname{im}(\chi_c\chi_u(A))$, $E^c(A) = \operatorname{im}(\chi_s\chi_u(A))$, and $E^u(A) = \operatorname{im}(\chi_s\chi_c(A))$.

(iv) The orthogonal complements of the stable, center, and unstable subspace are given by

$$\begin{aligned} E^s(A)^\perp &= E^c\big(A^\mathsf{T}\big) \oplus E^u\big(A^\mathsf{T}\big), \\ E^c(A)^\perp &= E^s\big(A^\mathsf{T}\big) \oplus E^u\big(A^\mathsf{T}\big), \text{ and} \\ E^u(A)^\perp &= E^s\big(A^\mathsf{T}\big) \oplus E^c\big(A^\mathsf{T}\big). \end{aligned} \qquad \Diamond$$

CHAPTER 3

ALGEBRAIC LYAPUNOV AND SYLVESTER EQUATIONS

Contents

In this chapter, we consider the *algebraic Sylvester equation* (ASE)

$$AX - XB = C, \tag{3.1}$$

for square matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{m\times m}$, and a right-hand side $C \in \mathbb{R}^{n\times m}$. The associated bounded linear operator

$$\mathcal{S}X = AX - XB, \tag{3.2}$$

is called a *Sylvester operator*. The *algebraic Lyapunov equation* (ALE)

$$AX + XA^{\mathsf{T}} = C \tag{3.3}$$

is a special case of the ASE (3.1).

The ASE and ALE play a key role in control theory, model order reduction, and stability analysis for linear systems; cf. [7,26,42]. The ASE also arises in finite difference discretizations of partial differential equations and numerical approximation of matrix functions; cf. [36,41,98] and [56, Sec. 4.6].

We organize this chapter as follows. Section 3.1 reviews the existence and uniqueness of solutions of the ASE (3.1). In Section 3.2, we study the Sylvester operator $\mathcal{S}$ and its spectral decomposition. A particular choice of the inner product ensures that the Sylvester operator $\mathcal{S}$ is normal. The spectral decomposition of $\mathcal{S}$ then yields a solution formula (Section 3.3, Theorem 3.6). Further solution representations are discussed in Section 3.3. Section 3.4 recalls the *alternating directions implicit* (ADI) method. Section 3.5 summarizes results on the singular value decay of the solution of the ASE and ALE. These results motivate us to consider low-rank approximations of the solution in the large-scale setting. In Section 3.6, we consider the ALE with a negative semidefinite right-hand side

$$AX + XA^{\mathsf{T}} = -FF^{\mathsf{T}}. \tag{3.4}$$

Section 3.7 reviews the low-rank ADI method for the numerical approximation of the solution of the ALE (3.4).

3.1 Existence and Uniqueness of Solutions

In this section, we are concerned with the existence and uniqueness of solutions of the ASE (3.1)

$$AX - XB = C.$$

We provide a proof of Theorem 3.1 because the argumentation of [59, Lem. 2.4.4.0, Thm. 2.4.4.1] is incomplete.

Theorem 3.1 (Existence and Uniqueness, [59, Lem. 2.4.4.0, Thm. 2.4.4.1]): The ASE (3.1) has a unique solution $X \in \mathbb{R}^{n\times m}$ if and only $\Lambda(A) \cap \Lambda(B) = \varnothing$, that is, if and only if A and B have no eigenvalue in common. ♢

Proof. The unique solvability of the ASE (3.1) is equivalent to $\ker(\mathcal{S}) = \{0\}$. If $X \in \ker(\mathcal{S})$, then

$$AX = XB \text{ and } A^k X = XB^k$$

for all $k \in \mathbb{N}_0$. Consequently,

$$p(A)X = Xp(B)$$

for any polynomial p. Let

$$\chi_B(t) = \prod_{i=1}^{m}(t - \lambda_i)$$

be the characteristic polynomial of B. The Cayley–Hamilton theorem yields $\chi_B(B) = 0$. We obtain

$$0 = \chi_B(A)X - X\chi_B(B) = \chi_B(A)X.$$

If $\Lambda(A) \cap \Lambda(B) = \varnothing$, then each of the factors $A - \lambda_i I$ is nonsingular. Hence, $\chi_B(A)$ is nonsingular and $X = 0$. Therefore, $\Lambda(A) \cap \Lambda(B) = \varnothing$ implies $\ker(\mathcal{S}) = \{0\}$.

Conversely, let $\lambda \in \Lambda(A) \cap \Lambda(B)$ be a common eigenvalue of A and B and $x, y \neq 0$ eigenvectors such that

$$Ax = \lambda x \text{ and } B^{\mathsf{T}}y = \lambda y.$$

Then, we have

$$\mathcal{S}xy^{\mathsf{T}} = Axy^{\mathsf{T}} - xy^{\mathsf{T}}B = \lambda xy^{\mathsf{T}} - \lambda xy^{\mathsf{T}} = 0.$$

Therefore, $0 \neq xy^{\mathsf{T}} \in \ker(\mathcal{S})$.

Finally, if X satisfies the ASE, then so does the complex conjugated $\overline{X}$, because the matrices A and B are real. The unique solvability implies that X must be real. □

Alternatively, the equivalent representation of the ASE (3.1) as a linear equation system

$$\left(I_m \otimes A - B^{\mathsf{T}} \otimes I_n\right) \operatorname{vec}(X) = \operatorname{vec}(C)$$

can be used to prove Theorem 3.1; cf. Section 2.9 and [1, Thm. 1.1.3].

3.2 Spectral Decomposition of the Sylvester Operator

We consider the Sylvester operator (3.2)

$$\mathcal{S}\colon \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m},\ \mathcal{S}X = AX - XB.$$

The Sylvester operator $\mathcal{S}$ has been thoroughly studied in [28,68,69,111]. The eigenvalues and eigenvectors of $\mathcal{S}$ are related to those of A and B; cf. [1, Rem. 1.1.2.], [7, Sec. 6.1.1], and [67].

We show that the Sylvester operator $\mathcal{S}$ (3.2) is normal if A and B are diagonalizable and if a suitably chosen inner product on a Hilbert space is considered. The inner product depends on the spectral decomposition of A and B. The eigenspaces of $\mathcal{S}$ can be constructed from the eigenspaces of A and B. This approach leads to a spectral decomposition of $\mathcal{S}$ and enables us to derive a solution formula (Theorem 3.6). In particular, this solution formula resembles the results of [42, Sec. 4.1.1] and [61].

Lemma 3.2 (Spectrum of the Sylvester Operator, [111, Ch. V, Cor. 1.4]):
For the spectrum of the Sylvester operator, it holds $\Lambda(\mathcal{S}) = \Lambda(A) - \Lambda(B)$. ◊

Lemma 3.3 (Sylvester Operator and Commuting Operators, [123]):
The Sylvester operator $\mathcal{S}$ splits into a sum of bounded linear commuting operators $\mathcal{S} = \mathcal{L} + \mathcal{R}$, where $\mathcal{L}X = AX$ and $\mathcal{R}X = -XB$. ◊

Lemma 3.4 ([15, Lem. 2]):
Assume that $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ are diagonalizable, and let $A = UD_AU^{-1}$ and $B^{\mathsf{T}} = VD_{B^{\mathsf{T}}}V^{-1}$ be the spectral decompositions of A and B^{T}. Furthermore, let $U = \left[u_1, \ldots, u_n\right] \in \mathbb{C}^{n\times n}$ and $V = \left[v_1, \ldots, v_m\right] \in \mathbb{C}^{m\times m}$ be the columns of U and V. It holds:

(i) $\langle X, Y\rangle_{U,V} := \left\langle U^{-1}XV^{-\mathsf{T}}, U^{-1}YV^{-\mathsf{T}}\right\rangle_{\mathrm{F}}$ is an inner product on $\mathbb{C}^{n\times m}$.

(ii) $\left(u_i v_j^\mathsf{T}\right)_{\substack{i=1,\dots,n\\ j=1,\dots,m}}$ is an orthonormal basis of $\mathbb{C}^{n\times m}$ with respect to $\langle\cdot,\cdot\rangle_{U,V}$.

(iii) The adjoint operator $\mathcal{S}^* : \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m}$ with respect to $\langle\cdot,\cdot\rangle_{U,V}$ is given by $\mathcal{S}^* X = U\overline{D_A}U^{-1}X - XV^{-\mathsf{T}}D_{B^\mathsf{T}}V^\mathsf{T}$.

(iv) $\mathcal{S}$ is a normal operator with respect to $\langle\cdot,\cdot\rangle_{U,V}$. ◊

Proof.

(i) It is straightforward to verify that $\langle\cdot,\cdot\rangle_{U,V}$ is an inner product on $\mathbb{C}^{n\times m}$.

(ii) Because of the identity

$$\left\langle u_i v_j^\mathsf{T}, u_k v_l^\mathsf{T}\right\rangle_{U,V} = \left\langle U^{-1}u_i v_j^\mathsf{T} V^{-\mathsf{T}}, U^{-1}u_k v_l^\mathsf{T} V^{-\mathsf{T}}\right\rangle_\mathrm{F} = \left\langle e_i e_j^\mathsf{T}, e_k e_l^\mathsf{T}\right\rangle_\mathrm{F} = \delta_{ik}\delta_{jl},$$

the matrices $u_i v_j^\mathsf{T} \in \mathbb{C}^{n\times m}$ are orthogonal with respect to $\langle\cdot,\cdot\rangle_{U,V}$ and therefore linearly independent. Because of $\dim(\mathbb{C}^{n\times m}) = n\cdot m$, the tuple $(u_i v_j^\mathsf{T})_{\substack{i=1,\dots,n\\ j=1,\dots,m}}$ forms an orthonormal basis of $\mathbb{C}^{n\times m}$.

(iii), (iv) We make use of the splitting $\mathcal{S} = \mathcal{L} + \mathcal{R}$ given by Lemma 3.3. The adjoints of $\mathcal{L}$ and $\mathcal{R}$ are given by

$$\mathcal{L}^* X = U\overline{D_A}U^{-1}X \text{ and } \mathcal{R}^* X = -XV^{-\mathsf{T}}\overline{D_{B^\mathsf{T}}}V^\mathsf{T},$$

because

$$\begin{aligned}\langle \mathcal{L}X, Y\rangle_{U,V} &= \left\langle U^{-1}AXV^{-\mathsf{T}}, U^{-1}YV^{-\mathsf{T}}\right\rangle_\mathrm{F} = \left\langle D_A U^{-1}XV^{-\mathsf{T}}, U^{-1}YV^{-\mathsf{T}}\right\rangle_\mathrm{F}\\ &= \left\langle U^{-1}XV^{-\mathsf{T}}, U^{-1}U\overline{D_A}U^{-1}YV^{-\mathsf{T}}\right\rangle_\mathrm{F} = \langle X, \mathcal{L}^*Y\rangle_{U,V},\end{aligned}$$

and

$$\begin{aligned}\langle \mathcal{R}X, Y\rangle_{U,V} &= -\left\langle U^{-1}XBV^{-\mathsf{T}}, U^{-1}YV^{-\mathsf{T}}\right\rangle_\mathrm{F} = -\left\langle U^{-1}XV^{-\mathsf{T}}D_{B^\mathsf{T}}, U^{-1}YV^{-\mathsf{T}}\right\rangle_\mathrm{F}\\ &= -\left\langle U^{-1}XV^{-\mathsf{T}}, U^{-1}YV^{-\mathsf{T}}\overline{D_{B^\mathsf{T}}}V^\mathsf{T}V^{-\mathsf{T}}\right\rangle_\mathrm{F} = \langle X, \mathcal{R}^*Y\rangle_{U,V},\end{aligned}$$

for all $X, Y \in \mathbb{C}^{n\times m}$. We have

$$\mathcal{S}^* = (\mathcal{L} + \mathcal{R})^* = \mathcal{L}^* + \mathcal{R}^*.$$

The operators $\mathcal{L}$, $\mathcal{R}$, $\mathcal{L}^*$, and $\mathcal{R}^*$ pairwise commute. Consequently, $\mathcal{S}$ and $\mathcal{S}^*$ commute as well. Therefore, $\mathcal{S}$ is normal with respect to $\langle\cdot,\cdot\rangle_{U,V}$. □

We formulate the Spectral Decomposition Theorem for the Sylvester operator $\mathcal{S}$. Utilizing Lemma 3.4, the proof basically requires standard arguments; cf. [72, Ch. 3.1].

Theorem 3.5 (Spectral Decomposition, [15, Lem. 3]):
Let the assumptions of Lemma 3.4 hold. Moreover, let $\alpha_1, \dots, \alpha_n \in \mathbb{C}$ and $\beta_1, \dots, \beta_m \in \mathbb{C}$ be the diagonal entries of D_A and D_{B^T}, respectively. Then we have:

(i) The matrix $u_i v_j^{\mathsf{T}} \in \mathbb{C}^{n\times m}$ is an eigenvector of $\mathcal{S}$ corresponding to the eigenvalue $\alpha_i - \beta_j$ for all $i = 1, \dots, n$ and $j = 1, \dots, m$.

(ii) The operator $\mathcal{P}_{i,j} \colon \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m}$, $\mathcal{P}_{i,j} X = \big\langle X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V}\, u_i v_j^{\mathsf{T}}$ is the orthogonal projection onto $\operatorname{span}\big\{u_i v_j^{\mathsf{T}}\big\} \subseteq \mathbb{C}^{n\times m}$ with respect to $\langle \cdot, \cdot \rangle_{U,V}$ for all $i = 1, \dots, n$ and $j = 1, \dots, m$.

(iii) The projectors are pairwise orthogonal, i.e., $\mathcal{P}_{i,j}\mathcal{P}_{k,l} = 0$ for all $i, k = 1, \dots, n$ and $j, l = 1, \dots, m$ such that $(i, j) \neq (k, l)$.

(iv) The projectors form a partition of the identity, i.e. $\operatorname{id}_{\mathbb{C}^{n\times m}} = \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{m} \mathcal{P}_{i,j}$.

(v) The Sylvester operator can be decomposed into $\mathcal{S} = \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{m} \big(\alpha_i - \beta_j\big)\mathcal{P}_{i,j}$ and

$$\mathcal{S}(X) = U\Bigg(\big(\alpha_i - \beta_j\big)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot U^{-1} X V^{-\mathsf{T}}\Bigg) V^{\mathsf{T}}. \qquad \lozenge$$

Proof.

(i) From $AU = UD_A$ and $B^{\mathsf{T}}V = VD_{B^{\mathsf{T}}}$, we deduce

$$\mathcal{S} u_i v_j^{\mathsf{T}} = A u_i v_j^{\mathsf{T}} - u_i v_j^{\mathsf{T}} B = \alpha_i u_i v_j^{\mathsf{H}} - \beta_j u_i v_j^{\mathsf{T}} = \big(\alpha_i - \beta_j\big) u_i v_j^{\mathsf{T}}.$$

(ii) $\mathcal{P}_{i,j}$ is a projection onto $\operatorname{span}\big\{u_i v_j^{\mathsf{T}}\big\}$, because Lemma 3.4 (ii) implies $\mathcal{P}_{i,j}^2 = \mathcal{P}_{i,j}$ and

$$\mathcal{P}_{i,j} X = \big\langle X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V}\, u_i v_j^{\mathsf{T}} \in \operatorname{span}\big\{u_i v_j^{\mathsf{T}}\big\}$$

for all $X \in \mathbb{C}^{n\times m}$. Moreover,

$$\begin{aligned} \big\langle \mathcal{P}_{i,j} X - X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V} &= \big\langle \mathcal{P}_{i,j} X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V} - \big\langle X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V} \\ &= \big\langle X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V} \big\langle u_i v_j^{\mathsf{T}}, u_i v_j^{\mathsf{T}} \big\rangle_{U,V} - \big\langle X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V} \\ &= 0, \end{aligned}$$

which means that $\mathcal{P}_{i,j}$ is the orthogonal projection onto $\operatorname{span}\big\{u_i v_j^{\mathsf{T}}\big\} \subseteq \mathbb{C}^{n\times m}$.

(iii) The assertion follows from Lemma 3.4 (ii).

(iv) According to Lemma 3.4 (ii), each $X \in \mathbb{C}^{n\times m}$ can be represented with respect to the orthonormal basis $\big(u_i v_j^{\mathsf{T}}\big)_{\substack{i=1,\dots,n\\ j=1,\dots,m}}$. Therefore,

$$X = \sum_{i=1}^{n} \sum_{j=1}^{m} \big\langle X, u_i v_j^{\mathsf{T}} \big\rangle_{U,V}\, u_i v_j^{\mathsf{T}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathcal{P}_{ij} X, \tag{3.5}$$

for all $X \in \mathbb{C}^{n\times m}$.

(v) We represent X as in Equation (3.5) and get

$$\begin{aligned}
\mathcal{S}X &= \mathcal{S}\sum_{i=1}^{n}\sum_{j=1}^{m}\langle X, u_i v_j^{\mathsf{T}}\rangle_{U,V}\, u_i v_j^{\mathsf{T}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\langle X, u_i v_j^{\mathsf{T}}\rangle_{U,V}\, \mathcal{S}u_i v_j^{\mathsf{T}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha_i - \beta_j)\,\langle X, u_i v_j^{\mathsf{T}}\rangle_{U,V}\, u_i v_j^{\mathsf{T}} && (3.6)\\
&= U\Big((\alpha_i - \beta_j)\,\langle X, u_i v_j^{\mathsf{T}}\rangle_{U,V}\Big)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} V^{\mathsf{T}} && (3.7)\\
&= U\Big((\alpha_i - \beta_j)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot \big(\langle U^{-1}XV^{-\mathsf{T}}, e_i e_j^{\mathsf{H}}\rangle_{\mathrm{F}}\big)_{\substack{i=1,\dots,n\\ j=1,\dots,m}}\Big) V^{\mathsf{T}} && (3.8)\\
&= U\Big((\alpha_i - \beta_j)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot U^{-1}XV^{-\mathsf{T}}\Big) V^{\mathsf{T}}. && (3.9)
\end{aligned}$$

□

3.3 Solution Representations

In this section, we consider explicit solution formulas for the ASE (3.1)

$$AX - XB = C.$$

First, we use Theorem 3.5 and obtain a solution formula based on the spectral decomposition of the Sylvester operator $\mathcal{S}$ (Theorem 3.6). After that, we review a well-known integral based solution formula.

Theorem 3.6 (Solution via Spectral Decomposition, [15, Lem. 3]):
Let the assumptions of Theorem 3.5 hold. Moreover, assume that $\Lambda(A) \cap \Lambda(B) = \varnothing$. Then the unique solution of the ASE (3.1) is given by

$$X = U\left(\left(\frac{1}{\alpha_i - \beta_j}\right)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot U^{-1}CV^{-\mathsf{T}}\right)V^{\mathsf{T}}. \tag{3.10}$$

◊

Proof. We apply Theorem 3.5 and show that the inverse of the Sylvester operator $\mathcal{S}$ is given by

$$\mathcal{S}^{-1} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\alpha_i - \beta_j}\mathcal{P}_{i,j}.$$

It holds that

$$\mathcal{S}\left(\sum_{k=1}^{n}\sum_{l=1}^{m}\frac{1}{\alpha_k-\beta_l}\mathcal{P}_{k,l}\right)\overset{\text{(v)}}{=}\left(\sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha_i-\beta_j)\mathcal{P}_{i,j}\right)\left(\sum_{k=1}^{n}\sum_{l=1}^{m}\frac{1}{\alpha_k-\beta_l}\mathcal{P}_{k,l}\right)$$
$$=\sum_{i,k=1}^{n}\sum_{j,l=1}^{m}\frac{\alpha_i-\beta_j}{\alpha_k-\beta_l}\mathcal{P}_{i,j}\mathcal{P}_{k,l}$$
$$\overset{\text{(ii), (iii)}}{=}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathcal{P}_{i,j}\overset{\text{(iv)}}{=}\mathrm{id}_{\mathbb{C}^{n\times m}}\,.$$

Therefore,

$$X=\mathcal{S}^{-1}C=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\alpha_i-\beta_j}\mathcal{P}_{i,j}C=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\alpha_i-\beta_j}\left\langle C,u_iv_j^{\mathsf{T}}\right\rangle_{U,V}u_iv_j^{\mathsf{T}}.$$

The claim follows using similar algebraic manipulations as in Theorem 3.5 Equations (3.6)–(3.9). □

The explicit solution representation of Theorem 3.7 is important, hence we provide the proof.

Theorem 3.7 (Integral Representation, [23, Thm. VII.2.3]):
Assume that A and $-B$ are stable. Then the unique solution of the ASE (3.1) is given by

$$X=-\int_{0}^{\infty}e^{tA}Ce^{-tB}\,\mathrm{d}t. \tag{3.11}$$

◊

Proof. The convergence of the integral (3.11) can be established by means of Theorem 2.24. We have

$$AX-XB=-\int_{0}^{\infty}Ae^{tA}Ce^{-tB}-e^{tA}Ce^{-tB}B\,\mathrm{d}t=-e^{tA}Ce^{-tB}\Big|_{0}^{\infty}=C.$$

□

If the spectra $\Lambda(A)$ and $\Lambda(B)$ are separated by a line, Theorem 3.7 may also be applied after a proper rotation. Theorem 3.7 also holds under weaker assumptions on $\Lambda(A)$ and $\Lambda(B)$; cf. [78, Thm. 5].

Other solution representations based on a series or a contour integral are presented in [23, Thm. VII.2.2, VII.2.4–VII.2.7]. In [38], annihilating polynomials are used to derive a solution representation. A very detailed overview of solution formulas is given in [78].

3.4 Alternating Direction Implicit Method

In preparation for Sections 3.5 and 3.7, we review the ADI method for the numerical approximation of the solution of the ASE (3.1)

$$AX - XB = C.$$

The ADI method is a splitting method and originates from to the iterative numerical solution of linear equation systems arising from the finite difference discretization of partial differential equations; cf. [92]. In this setting, the linear systems to be solved are of the form $(L + R)x = b$, and the matrices L and R commute. The ASE (3.1) is a linear equation defined by two commuting linear transformations: $\mathcal{S} = \mathcal{L} + \mathcal{R}$; cf. Lemma 3.3. Therefore, the ASE is also known as a *model problem* for the ADI method; cf. [123] and [124, Sec. 3].

The solutions of the linear systems

$$\left.\begin{array}{rcl} (A - q_k I)X_{k-\frac{1}{2}} & = & X_{k-1}(B - q_k I) + C, \\ X_k(B - p_k I) & = & (A - p_k I)X_{k-\frac{1}{2}} - C, \end{array}\right\} \tag{3.12}$$

define the iterates of the ADI method. The *shift parameters* p_k and q_k are assumed to be no eigenvalue of A and B, respectively. Their choice is crucial for the convergence of the method. We reformulate the two-step iteration (3.12) to an one-step iteration and represent the error explicitly. We use the explicit representation of the error in Section 3.5, therefore, we provide the proof.

Lemma 3.8 (ADI One-Step Iteration and Error Representation, e.g. [119]): Let X be any solution of the ASE (3.1), and let $p_k \in \mathbb{C} \setminus \Lambda(A)$ and $q_k \in \mathbb{C} \setminus \Lambda(B)$ be shift parameters. The two-step iteration (3.12) is equivalent to the one-step iteration

$$\begin{aligned} X_k = {} & (A - p_k I)(A - q_k I)^{-1} X_{k-1} (B - p_k I)^{-1} (B - q_k I) \\ & + (q_k - p_k)(A - q_k I)^{-1} C (B - p_k I)^{-1}. \end{aligned} \tag{3.13}$$

The error fulfills the recurrence relation

$$X - X_k = (A - p_k I)(A - q_k I)^{-1}(X - X_{k-1})(B - p_k I)^{-1}(B - q_k I). \tag{3.14}$$

◇

Proof. We multiply Equation (3.12) with $(A - q_k I)^{-1}$ and $(B - p_k I)^{-1}$ from the left and right, respectively. We obtain

$$\begin{aligned} X_{k-\frac{1}{2}} &= (A - q_k I)^{-1} X_{k-1}(B - q_k I) + (A - q_k I)^{-1} C, \\ X_k &= (A - p_k I) X_{k-\frac{1}{2}} (B - p_k I)^{-1} - C(B - p_k I)^{-1}. \end{aligned}$$

We merge the latter two equations and obtain

$$\begin{aligned}X_k &= (A - p_k I)(A - q_k I)^{-1} X_{k-1} (B - p_k I)^{-1} (B - q_k I) \\ &\quad + \big((A - p_k I)(A - q_k I)^{-1} - I\big) C (B - p_k I)^{-1}.\end{aligned}$$

The identity

$$(A - p_k I)(A - q_k I)^{-1} = I + (q_k - p_k)(A - q_k I)^{-1}$$

leads to the one-step iteration (3.13). Finally, if X is a solution of the ASE (3.1), then

$$\begin{aligned}(A - q_k I)X &= X(B - q_k I) + C, \\ X(B - p_k I) &= (A - p_k I)X - C.\end{aligned}$$

By using the same algebraic manipulations, it follows that X satisfies (3.13), and we obtain the recurrence relation (3.14) □

3.5 Bounds on the Singular Value Decay

If n and m are large, the numerical approximation of the solution $X \in \mathbb{R}^{n\times m}$ of the ASE (3.1)

$$AX - XB = C$$

comes with possibly high memory demands. Here, one may try to find a low-rank approximation $X \approx UV^{\mathsf{T}}$ where U and V are tall-and-skinny matrices. The absolute and relative error of the best possible approximation of rank at most k is given by

$$\sigma_{k+1}(X) = \inf_{\substack{Y\in\mathbb{C}^{n\times m}\\ \operatorname{rank}(Y)\le k}} \|X - Y\|_2 \text{ and } \frac{\sigma_{k+1}(X)}{\sigma_1(X)} = \inf_{\substack{Y\in\mathbb{C}^{n\times m}\\ \operatorname{rank}(Y)\le k}} \frac{\|X - Y\|_2}{\|X\|_2},$$

respectively; cf. Theorem 2.15. Therefore, a quick singular value decay of the solution X ensures that the solution can be well approximated by low-rank matrices.

A very detailed overview of singular value bounds can be found in [42, Sec. 2.2, Sec. 3.2, Sec. 4.2] and [43, 77, 88]. These bounds usually require the right-hand side C to be of full rank.

Mostly, in the large-scale setting, the right-hand side C is given in low-rank form $C = C_1 C_2^{\mathsf{T}}$, where C_1 and C_2 are tall-and-skinny matrices. Hence, the rank of C is much smaller than n and m. Therefore, the derived bounds often may not become applicable.

If the rank of the right-hand side C is low, a strategy to obtain bounds of the singular value decay can be summarized as follows. Let X_k be an approximation of rank at most k to the solution X of the ASE (3.1). The approximation error must be larger or equal to the error of the best possible approximation of rank at most k. Theorem 2.15 implies that

$$\sigma_{k+1}(X) = \inf_{\substack{Y\in\mathbb{C}^{n\times m}\\ \operatorname{rank}(Y)\le k}} \|X - Y\|_2 \le \|X - X_k\|_2 .$$

Finally, any bound on the absolute error $\|X - X_k\|_2$ bounds the singular value $\sigma_{k+1}(X)$.

In this context, approximations based on the application of quadrature rules to the solution representation Equation (3.11) of Theorem 3.7 were studied; cf. [17, 44–46, 71]. Quadrature rule based techniques also apply in an infinite-dimensional setting; cf. [48, 91, 114].

Singular value decay bounds based on the ADI method were presented in [12, 13, 18, 41, 94, 103, 108, 116, 119, 120]. We want to sketch the techniques and summarize at the end of this section known results. We utilize the ADI method (Lemma 3.8) and consider the rational function

$$\Phi(z) = \prod_{i=1}^{k} \frac{z - p_i}{z - q_i},$$

where $p_1, \ldots, p_k \in \mathbb{C} \setminus \Lambda(B)$ and $q_1, \ldots, q_k \in \mathbb{C} \setminus \Lambda(A)$. Lemma 3.8 yields

$$\|X - X_k\|_2 = \left\|\Phi(A)(X - X_0)\Phi(B)^{-1}\right\|_2 \leq \|\Phi(A)\|_2 \left\|\Phi(B)^{-1}\right\|_2 \|X - X_0\|_2 .$$

For simplicity, we assume A and B to be normal and get

$$\|\Phi(A)\|_2 = \max_{\lambda \in \Lambda(A)} |\Phi(\lambda)| \quad \text{and} \quad \left\|\Phi(B)^{-1}\right\|_2 = \max_{\lambda \in \Lambda(B)} \left|\Phi(\lambda)^{-1}\right| .$$

The shift parameters were chosen arbitrarily, thus

$$\|X - X_k\|_2 \leq \inf_{\substack{p_1,\ldots,p_k \in \mathbb{C}\setminus\Lambda(B) \\ q_1,\ldots,q_k \in \mathbb{C}\setminus\Lambda(A)}} \max_{\lambda \in \Lambda(A)} \prod_{i=1}^{k} \left| \frac{\lambda - p_i}{\lambda - q_i} \right| \max_{\lambda \in \Lambda(B)} \prod_{i=1}^{k} \left| \frac{\lambda - q_i}{\lambda - p_i} \right| \|X - X_0\|_2 .$$

We choose $X_0 = 0$ and examine the one-step iteration (3.13). If $\operatorname{rank}(C) \leq c$, then $\operatorname{rank}(X_k) \leq kc$. Theorem 2.15 yields

$$\sigma_{kc+1}(X) \leq \inf_{\substack{p_1,\ldots,p_k \in \mathbb{C}\setminus\Lambda(B) \\ q_1,\ldots,q_k \in \mathbb{C}\setminus\Lambda(A)}} \max_{\lambda \in \Lambda(A)} \prod_{i=1}^{k} \left| \frac{\lambda - p_i}{\lambda - q_i} \right| \max_{\lambda \in \Lambda(B)} \prod_{i=1}^{k} \left| \frac{\lambda - q_i}{\lambda - p_i} \right| \sigma_1(X), \tag{3.15}$$

for all $k \geq 1$ such that $0 \leq kc < \min\{n, m\}$. If the infimum of (3.15) is strictly smaller than 1, then the obtained bound is nontrivial.

3.5.1 Algebraic Lyapunov Equation and Singular Value Decay

Penzl considered the ALE (3.3) with a symmetric coefficient matrix A. A specific choice of shift parameters leads to a bound depending on the condition number of A.

Theorem 3.9 (Symmetric Coefficient Matrix, [94, Thm. 1]):
Assume that $A \in \mathbb{R}^{n \times n}$ is symmetric and stable, and $\operatorname{rank}(C) \leq c$. Let X be the unique solution of the ALE (3.3). Then

$$\sigma_{kc+1}(X) \leq \left(\prod_{i=1}^{k} \frac{\kappa_2(A)^{\frac{2i-1}{2k}} - 1}{\kappa_2(A)^{\frac{2i-1}{2k}} + 1} \right)^2 \sigma_1(X),$$

for all $k \geq 1$ such that $0 \leq kc < n$. ◊

Theorem 3.9 was initially formulated for symmetric negative semidefinite matrix C. The proof under the more general assumptions is analogous; cf. [94, Proof of Thm. 1].

Sorensen and Zhou considered the ALE (3.3) with diagonalizable coefficient matrix.

Theorem 3.10 (Diagonalizable Coefficient Matrix, [108, Thm. 2.1]): Assume that $A \in \mathbb{R}^{n\times n}$ is diagonalizable and stable, and $\operatorname{rank}(C) \leq c$. Moreover, suppose that the columns of $V \in \mathbb{C}^{n\times n}$ form a system of linearly independent right eigenvectors of A. Let X be the unique solution of the ALE (3.3). Then

$$\sigma_{kc+1}(X) \leq \kappa_2(V)^2 \inf_{q_1,\ldots,q_k\in\mathbb{C}\setminus\Lambda(A)} \max_{\lambda\in\Lambda(A)} \prod_{i=1}^{k} \left|\frac{\lambda+\overline{q_i}}{\lambda-q_i}\right|^2 \sigma_1(X), \tag{3.16}$$

for all $k \geq 1$ such that $0 \leq kc < n$. ◊

For any $\lambda \in \mathbb{C}_-$ and $q \in \mathbb{C}_+$ it holds that

$$\left|\frac{\lambda+\overline{q}}{\lambda-q}\right| < 1.$$

Therefore, if the condition number of V is not too large, the bound (3.16) is nontrivial. If A is unitarily diagonalizable (normal), then V can be chosen unitarily with $\kappa_2(V) = 1$.

3.5.2 Algebraic Sylvester Equation and Singular Value Decay

Similar to the one-step iteration (3.13), Beckermann and Townsend utilize rational functions and get a bound depending on the *Zolotarev number*; cf. [13, Thm. 2.1]. The Zolotarev number is related to an extremal problem for rational functions on sets E and F in the complex plane. If the sets E and F have a simple geometric shape, e.g., an interval or a ball, the Zolotarev number is explicitly known; cf. [2, §50, §51], [110].

Theorem 3.11 ([13, Thm. 2.1, Sec. 3.1–3.2], [41, Thm. 2.1]): Assume that $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ are symmetric, and $\operatorname{rank}(C) \leq c$. Moreover, let $\alpha_1, \alpha_2, \beta_1$ and β_2 be real numbers such that $\alpha_1 < \alpha_2$, $\beta_1 < \beta_2$, $[\alpha_1,\alpha_2] \cap [\beta_1,\beta_2] = \varnothing$, $\Lambda(A) \subseteq [\alpha_1,\alpha_2]$, and $\Lambda(B) \subseteq [\beta_1,\beta_2]$ hold. Let X be the unique solution of the ASE (3.1), and let $\gamma = \frac{|\beta_1-\alpha_1||\beta_2-\alpha_2|}{|\beta_1-\alpha_2||\beta_2-\alpha_1|}$. Then

$$\sigma_{kc+1}(X) \leq 4\exp\left(-\frac{k\pi^2}{\ln(16\gamma)}\right)\sigma_1(X),$$

for all $k \geq 1$ such that $0 \leq kc < \min\{n, m\}$. ◊

Theorem 3.12 ([13, Thm. 2.1, Sec. 3.3], [110, p. 123], [116, Thm. 1]): Assume that $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ are normal, and $\operatorname{rank}(C) \leq c$. Moreover, let $E = \left\{z \in \mathbb{C} \;\middle|\; \left|z - \frac{b+a}{2}\right| \leq \frac{b-a}{2}\right\}$ be a closed ball such that $0 < a < b$, $\Lambda(A) \subseteq -E$, and $\Lambda(B) \subseteq E$ hold. Let X be the unique solution of the ASE (3.1). Then

$$\sigma_{kc+1}(X) \leq \left(\frac{1-\sqrt{a/b}}{1+\sqrt{a/b}}\right)^{2k} \sigma_1(X),$$

for all $k \geq 1$ such that $0 \leq kc < \min\{n, m\}$. ◊

The influence of the deviation of the coefficient matrices A and B from normality is discussed in [9,103] and [13, Cor. 2.2]. The study of Cauchy and Krylov matrices related to the ASE (3.1) yield a bound on the singular value decay as well; cf. [8,89].

3.6 Algebraic Lyapunov Equation

In this section, we consider the ALE (3.4)

$$AX + XA^{\mathsf{T}} = -FF^{\mathsf{T}}.$$

Most importantly, if A is stable, the unique solution of the ALE (3.4) is *symmetric positive semidefinite* (spsd) and its image or equivalently kernel can be explicitly characterized. The proof in [26, Sec. 13, Thm. 3] focuses on the differential Lyapunov equation. For reasons of clarity, we provide the proof.

Theorem 3.13 (Image of the Solution, [26, Sec. 13, Thm. 3]):
Assume that $A \in \mathbb{R}^{n\times n}$ is stable, then the unique solution X of the ALE (3.4) is spsd and

$$\operatorname{im}(X) = \operatorname{im}(\mathcal{K}(A, F)). \qquad \diamondsuit$$

Proof.
X is symmetric:
According to Theorem 3.1, the ALE (3.4) admits a unique solution X. The right-hand side $-FF^{\mathsf{T}}$ is symmetric. Hence, X^{T} satisfies the equation as well, and the unique solvability implies the symmetry of X.

X is spsd:
Utilizing Theorem 3.6, we have

$$x^{\mathsf{T}}Xx = \int\limits_0^\infty x^{\mathsf{T}}e^{tA}FF^{\mathsf{T}}e^{tA^{\mathsf{T}}}x \,\mathrm{d}t = \int\limits_0^\infty \left\|F^{\mathsf{T}}e^{tA^{\mathsf{T}}}x\right\|_2^2 \mathrm{d}t \geq 0, \qquad (3.17)$$

for all $x \in \mathbb{R}^n$. Hence, X is spsd.

$\ker(X) \subseteq \ker\bigl(\mathcal{K}(A, F)^{\mathsf{T}}\bigr)$:
If $x \in \ker(X)$, then the integrand of Equation (3.17) must vanish. Therefore, we have

$$F^{\mathsf{T}}e^{tA^{\mathsf{T}}}x = 0$$

for all $t \geq 0$. We differentiate the latter equation with respect to t, evaluate at $t = 0$ and obtain

$$F^{\mathsf{T}}{A^{\mathsf{T}}}^k x = 0$$

for any power $k \in \mathbb{N}_0$. Therefore, $x \in \ker\big(\mathcal{K}(A, F)^\mathsf{T}\big)$.

$\underline{\ker\big(\mathcal{K}(A, F)^\mathsf{T}\big) \subseteq \ker(X)}$:
Conversely, if $x \in \ker\big(\mathcal{K}(A, F)^\mathsf{T}\big)$, then $F^\mathsf{T} {A^\mathsf{T}}^k x = 0$ for any power $k \in \mathbb{N}_0$. We have

$$F^\mathsf{T} e^{tA^\mathsf{T}} x = \sum_{k=0}^{\infty} \tfrac{t^k}{k!} F^\mathsf{T} {A^\mathsf{T}}^k x = 0,$$

and Equation (3.17) yields $x \in \ker(X)$.

$\underline{\operatorname{im}(X) = \operatorname{im}(\mathcal{K}(A, F))}$:
Finally, we consider the orthogonal complements of the spaces and obtain

$$\operatorname{im}(X) = \ker(X)^\perp = \ker\big(\mathcal{K}(A, F)^\mathsf{T}\big)^\perp = \operatorname{im}(\mathcal{K}(A, F)). \qquad \square$$

In [38, Thm. 2] the image and the kernel of the solution of the ASE (3.1) has been studied.

3.7 Low-rank Alternating Direction Implicit Method

As in the previous Section 3.6, we consider the ALE (3.4)

$$AX + XA^\mathsf{T} = -FF^\mathsf{T}.$$

We assume that F has only a few columns, and focus on the large-scale case (n is large). Theorem 3.13 yields that the unique solution X of the ALE (3.4) is spsd. The results on the singular value decay of Section 3.5 motivate us to consider a low-rank approach

$$X \approx ZZ^\mathsf{T}$$

for the numerical solution, where the low-rank factor Z has only a moderate number of columns. Here, the ADI method as a one-step iteration (3.13) can be reformulated in a low-rank fashion suitable for the large-scale ALE (3.4); cf. [20, 74, 82, 93]. Low-rank ADI methods for the ASE (3.1) have also been studied; cf. [21,74]. Efficient residual approximation and reduced complex arithmetic have led to considerable computational improvements; cf. [74]. A variant of the low-rank ADI method for the generalized ALE

$$AXM^\mathsf{T} + MXA^\mathsf{T} = -FF^\mathsf{T} \tag{3.18}$$

is given in Algorithm 3.1. Algorithm 3.1 Line 3 requires a shift parameter α in each iteration. There are various approaches and heuristics to this in the literature; cf. [103, Sec. 3.2] and [124, Ch. 1, Ch. 4]. We present a heuristic approach given in [19, Sec. 4.5.1] and [76, Sec. 2.1.3].

Algorithm 3.1: Low-rank ADI Method for the ALE (3.18) [74, Alg. 4.3].

```
Input: matrices $A, M \in \mathbb{R}^{n \times n}$, $F \in \mathbb{R}^{n \times f}$ defining Equation (3.18), and a
       tolerance $0 < \varepsilon_{\mathtt{rel}} \ll 1$ for the relative residual
Assumptions: $M$ is nonsingular, $M^{-1}A$ is stable, and $f \ll n$
Output: real matrix $Z$ such that $ZZ^{\mathsf{T}} \approx X$, the absolute and relative residual
        $r_{\mathtt{abs}}$ and $r_{\mathtt{rel}}$

% initialization:
$W := F$;   $Z := [\,]$;   $r_{\mathtt{abs}} := \|W\|_2^2$;   $\gamma_F := r_{\mathtt{abs}}$;
% iterate until relative residual is smaller than $\varepsilon_{\mathtt{rel}}$:
while $r_{\mathtt{abs}} \geq \varepsilon_{\mathtt{rel}} \gamma_F$ do
    obtain new shift parameter $\alpha$;
    $V := (A + \alpha M)^{-1} W$;
    if $\Im(\alpha) = 0$ then
        $W := W - 2\Re(\alpha) M V$;
        $Z := \left[Z, \sqrt{-2\alpha} V\right]$;
    else
        $\gamma := 2\sqrt{-\Re(\alpha)}$;   $\delta := \Re(\alpha)/\Im(\alpha)$;
        $\widehat{V} := \gamma(\Re(V) + \delta\Im(V))$;
        $W := W + \gamma M \widehat{V}$;
        $Z := \left[Z, \widehat{V}, \gamma\sqrt{1+\delta^2}\Im(V)\right]$;
    % update absolute residual:
    $r_{\mathtt{abs}} := \|W\|_2^2$;
% set relative residual:
$r_{\mathtt{rel}} := r_{\mathtt{abs}} / \gamma_F$;
return $Z$, $r_{\mathtt{abs}}$, $r_{\mathtt{rel}}$;
```

Heuristic 3.2: Shift Parameter Heuristic for Algorithm 3.1 line 3; [19, Sec. 4.5.1], [76, Sec. 2.1.3].

```
Input: matrices $A, M \in \mathbb{R}^{n \times n}, W \in \mathbb{R}^{n \times f}, Z \in \mathbb{R}^{n \times z}$ as in Algorithm 3.1 line 3
       and a number $l \in \mathbb{N}$
Output: shift parameter $\alpha \in \mathbb{C}_-$
if $Z$ is empty then
    $l := f$;   $\widehat{Z} := W$;
else
    $l := \min\{l, z\}$;
    set $\widehat{Z}$ to the last $l$ columns of $Z$;
compute a reduced QR decomposition of $\hat{Z}$:
  $QR := \widehat{Z}$;
$A_Q := Q^\mathsf{T} A Q$;   $M_Q := Q^\mathsf{T} M Q$;   $R_Q := Q^\mathsf{T} W$;
$H := \begin{bmatrix} A_Q^\mathsf{T} & 0 \\ R_Q R_Q^\mathsf{T} & -A_Q \end{bmatrix}$;   $E := \begin{bmatrix} M_Q^\mathsf{T} & 0 \\ 0 & M_Q \end{bmatrix}$;
compute the generalized eigenvalues and eigenvectors of $(H, E)$:
  $HV = EVD$ with $V = \begin{bmatrix} v_1, \dots, v_{2l} \\ w_1, \dots, w_{2l} \end{bmatrix}$ and $D = \operatorname{diag}(\alpha_1, \dots, \alpha_{2l})$;
$\alpha := -1$;   $\gamma := 0$;
for $i = 1, \dots, 2l$ do
    if $\Re(\alpha_i) < 0$ then
        $\tau := \|w_i\|_2^2 \,/\, \left|w_i^\mathsf{H} M_Q v_i\right|$;
        if $\tau \geq \gamma$ then
            $\gamma := \tau$;   $\alpha := \alpha_i$;
return $\alpha$;
```

CHAPTER 4

DIFFERENTIAL LYAPUNOV AND SYLVESTER EQUATIONS

Contents

In this chapter, we consider the *differential Sylvester equation* (DSE)

$$\dot{X}(t) = AX(t) - X(t)B + C, \tag{4.1a}$$

$$X(0) = X_0, \tag{4.1b}$$

for square matrices $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{m\times m}$, a right-hand side $C \in \mathbb{R}^{n\times m}$, and an initial value $X_0 \in \mathbb{R}^{n\times m}$. The *differential Lyapunov equation* (DLE)

$$\dot{X}(t) = AX(t) + X(t)A^{\mathsf{T}} + FF^{\mathsf{T}}, \tag{4.2a}$$

$$X(0) = X_0, \tag{4.2b}$$

where $F \in \mathbb{R}^{n\times f}$, is a special case of the ASE (3.3). Stationary points of the DSE and DLE are solutions of the corresponding algebraic equations presented in Chapter 3.

Similar to the ASE and ALE, the DSE and DLE find applications in control theory, model order reduction; cf. [7, 26, 42].

We organize this chapter as follows. In Section 4.1, we consider the more general nonautonomous version of the DSE and present a theorem about the existence and uniqueness of solutions. We also present some solution formulas for the DSE (4.1). In Section 4.2, we focus on the DLE (4.2) and develop a Galerkin approach for the numerical approximation of the solution. Finally, Section 4.3 reviews a low-rank version of the *backward differentiation formula* (BDF) method for the DLE (4.2).

4.1 Differential Sylvester Equation

In Section 4.1.1, we review the existence and uniqueness of solutions of the DSE (4.1). Section 4.1.2 summarizes some explicit solution formulas.

4.1.1 Existence and Uniqueness of Solutions

Here, we recall a well-known result on the existence and uniqueness of solutions, which applies to the more general nonautonomous DSE. The proof provided by [1, Thm. 1.1.5] is relatively concise. Therefore, we give a more detailed version of the proof.

Theorem 4.1 (Existence and Uniqueness, [1, Thm. 1.1.5]):
Let $\mathbb{I} \subseteq \mathbb{R}$ be an open interval with $t_0 \in \mathbb{I}$, $A \colon \mathbb{I} \to \mathbb{R}^{n\times n}$, $B \colon \mathbb{I} \to \mathbb{R}^{m\times m}$, $C \colon \mathbb{I} \to \mathbb{R}^{n\times m}$ continuous matrix-valued functions, and $X_0 \in \mathbb{R}^{n\times m}$. The nonautonomous DSE

$$\dot{X}(t) = A(t)X(t) - X(t)B(t) + C(t), \tag{4.3a}$$

$$X(t_0) = X_0, \tag{4.3b}$$

has a unique solution $X \colon \mathbb{I} \to \mathbb{R}^{n\times m}$ given by

$$X(t) = \Phi(t)X_0\Psi(t)^{\mathsf{T}} + \int\limits_{t_0}^{t} \Phi(t)\Phi(s)^{-1}C(s)\Psi(s)^{-\mathsf{T}}\Psi(t)^{\mathsf{T}}\,\mathrm{d}s, \tag{4.4}$$

where $\Phi(t)$ and $\Psi(t)$ are the fundamental matrices of the systems

$$\dot{x}(t) = A(t)x(t) \text{ and } \dot{x}(t) = -B(t)^{\mathsf{T}}x(t),$$

respectively, such that $\Phi(t_0) = I$ and $\Psi(t_0) = I$. ◇

Proof.
Existence:
Theorem 2.31 yields that $\Phi(t)$ and $\Psi(t)$ are nonsingular for all $t \in \mathbb{I}$. Utilizing $\Phi(t_0) = I$, $\Psi(t_0) = I$, and Equation (4.4), it follows that $X(t_0) = X_0$. By differentiating Equation (4.4), it can be verified that (4.4) defines a solution of the nonautonomous DSE (4.3) on the interval $\mathbb{I}$.

Uniqueness:
If $\tilde{X}\colon \mathbb{J} \to \mathbb{R}^{n\times m}$ solves the nonautonomous DSE (4.3) such that $\mathbb{J} \subseteq \mathbb{I}$ is an open interval with $t_0 \in \mathbb{J}$, then the difference $E(t) := X(t) - \tilde{X}(t)$ satisfies

$$\begin{aligned}\dot{E}(t) &= A(t)E(t) - E(t)B(t),\\ E(t_0) &= 0.\end{aligned}$$

We observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\Phi(t)^{-1}E(t)\Psi(t)^{-\mathsf{T}}\right) = 0.$$

Therefore, $\Phi(t)^{-1}E(t)\Psi(t)^{-\mathsf{T}}$ is constant on $\mathbb{J}$. Because of $E(t_0) = 0$, we have

$$\Phi(t)^{-1}E(t)\Psi(t)^{-\mathsf{T}} = 0$$

for all $t \in \mathbb{J}$. Consequently, $X(t) = \tilde{X}(t)$ for all $t \in \mathbb{J}$. □

Alternatively, we can utilize the equivalent representation

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}\operatorname{vec}(X(t)) &= \left(I_n \otimes A(t) - B(t)^{\mathsf{T}} \otimes I_m\right)\operatorname{vec}(X(t)) + \operatorname{vec}(C(t)),\\ \operatorname{vec}(X(t_0)) &= \operatorname{vec}(X_0),\end{aligned}$$

to ensure the existence and uniqueness of solutions; cf. Theorem 2.32 and Section 2.9.

4.1.2 Solution Representations

In this section, we focus on the autonomous DSE (4.1)

$$\begin{aligned}\dot{X}(t) &= AX(t) - X(t)B + C,\\ X(0) &= X_0\end{aligned}$$

and explicit solution representations. Using the Sylvester operator (3.2), the DSE (4.1) rewrites as

$$\begin{aligned}\dot{X}(t) &= \mathcal{S}X(t) + C,\\ X(0) &= X_0.\end{aligned}$$

4.1.2.1 Exponential of the Sylvester Operator

We consider the exponential of the Sylvester operator $e^{t\mathcal{S}}$. We recall that $\mathcal{S} = \mathcal{L} + \mathcal{R}$ and $\mathcal{L}$ and $\mathcal{R}$ commute; cf. Lemma 3.3.

Lemma 4.2 (Exponential of the Sylvester Operator $\mathcal{S}$):
For the Sylvester operator $\mathcal{S}$ and its partial realizations $\mathcal{L}X = AX$ and $\mathcal{R}X = -XB$, it holds that:

$$e^{t\mathcal{S}} = e^{t\mathcal{L}}e^{t\mathcal{R}} = e^{t\mathcal{L}}e^{t\mathcal{R}} \tag{4.5}$$

for all $t \in \mathbb{R}$. ◊

Proof. See Lemma 3.3 and Section 2.5. □

Next, we give an explicit formula for the action of the exponential of the Sylvester operator in terms of the spectral decomposition of A and B.

Theorem 4.3 (Exponential of $\mathcal{S}$ and Spectral Decomposition, [15, Lem. 3]): Assume that $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ are diagonalizable. Let $A = UD_AU^{-1}$ and $B^\mathsf{T} = VD_{B^\mathsf{T}}V^{-1}$ be the spectral decompositions of A and B^T, $\alpha_1, \ldots, \alpha_m \in \mathbb{C}$ and $\beta_1, \ldots, \beta_n \in \mathbb{C}$ be the diagonal entries of D_A and D_{B^T}, then

$$e^{t\mathcal{S}}X = U\left(\left(e^{t(\alpha_i-\beta_j)}\right)_{\substack{i=1,\ldots,n\\ j=1,\ldots,m}} \odot U^{-1}XV^{-\mathsf{T}}\right)V^\mathsf{T}, \tag{4.6}$$

for all $t \in \mathbb{R}$. ◊

Proof. Using Theorem 3.5, we can decompose $\mathcal{S}$ into a linear combination of commuting projectors

$$\mathcal{S} = \sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha_i - \beta_j)\mathcal{P}_{i,j}.$$

We have $\mathcal{P}_{i,j}\mathcal{P}_{k,l} = 0$ for all $(i,j) \neq (k,l)$, and $\mathcal{P}_{i,j}^2 = \mathcal{P}_{i,j}$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Therefore,

$$\mathcal{S}^k = \left(\sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha_i - \beta_j)\mathcal{P}_{i,j}\right)^k = \sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha_i - \beta_j)^k\mathcal{P}_{i,j}$$

for all $k \in \mathbb{N}$. Hence,

$$\begin{aligned} e^{t\mathcal{S}}X &= \sum_{i=1}^{n}\sum_{j=1}^{m} e^{t\left(\alpha_i-\beta_j\right)}\mathcal{P}_{i,j}X = \sum_{i=1}^{n}\sum_{j=1}^{m} e^{t\left(\alpha_i-\beta_j\right)}\mathcal{P}_{i,j}X \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} e^{t\left(\alpha_i-\beta_j\right)}\left\langle X, u_iv_j^\mathsf{T}\right\rangle_{U,V} u_iv_j^\mathsf{T}. \end{aligned}$$

Finally, Equation (4.6) follows by using similar algebraic manipulations as in Theorem 3.5 Equations (3.6)–(3.9). □

4.1.2.2 Variation of Constants Formula

The specification of Equation (4.4) to the autonomous case with constant coefficients is straightforward by simply replacing the state transition matrix with the matrix exponential; cf. Theorems 2.32 and 2.33. Utilizing Lemma 4.2, we obtain the solution formula

$$\begin{aligned} X(t) &= e^{tA}X_0e^{-tB} + \int_0^t e^{(t-s)A}Ce^{-(t-s)B}\,\mathrm{d}s = e^{t\mathcal{L}}e^{t\mathcal{R}}X_0 + \int_0^t e^{(t-s)\mathcal{L}}e^{(t-s)\mathcal{R}}C\,\mathrm{d}s \\ &= e^{t\mathcal{S}}X_0 + \int_0^t e^{(t-s)\mathcal{S}}C\,\mathrm{d}s. \end{aligned}$$

Theorem 4.4 (Variation of Constants Formula, [1, Cor. 1.1.6]):
The DSE (4.1) has a unique solution $X\colon \mathbb{R} \to \mathbb{R}^{m\times n}$ given by

$$X(t) = e^{tA}X_0e^{-tB} + \int\limits_0^t e^{(t-s)A}Ce^{-(t-s)B}\,\mathrm{d}s. \tag{4.7}$$

◊

We combine Theorem 4.3 and Equation (4.7), and we find that, under the assumptions of Theorem 4.3, the solution of the DSE has the form

$$\begin{aligned} X(t) &= e^{t\mathcal{S}}X_0 + \int\limits_0^t e^{(t-s)\mathcal{S}}C\,\mathrm{d}s \\ &= U\Big(\left(e^{t(\alpha_i-\beta_j)}\right)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot U^{-1}X_0V^{-\mathsf{T}}\Big)V^{\mathsf{T}} \\ &\quad + \int\limits_0^t U\Big(\left(e^{(t-s)(\alpha_i-\beta_j)}\right)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot U^{-1}CV^{-\mathsf{T}}\Big)V^{\mathsf{T}}\,\mathrm{d}s \\ &= U\Big(\left(e^{t(\alpha_i-\beta_j)}\right)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot U^{-1}X_0V^{-\mathsf{T}}\Big)V^{\mathsf{T}} \\ &\quad + U\left(\left(\int\limits_0^t e^{(t-s)(\alpha_i-\beta_j)}\,\mathrm{d}s\right)_{\substack{i=1,\dots,n\\ j=1,\dots,m}} \odot U^{-1}CV^{-\mathsf{T}}\right)V^{\mathsf{T}}, \end{aligned} \tag{4.8}$$

with the involved scalar integrals given explicitly as:

$$\int\limits_0^t e^{(t-s)(\alpha_i+\beta_j)}\,\mathrm{d}s = \begin{cases} \frac{e^{t(\alpha_i+\beta_j)}-1}{\alpha_i+\beta_j} & \text{if } \alpha_i+\beta_j \neq 0, \\ t & \text{if } \alpha_i+\beta_j = 0 \end{cases}.$$

Formula (4.8) resembles the results presented in [99]. The variation of constants formula (Theorem 4.4) leads to another solution formula based on the ASE's unique solution.

Theorem 4.5 ([26, Sec. 13]):
Assume $\Lambda(A) \cap \Lambda(B) = \varnothing$, then the unique solution of the DSE has the form

$$X(t) = e^{t\mathcal{S}}X_0 - \mathcal{S}^{-1}C + e^{t\mathcal{S}}\mathcal{S}^{-1}C. \tag{4.9}$$

◊

Proof. Because of $\Lambda(A) \cap \Lambda(B) = \varnothing$ and Theorem 3.1, the inverse $\mathcal{S}^{-1}$ exists, and we can rewrite the solution formula (4.7) as

$$\begin{aligned} X(t) &= e^{t\mathcal{S}}X_0 + \int\limits_0^t e^{(t-s)\mathcal{S}}C\,\mathrm{d}s = e^{t\mathcal{S}}X_0 + \left(-\mathcal{S}^{-1}e^{(t-s)\mathcal{S}}C\Big|_0^t\right) \\ &= e^{t\mathcal{S}}X_0 - \mathcal{S}^{-1}(C) + \mathcal{S}^{-1}e^{t\mathcal{S}}C, \end{aligned}$$

and confirm that $X(0) = X_0$ holds. □

4.2 Galerkin Approach for the Differential Lyapunov Equation

In this section, we develop a Galerkin approach for the numerical approximation of the solution of the DLE (4.2)

$$\dot{X}(t) = AX(t) + X(t)A^{\mathsf{T}} + FF^{\mathsf{T}}, \tag{4.10a}$$
$$X(0) = 0, \tag{4.10b}$$

with zero initial conditions. We consider the large-scale case, that is, the state-space dimension n is large, and F has only a few columns.

4.2.1 Solution Formula and Matrix Exponential

The solution representation of Theorem 4.5 serves as a motivation for our approach. If A is stable, then

$$X(t) = X_{\infty} - e^{tA}X_{\infty}e^{tA^{\mathsf{T}}} \tag{4.11}$$

represents the solution of the DLE (4.10), where $X_{\infty} \in \mathbb{R}^{n\times n}$ is the unique spsd solution of the ALE

$$AX_{\infty} + X_{\infty}A^{\mathsf{T}} = -FF^{\mathsf{T}}. \tag{4.12}$$

4.2.2 Approximation of the Matrix Exponential Action and Spectral Decomposition

Here, we focus on the action of the matrix exponential on the solution of the ALE (4.12)

$$e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}.$$

We consider a spectral decomposition of X_{∞} of the form

$$X_{\infty} = Q_{\infty}D_{\infty}Q_{\infty}^{\mathsf{T}}, \tag{4.13a}$$
$$Q_{\infty} = \left[q_1, \ldots, q_n\right] \in \mathbb{R}^{n\times n}, \tag{4.13b}$$
$$D_{\infty} = \operatorname{diag}(\lambda_1^{\downarrow}(X_{\infty}), \ldots, \lambda_n^{\downarrow}(X_{\infty})) \in \mathbb{R}^{n\times n}, \tag{4.13c}$$

where $q_1, \ldots, q_n \in \mathbb{R}^n$ is a system of orthonormal eigenvectors corresponding to the nonnegative eigenvalues $\lambda_1^{\downarrow}(X_{\infty}), \ldots, \lambda_n^{\downarrow}(X_{\infty}) \in \mathbb{R}$. For sufficiently quick eigenvalue decay, the solution of the ALE (4.12) can be well approximated using the eigenvectors corresponding to the k-largest eigenvalues, i.e.,

$$X_{\infty} \approx Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}X_{\infty}Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}} \text{ and } \left\|X_{\infty} - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}X_{\infty}Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right\|_2 = \lambda_{k+1}^{\downarrow}(X_{\infty}),$$

where $Q_{\infty,k} := \left[q_1, \ldots, q_k\right] \in \mathbb{R}^{n\times k}$; cf. Theorem 2.15 and Section 3.5. This motivates us to study a similar approximation of the time-dependent part of the solution representation (4.11) given by

$$e^{tA} X_\infty e^{tA^\mathsf{T}} \approx Q_{\infty,k} Q_{\infty,k}^\mathsf{T} e^{tA} X_\infty e^{tA^\mathsf{T}} Q_{\infty,k} Q_{\infty,k}^\mathsf{T}.$$

Furthermore, we consider the linear space

$$\mathcal{Q}_k := \left\{ Q_{\infty,k} Y Q_{\infty,k}^\mathsf{T} \mid Y \in \mathbb{R}^{k\times k} \right\}$$

together with its orthogonal projection with respect to the Frobenius inner product

$$\mathcal{P}_k \colon \mathbb{R}^{n\times n} \to \mathcal{Q}_k,\ \mathcal{P}_k X = Q_{\infty,k} Q_{\infty,k}^\mathsf{T} X Q_{\infty,k} Q_{\infty,k}^\mathsf{T}.$$

Next, we give a bound on the projection error.

Lemma 4.6 (Approximation of the Matrix Exponential Action): Assume that A is stable and let X_∞ be the unique spsd solution of the ALE (4.12). Then for all $k = 1, \ldots, n-1$ and all $t \geq 0$, the approximation error is bounded by

$$\left\| e^{tA} X_\infty e^{tA^\mathsf{T}} - \mathcal{P}_k e^{tA} X_\infty e^{tA^\mathsf{T}} \right\|_2 \leq 2\sqrt{\lambda_{k+1}^{\downarrow}(X_\infty)\lambda_1^{\downarrow}(X_\infty)}. \tag{4.14}$$

◊

Proof. As X_∞ is spsd, we have $0 \preccurlyeq e^{tA} X_\infty e^{tA^\mathsf{T}}$. According to Theorem 3.7, the solution X_∞ can be represented as

$$X_\infty = \int\limits_0^\infty e^{sA} F F^\mathsf{T} e^{sA^\mathsf{T}} \,\mathrm{d}s.$$

With that, we get

$$\begin{aligned} 0 \preccurlyeq e^{tA} X_\infty e^{tA^\mathsf{T}} &= \int\limits_0^\infty e^{(s+t)A} F F^\mathsf{T} e^{(s+t)A^\mathsf{T}} \,\mathrm{d}s = \int\limits_t^\infty e^{sA} F F^\mathsf{T} e^{sA^\mathsf{T}} \,\mathrm{d}s \\ &\preccurlyeq \int\limits_0^\infty e^{sA} F F^\mathsf{T} e^{sA^\mathsf{T}} \,\mathrm{d}s = X_\infty. \end{aligned}$$

Theorem 2.4 (ii) yields

$$\begin{aligned} 0 \preccurlyeq e^{tA} X_\infty e^{tA^\mathsf{T}} &\preccurlyeq X_\infty = \mathcal{P}_k X_\infty + X_\infty - \mathcal{P}_k X_\infty \\ &\preccurlyeq \mathcal{P}_k X_\infty + \left\| X_\infty - \mathcal{P}_k X_\infty \right\|_2 I = \mathcal{P}_k X_\infty + \lambda_{k+1}^{\downarrow}(X_\infty) I \\ &= Q_{\infty,k} Q_{\infty,k}^\mathsf{T} X_\infty Q_{\infty,k} Q_{\infty,k}^\mathsf{T} + \lambda_{k+1}^{\downarrow}(X_\infty) I. \end{aligned}$$

We apply the projection matrix $I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T}$ and get

$$0 \preccurlyeq \left(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T}\right) e^{tA} X_\infty e^{tA^\mathsf{T}} \left(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T}\right) \preccurlyeq \lambda_{k+1}^{\downarrow}(X_\infty)\left(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T}\right),$$

where we have used that $\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)Q_{\infty,k}$ vanishes. Next, Theorem 2.4 (v) yields

$$\begin{aligned}\left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)e^{tA}X_{\infty}^{1/2}\right\|_2^2 &= \left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)\right\|_2 \\ &\leq \left\|\lambda_{k+1}^{\downarrow}(X_{\infty})\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)\right\|_2 \\ &= \lambda_{k+1}^{\downarrow}(X_{\infty}).\end{aligned}$$

Finally,

$$\begin{aligned}\left\|e^{tA}X_{\infty}e^{tA^{\mathsf{T}}} - \mathcal{P}_k e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}\right\|_2 &= \left\|e^{tA}X_{\infty}e^{tA^{\mathsf{T}}} - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right\|_2 \\ &= \left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}\right\|_2 + \\ &\quad \left\|Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)\right\|_2 \\ &\leq 2\left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}\right\|_2 \\ &\leq 2\left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)e^{tA}X_{\infty}^{1/2}\right\|_2\left\|X_{\infty}^{1/2}e^{tA^{\mathsf{T}}}\right\|_2 \\ &= 2\left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)e^{tA}X_{\infty}^{1/2}\right\|_2\sqrt{\left\|e^{tA}X_{\infty}e^{tA^{\mathsf{T}}}\right\|_2} \\ &\leq 2\sqrt{\lambda_{k+1}^{\downarrow}(X_{\infty})\left\|X_{\infty}\right\|_2} = 2\sqrt{\lambda_{k+1}^{\downarrow}(X_{\infty})\lambda_1^{\downarrow}(X_{\infty})}. \qquad \square\end{aligned}$$

If the eigenvalues of the solution X_{∞} decay quick enough, Bound (4.14) ensures that the time-dependent part can be well-approximated in a low-rank fashion using a system of orthonormal eigenvectors corresponding to the k-largest eigenvalues.

4.2.3 Approximation of the Matrix Exponential Action and Approximate Solution

We consider a low-rank approximation ZZ^{T} of the solution of the ALE (4.12), i.e.,

$$AZZ^{\mathsf{T}} + ZZ^{\mathsf{T}}A^{\mathsf{T}} + FF^{\mathsf{T}} = R,$$

where R is the residual. If $Z = QSV^{\mathsf{T}}$ is an compact SVD of the low-rank factor Z, then the columns of the matrix Q form a system of orthonormal eigenvectors of the approximation ZZ^{T}. We replace in Equation (4.11) the exact solution X_{∞} by its low-rank approximation ZZ^{T} and get

$$X(t) \approx ZZ^{\mathsf{T}} - e^{tA}ZZ^{\mathsf{T}}e^{tA^{\mathsf{T}}}.$$

We focus on the action of the matrix exponential $e^{tA}Z$ and its approximation by projection $QQ^{\mathsf{T}}e^{tA}Z$.

Lemma 4.7 (Approximation of the Matrix Exponential Action):
Assume that $\mu_2[A] < 0$. Let ZZ^T be an approximate solution of the ALE (4.12) with residual R and $Z = QSV^\mathsf{T}$ be its compact SVD. Then for all $t \geq 0$, the projection error is bounded by

$$\left\| e^{tA}Z - QQ^\mathsf{T}e^{tA}Z \right\|_2 \leq \sqrt{\|R\|_2 \int\limits_0^t e^{2s\mu_2[A]}\,\mathrm{d}s}. \tag{4.15}$$

◊

Proof. The proof is similar to that of Lemma 4.6. Theorem 3.7 yields

$$ZZ^\mathsf{T} = \int\limits_0^\infty e^{sA}\left(FF^\mathsf{T} - R\right)e^{sA^\mathsf{T}}\,\mathrm{d}s.$$

Theorem 2.4 (ii) and Theorem 2.26 (vi) yield

$$\begin{aligned}
0 &\preccurlyeq e^{tA}ZZ^\mathsf{T}e^{tA^\mathsf{T}} = \int\limits_t^\infty e^{sA}\left(FF^\mathsf{T} - R\right)e^{sA^\mathsf{T}}\,\mathrm{d}s \preccurlyeq \int\limits_0^\infty e^{sA}FF^\mathsf{T}e^{sA^\mathsf{T}}\,\mathrm{d}s - \int\limits_t^\infty e^{sA}Re^{sA^\mathsf{T}}\,\mathrm{d}s \\
&= \int\limits_0^\infty e^{sA}\left(FF^\mathsf{T} - R\right)e^{sA^\mathsf{T}}\,\mathrm{d}s + \int\limits_0^t e^{sA}Re^{sA^\mathsf{T}}\,\mathrm{d}s \preccurlyeq ZZ^\mathsf{T} + \int\limits_0^t e^{sA}Re^{sA^\mathsf{T}}\,\mathrm{d}s \\
&\preccurlyeq ZZ^\mathsf{T} + \left\| \int\limits_0^t e^{sA}Re^{sA^\mathsf{T}}\,\mathrm{d}s \right\|_2 I \preccurlyeq ZZ^\mathsf{T} + \|R\|_2 \int\limits_0^t e^{2s\mu_2[A]}\,\mathrm{d}sI.
\end{aligned}$$

We apply the projection matrix $I - QQ^\mathsf{T}$ to the latter inequality from the left and right and get

$$0 \preccurlyeq \left(I - QQ^\mathsf{T}\right)e^{tA}ZZ^\mathsf{T}e^{tA^\mathsf{T}}\left(I - QQ^\mathsf{T}\right) \preccurlyeq \|R\|_2 \int\limits_0^t e^{2s\mu_2[A]}\,\mathrm{d}s\left(I - QQ^\mathsf{T}\right).$$

Finally, Theorem 2.4 (v) yields

$$\begin{aligned}
\left\| e^{tA}Z - QQ^\mathsf{T}e^{tA}Z \right\|_2^2 &= \left\| \left(I - QQ^\mathsf{T}\right)e^{tA}Z \right\|_2^2 = \left\| \left(I - QQ^\mathsf{T}\right)e^{tA}ZZ^\mathsf{T}e^{tA^\mathsf{T}}\left(I - QQ^\mathsf{T}\right) \right\|_2 \\
&\leq \left\| \|R\|_2 \int\limits_0^t e^{2s\mu_2[A]}\,\mathrm{d}s\left(I - QQ^\mathsf{T}\right) \right\|_2 = \|R\|_2 \int\limits_0^t e^{2s\mu_2[A]}\,\mathrm{d}s.
\end{aligned}$$

□

If the logarithmic norm of A is nonnegative, then Lemma 4.7 holds as well, but the integral term dominates for large times. Alternatively, the bound on the matrix exponential of Theorem 2.24 can be applied.

4.2.4 Galerkin Approach

The bounds on the projection error (Lemmas 4.6 and 4.7) motivate the Galerkin approach $Q\tilde{X}(t)Q^{\mathsf{T}}$ for the approximation of the action of the matrix exponential

$$e^{tA}ZZ^{\mathsf{T}}e^{tA^{\mathsf{T}}},$$

where $Z = QSV^{\mathsf{T}}$ is the compact SVD of the low-rank factor Z. The action of the matrix exponential satisfies the DLE

$$\begin{aligned}\dot{X}(t) &= AX(t) + X(t)A^{\mathsf{T}},\\ X(0) &= ZZ^{\mathsf{T}}.\end{aligned}$$

Combining the Galerkin approach and the latter equation yields the defect equation

$$D(t) := Q\dot{\tilde{X}}(t)Q^{\mathsf{T}} - AQ\tilde{X}(t)Q^{\mathsf{T}} - Q\tilde{X}(t)Q^{\mathsf{T}}A^{\mathsf{T}}.$$

We impose the Galerkin condition

$$\left\langle D(t), QYQ^{\mathsf{T}}\right\rangle_{\mathrm{F}} = 0 \text{ for all } Y \in \mathbb{R}^{k\times k}$$

and get

$$\dot{\tilde{X}}(t) = Q^{\mathsf{T}}AQ\tilde{X}(t) + \tilde{X}(t)Q^{\mathsf{T}}A^{\mathsf{T}}Q.$$

If we require $\tilde{X}(0) = S^2$, then

$$Q\tilde{X}(0)Q^{\mathsf{T}} = QS^2Q^{\mathsf{T}} = QSV^{\mathsf{T}}VSQ^{\mathsf{T}} = ZZ^{\mathsf{T}} = e^{tA}ZZ^{\mathsf{T}}e^{tA^{\mathsf{T}}}\Big|_{t=0}.$$

Hence, the initial condition is satisfied exactly by the Galerkin approach. Furthermore, this gives

$$\tilde{X}(t) = e^{tQ^{\mathsf{T}}AQ}S^2e^{tQ^{\mathsf{T}}A^{\mathsf{T}}Q}$$

and the Galerkin approximation

$$Qe^{tQ^{\mathsf{T}}AQ}S^2e^{tQ^{\mathsf{T}}A^{\mathsf{T}}Q}Q^{\mathsf{T}} \approx e^{tA}ZZ^{\mathsf{T}}e^{tA^{\mathsf{T}}}.$$

The matrices Q and Z have the same number of columns. Hence, if the low-rank factor $Z \in \mathbb{R}^{n\times k}$ has only a few columns ($k \ll n$), then the matrix exponential $e^{tQ^{\mathsf{T}}AQ}$ of size $k \times k$ can be approximated, e.g., by the scaling and squaring method. Here, the main issue is that the matrix $Q^{\mathsf{T}}AQ$ might not be stable. If the logarithmic norm of A is negative, then $Q^{\mathsf{T}}AQ$ is stable; cf. Section 2.7. Nevertheless, generally, the Galerkin approach does not preserve stability.

Remark 4.8:
We note that similar arguments also hold for the generalized DLE

$$M\dot{X}(t)M^{\mathsf{T}} = AX(t)M^{\mathsf{T}} + MX(t)A^{\mathsf{T}} + FF^{\mathsf{T}}, \tag{4.17a}$$

$$X(0) = 0, \tag{4.17b}$$

with $M \in \mathbb{R}^{n\times n}$ nonsingular. ◊

In summary, the proposed approach reads as written down in Algorithm 4.1.

Algorithm 4.1: Galerkin Approach for the DLE (4.17) (`ALE-Galerkin-DLE`).

```
Input: $M \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times n}$, $F \in \mathbb{R}^{n\times f}$, truncation tolerance $tol_{\texttt{trunc}} > 0$, step
       size $h > 0$, final time $t_f > 0$.
Assumptions: $M^{-1}A$ is stable.
Output: $X(kh) \approx ZZ^{\mathsf{T}} - Q\tilde{X}_kQ^{\mathsf{T}}$ that approximates the solution to the
        generalized DLE.

  % solve the ALE:
1 $A^{\mathsf{T}}X_\infty M + M^{\mathsf{T}}X_\infty A + C^{\mathsf{T}}C = 0$ for $X_\infty \approx ZZ^{\mathsf{T}}$;
  % compute compact SVD and truncate:
2 use Algorithm B.1 with $(Z, tol_{\texttt{trunc}})$ as input and obtain the truncated
    compact SVD $(Z, Q, S)$ as output;
  % compute matrix:
3 $\tilde{A} := Q^{\mathsf{T}}M^{-\mathsf{T}}A^{\mathsf{T}}Q$;
  % compute matrix exponential:
4 $\Theta_h := \texttt{expm}(h\tilde{A})$;
  % iterate up to final time:
5 $k := 0$;   $z_0 := S$;   $\tilde{X}_0 := z_0z_0^{\mathsf{T}}$;
6 while $kh < t_f$ do
7     $z_{k+1} := \Theta_h z_k$;
8     $k := k + 1$;
9     $\tilde{X}_k := z_kz_k^{\mathsf{T}}$;
```

4.2.4.1 Numerical Results

To quantify and illustrate the performance of `ALE-Galerkin-DLE` (Algorithm 4.1), we consider DLEs arising from the `RAIL` benchmark problem (Table A.1). Namely, we consider the DLEs:

$$M\dot{X}(t)M^{\mathsf{T}} = AX(t)M^{\mathsf{T}} + MX(t)A^{\mathsf{T}} + BB^{\mathsf{T}}, \quad X(0) = 0 \tag{\texttt{RAIL\_N}}$$

and

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A + C^{\mathsf{T}}C, \quad X(0) = 0. \tag{\texttt{RAIL\_T}}$$

We carried out all computations on the Silver Node (Appendix A). `ALE-Galerkin-DLE` (Algorithm 4.1 line 1) requires a low-rank approximation ZZ^{T} of the solution of the corresponding ALE. For this task, we have used the low-rank ADI method Algorithm 3.1. We report the absolute and relative residuals

$$\left\|AZZ^{\mathsf{T}}M^{\mathsf{T}} + MZZ^{\mathsf{T}}A^{\mathsf{T}} + BB^{\mathsf{T}}\right\|_2 \text{ or } \left\|A^{\mathsf{T}}ZZ^{\mathsf{T}}M + M^{\mathsf{T}}ZZ^{\mathsf{T}}A + C^{\mathsf{T}}C\right\|_2$$

and

$$\frac{\|AZZ^{\mathsf{T}}M^{\mathsf{T}} + MZZ^{\mathsf{T}}A^{\mathsf{T}} + BB^{\mathsf{T}}\|_2}{\|BB^{\mathsf{T}}\|_2} \text{ or } \frac{\|A^{\mathsf{T}}ZZ^{\mathsf{T}}M + M^{\mathsf{T}}ZZ^{\mathsf{T}}A + C^{\mathsf{T}}C\|_2}{\|C^{\mathsf{T}}C\|_2}.$$

For the truncation tolerance $tol_{\mathtt{trunc}}$, we have used machine precision $\varepsilon_{\mathtt{mach}}$ and the coarser value $\sqrt{\varepsilon_{\mathtt{mach}}}$. The achieved values for the different test setups as well as the number of columns of the corresponding of the truncated low-rank factor Z are listed in Table 4.1. Table 4.2 reports the computational timings for the numerical solution of the ALE. The step sizes are given in Table 4.3.

Instance	$tol_{\mathtt{trunc}}$	Size of Galerkin system	Absolute residual	Relative residual
RAIL_N	$\sqrt{\varepsilon_{\mathtt{mach}}}$	214	$6.95 \cdot 10^{-27}$	$3.17 \cdot 10^{-13}$
	$\varepsilon_{\mathtt{mach}}$	419	$7.22 \cdot 10^{-27}$	$3.29 \cdot 10^{-13}$
RAIL_T	$\sqrt{\varepsilon_{\mathtt{mach}}}$	190	$3.26 \cdot 10^{-14}$	$2.72 \cdot 10^{-15}$
	$\varepsilon_{\mathtt{mach}}$	297	$3.15 \cdot 10^{-14}$	$2.62 \cdot 10^{-15}$

Tab. 4.1: Residuals for the Numerical Approximation of the Solution of the ALE.

Instance	Time to solve ALE (s)
RAIL_N	1.84
RAIL_T	0.75

Tab. 4.2: Time to solve the ALE.

Instance	Step sizes h
RAIL_N	$\{2^0, 2^{-1}, \dots, 2^{-8}\}$
RAIL_T	$\{2^0, 2^{-1}, \dots, 2^{-8}\}$

Tab. 4.3: Step Sizes h.

Comparison with BDF Methods

We compare the Galerkin approach ALE-Galerkin-DLE (Algorithm 4.1) with the low-rank BDF/ADI method of orders $1, 2, \dots, 6$ abbreviated by BDF1, BDF2, BDF3, BDF4, BDF5, and BDF6, respectively; cf. Section 4.3. Because of memory limitations, we consider the BDF/ADI methods to the DLEs RAIL_N and RAIL_T on the time interval $[0, 100]$. For the Galerkin approach ALE-Galerkin-DLE, we have used the larger interval $[0, 4512]$.

The computational timings for both methods are given in Figures D.1 and D.2. Although the time intervals for both setups are different, the Galerkin approach ALE-Galerkin-DLE outperforms the BDF/ADI methods in terms of computational time for both problem instances RAIL_N and RAIL_T.

For the accuracy evaluation, we consider the absolute and relative errors

$$\|X(t) - X_{\mathtt{ref}}(t)\|_2 \quad \text{and} \quad \frac{\|X(t) - X_{\mathtt{ref}}(t)\|_2}{\|X_{\mathtt{ref}}(t)\|_2}.$$

We utilized the spectral decomposition of (A, M) and solution representation (4.8) to construct the reference solution $X_{\mathtt{ref}}(t)$. In order to measure the approximation quality of the trial space of the `ALE-Galerkin-DLE` method, we report the absolute and relative error of the best approximation of the reference solution. More precisely, the best approximation is given by

$$X_{\mathtt{best}}(t) := QQ^{\mathsf{T}} X_{\mathtt{ref}}(t) QQ^{\mathsf{T}},$$

where Q is the matrix of (Algorithm 4.1 line 2). Indeed, $X_{\mathtt{best}}(t)$ is the orthogonal projection with respect to the Frobenius inner product of $X_{\mathtt{ref}}(t)$ onto the trial space. The numerical errors for the `ALE-Galerkin-DLE` method are presented in Figures D.3a–D.3d and D.6a–D.6d. We report the convergence to the stationary point and the norm of the reference solution in Figures D.4, D.5, D.7, and D.8. Figures D.9a–D.11f show the numerical errors for the `BDF/ADI` methods. Based on the numerical results, the following observations can be made. At first, the best approximation $X_{\mathtt{best}}(t)$ for the truncation tolerance $\varepsilon_{\mathtt{mach}}$ is very accurate; cf. Figures D.3d and D.6d. Hence, the choice of the trial space of the `ALE-Galerkin-DLE` method is justified. The reference solution of `RAIL_T` itself is large in norm what makes the absolute error comparatively large; cf. Figure D.7 with Figures D.6c and D.11e. Interestingly, the error of the `ALE-Galerkin-DLE` approximations increases with decreasing h; cf. Figures D.3a, D.3c, D.6a, and D.6c. A decrease of the step size results in an increased number of time steps. Hence, the rounding errors possibly accumulate during time stepping; cf. Algorithm 4.1 lines 5–9. The relative error of the `BDF/ADI` approximation and the `ALE-Galerkin-DLE` with $tol_{\mathtt{trunc}} = \sqrt{\varepsilon_{\mathtt{mach}}}$ approximation are nearly at the same level; cf. Figures D.3d and D.9f and Figures D.6d and D.11f. But the `ALE-Galerkin-DLE` ($tol_{\mathtt{trunc}} = \varepsilon_{\mathtt{mach}}$) approximation is more accurate than the `BDF/ADI` approximation.

4.3 Backward Differentiation Formula for the Differential Lyapunov Equation

This section follows [80]. We review BDFs for the DLE

$$\dot{X}(t) = AX(t) + X(t)A^{\mathsf{T}} + FF^{\mathsf{T}}, \tag{4.18a}$$
$$X(0) = Z_0 Z_0^{\mathsf{T}}, \tag{4.18b}$$

where $A \in \mathbb{R}^{n\times n}$, $F \in \mathbb{R}^{n\times f}$, and $Z_0 \in \mathbb{R}^{n\times z}$. We consider the large-scale case and assume that Z_0 and F have only a few columns and n is large. Let $h > 0$ be the step size. We define the step size $h = t_k - t_{k-1}$. The s-step BDF method applied to the DLE (4.18) is given by

$$X_k - \sum_{j=1}^{s} \alpha_j X_{k-j} = h\beta\big(AX_k + X_k A^{\mathsf{T}} + FF^{\mathsf{T}}\big),$$

where α_j and β are coefficients of the BDF method; cf. [51, Ch. III.1, Eqn. (1.22")].

s	β	α_1	α_2	α_3	α_4	α_5	α_6
1	1	1					
2	$\frac{2}{3}$	$\frac{4}{3}$	$-\frac{1}{3}$				
3	$\frac{6}{11}$	$\frac{18}{11}$	$-\frac{9}{11}$	$\frac{2}{11}$			
4	$\frac{12}{25}$	$\frac{48}{25}$	$-\frac{36}{25}$	$\frac{16}{25}$	$-\frac{3}{25}$		
5	$\frac{60}{137}$	$\frac{300}{137}$	$-\frac{300}{137}$	$\frac{200}{137}$	$-\frac{75}{137}$	$\frac{12}{137}$	
6	$\frac{60}{147}$	$\frac{120}{49}$	$-\frac{150}{49}$	$\frac{400}{147}$	$-\frac{75}{49}$	$\frac{24}{49}$	$-\frac{10}{147}$

Tab. 4.4: Coefficients of the s-step BDF Method.

The iterate X_k approximates $X(t_k)$. The parameter s is the order of the BDF method. We recall that for $s > 6$, the method is not numerically stable, and for $s = 1$, the BDF method coincides with the implicit Euler method. A minor rearrangement shows that the current iterate X_k can be obtained as the solution of the ALE

$$\left(h\beta A - \tfrac{1}{2}I\right)X_k + X_k\left(h\beta A - \tfrac{1}{2}I\right)^{\mathsf{T}} = -h\beta FF^{\mathsf{T}} - \sum_{j=1}^{s}\alpha_j X_{k-j}. \tag{4.19}$$

Since for $s \geq 2$, some coefficients α_j are negative, the ALE (4.19) has a symmetric but possibly indefinite right-hand side, making the standard low-rank ADI method infeasible; cf. Section 3.7. For this reason, a low-rank LDL^{T}-decomposition for the numerical approximation is proposed, and suitable modifications of the low-rank ADI method have been developed; [80]. Assume that $X_i \approx L_i D_i L_i^{\mathsf{T}}$ for $i = 0, \dots, k$, $L_i \in \mathbb{R}^{n \times l_i}$, $D_i \in \mathbb{R}^{l_i \times l_i}$ and $l_i \ll n$, then the right-hand side can be factored as

$$\begin{aligned} -h\beta FF^{\mathsf{T}} - \sum_{j=1}^{s}\alpha_j X_{k-j} &\approx -G_k S_k G_k^{\mathsf{T}}, \\ G_k &= \left[F, L_{k-1}, \dots, L_{k-s}\right], \\ S_k &= \begin{bmatrix} h\beta I & & & \\ & \alpha_1 D_{k-1} & & \\ & & \ddots & \\ & & & \alpha_s D_{k-s} \end{bmatrix}. \end{aligned}$$

The LDL^{T}-type low-rank ADI method can be used to approximate $X_k \approx L_k D_k L_k^{\mathsf{T}}$.

Remark 4.9:

The BDF method also extends to the generalized DLE

$$\begin{aligned} M\dot{X}(t)M^{\mathsf{T}} &= AX(t)M^{\mathsf{T}} + MX(t)A^{\mathsf{T}} + FF^{\mathsf{T}}, \\ X(0) &= Z_0 Z_0^{\mathsf{T}} \end{aligned}$$

and applies to the nonautonomous case as well. ◊

CHAPTER 5

ALGEBRAIC RICCATI EQUATION

Contents

We consider the *algebraic Riccati equation* (ARE)

$$A^{\mathsf{T}}X + XA - XBB^{\mathsf{T}}X + C^{\mathsf{T}}C = 0, \tag{5.1}$$

for matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times b}$, and $C \in \mathbb{R}^{c\times n}$.

The ARE (5.1) arises frequently in control and systems theory. In particular, the *linear-quadratic regulator* and the *feedback stabilization* problem of the linear-time invariant system

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t),\\ y(t) &= Cx(t)\end{aligned}$$

are important applications; cf. [1, Ch. 8], [84, Ch. 4], and [66, Ch. 9].

We organize this chapter as follows. In Section 5.1, we review the one-to-one correspondence between solutions of the ARE (5.1) and invariant subspaces of the associated real Hamiltonian matrix

$$\mathcal{H} = \begin{bmatrix} A & -BB^{\mathsf{T}} \\ -C^{\mathsf{T}}C & -A^{\mathsf{T}} \end{bmatrix}. \tag{5.3}$$

Usually, the ARE (5.1) admits multiple solutions. Therefore, we focus in Section 5.2 on the *stabilizing* solution, and review existence and uniqueness conditions. Furthermore, we review spsd solutions of the ARE (5.1) and study the image or equivalently the kernel of the stabilizing solution.

5.1 Solutions and Subspaces

First, we review the definition of a graph subspace. Next, we state the one-to-one correspondence between solution of the ARE (5.1) and invariant graph subspace of the Hamiltonian matrix $\mathcal{H}$ (5.3).

Definition 5.1 (Graph Subspace, [1, Def. 2.1.1]):
A linear subspace $S \subseteq \mathbb{C}^{2n}$ is called a *graph subspace* if there are matrices $U, V \in \mathbb{C}^{n\times n}$ with U nonsingular such that

$$S = \operatorname{im}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right). \qquad \lozenge$$

Clearly, a graph subspace is n-dimensional. Because of

$$\operatorname{im}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) = \operatorname{im}\left(\begin{bmatrix} U \\ V \end{bmatrix} U^{-1}\right) = \operatorname{im}\left(\begin{bmatrix} I \\ VU^{-1} \end{bmatrix}\right),$$

every graph subspace has the form

$$\mathcal{G}(X) := \operatorname{im}\left(\begin{bmatrix} I \\ X \end{bmatrix}\right) \tag{5.4}$$

for some matrix $X \in \mathbb{C}^{n\times n}$.

Theorem 5.2 (Solutions and Invariant Subspaces [1, Sec. 2.1, Thm. 2.1.2]):
The matrix $X \in \mathbb{C}^{n\times n}$ satisfies the ARE (5.1) if and only if $\mathcal{G}(X)$ is an $\mathcal{H}$-invariant subspace. $\lozenge$

Proof. If $\mathcal{G}(X)$ is an $\mathcal{H}$-invariant subspace, then there is a matrix $T \in \mathbb{C}^{n\times n}$ such that

$$\mathcal{H}\begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} T. \tag{5.5}$$

Combining the first and second block row of the latter equation gives

$$-C^{\mathsf{T}}C - A^{\mathsf{T}}X = XT = X\left(A - BB^{\mathsf{T}}X\right),$$

which is equivalent to the ARE (5.1).

On the other hand, if X satisfies the ARE (5.1), then it can be verified that Equation (5.5) holds with

$$T = A - BB^{\mathsf{T}}X.$$

Hence, $\mathcal{G}(X)$ is an $\mathcal{H}$-invariant subspace. $\square$

Theorem 5.2 shows that there is a one-to-one correspondence between the solutions of the ARE (5.1) and all invariant graph subspaces of the Hamiltonian matrix $\mathcal{H}$. In particular, if

$$\operatorname{im}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right)$$

is an n-dimensional $\mathcal{H}$-invariant graph subspace, then VU^{-1} defines a solution of the ARE (5.1). Furthermore, the solution depends only on the subspace rather than on the chosen basis. More specifically, assume that

$$\operatorname{im}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) = \operatorname{im}\left(\begin{bmatrix} \tilde{U} \\ \tilde{V} \end{bmatrix}\right)$$

for some matrices $\tilde{U}, \tilde{V} \in \mathbb{C}^{n \times n}$. The n-dimensional linear spaces are equal, hence, there is a nonsingular matrix T such that

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \tilde{U} \\ \tilde{V} \end{bmatrix} T.$$

Because of U and T are nonsingular, the matrix $\tilde{U}$ is nonsingular and

$$VU^{-1} = \left(\tilde{V}T\right)\left(\tilde{U}T\right)^{-1} = \tilde{V}\tilde{U}.$$

5.2 Stabilizing Solutions

If $X \in \mathbb{C}^{n \times n}$ is a solution of the ARE (5.1), then

$$\mathcal{H}\begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix}\left(A - BB^{\mathsf{T}}X\right);$$

cf. proof of Theorem 5.2. The ARE (5.1) may have several solutions or even uncountable many; cf. [1, Ex. 2.1.5]. Therefore, we require in addition that $A - BB^{\mathsf{T}}X$ is stable.

Definition 5.3 (Stabilizing Solution, [84, Sec. 5.3]):
A solution $X \in \mathbb{C}^{n \times n}$ of the ARE (5.1) is called *stabilizing* if $A - BB^{\mathsf{T}}X$ is stable. ◊

5.2.1 Uniqueness of Stabilizing Solutions

Here, we review the uniqueness and some properties of the stabilizing solution of the ARE (5.1). We provide a proof of Theorem 5.4 because the argumentation in [84, Thm. 5.5] is partly incomplete and slightly more technical.

Theorem 5.4 (Uniqueness of Stabilizing Solutions, [84, Thm. 5.5, Thm. 5.9]):
Assume that $X \in \mathbb{C}^{n \times n}$ is a stabilizing solution of the ARE (5.1). Then it follows that:

(i) X is symmetric.

(ii) X is unique.

(iii) X is real.

(iv) X is spsd.

(v) X is maximal: If $\tilde{X}$ is a real symmetric solution of the ARE (5.1), then $X \succcurlyeq \tilde{X}$. ◊

Proof. (i) If X satisfies the ARE (5.1), then X^T does too. With that, we subtract

$$\begin{aligned} A^\mathsf{T} X + XA - XBB^\mathsf{T} X + C^\mathsf{T} C &= 0, \\ A^\mathsf{T} X^\mathsf{T} + X^\mathsf{T} A - X^\mathsf{T} BB^\mathsf{T} X^\mathsf{T} + C^\mathsf{T} C &= 0 \end{aligned}$$

and obtain the ALE

$$\left(A - BB^\mathsf{T} X\right)^\mathsf{T} (X - X^\mathsf{T}) + (X - X^\mathsf{T})\left(A - BB^\mathsf{T} X\right) = 0.$$

Because $A - BB^\mathsf{T} X$ is stable, the equation is uniquely solvable. Therefore, X is symmetric.

(ii) Let $X_1, X_2 \in \mathbb{C}^{n \times n}$ be stabilizing solutions of the ARE (5.1). Again, we subtract

$$\begin{aligned} A^\mathsf{T} X_1 + X_1 A - X_1 BB^\mathsf{T} X_1 + C^\mathsf{T} C &= 0, \\ A^\mathsf{T} X_2 + X_2 A - X_2 BB^\mathsf{T} X_2 + C^\mathsf{T} C &= 0, \end{aligned}$$

use the symmetry of X_1 and X_2, and get the ASE

$$\left(A - BB^\mathsf{T} X_2\right)^\mathsf{T} (X_1 - X_2) + (X_1 - X_2)\left(A - BB^\mathsf{T} X_1\right) = 0.$$

As $A - BB^\mathsf{T} X_i$ is stable, it follows that $X_1 = X_2$.

(iii) If X is a stabilizing solution, then $\overline{X}$ is also a stabilizing solution. The uniqueness implies that X is real.

(iv) If X is a stabilizing solution, then X is symmetric and satisfies the ALE

$$\left(A - BB^\mathsf{T} X\right)^\mathsf{T} X + X\left(A - BB^\mathsf{T} X\right) = -XBB^\mathsf{T} X - C^\mathsf{T} C.$$

Theorem 3.13 yields that X is spsd.

(v) The difference of

$$\begin{aligned} A^\mathsf{T} X + XA - XBB^\mathsf{T} X + C^\mathsf{T} C &= 0, \\ A^\mathsf{T} \tilde{X} + \tilde{X} A - \tilde{X} BB^\mathsf{T} \tilde{X} + C^\mathsf{T} C &= 0 \end{aligned}$$

rewrites as an ALE

$$\left(A - BB^\mathsf{T} X\right)^\mathsf{T} (X - \tilde{X}) + (X - \tilde{X})\left(A - XBB^\mathsf{T} X\right) = -(\tilde{X} - X)BB^\mathsf{T}(\tilde{X} - X).$$

Theorem 3.13 gives $X \succcurlyeq \tilde{X}$. □

5.2.2 Existence of Stabilizing Solutions

We focus on the existence of the stabilizing solution of the ARE (5.1). First, we review the definition of the pair (A, B) to be stabilizable and give some equivalent characterizations. Then, we review an existence theorem for the stabilizing solution of the ARE (5.1).

Definition 5.5 (Stabilizable, [1, Def. A.1.4]):
Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times b}$. The pair (A, B) is called *stabilizable*, if there exists a matrix $K \in \mathbb{R}^{b\times n}$ such that $A - BK$ is stable. $\Diamond$

Theorem 5.6 (Stabilizable, [1, Thm. A.1.5], [100, Thm. 18.28, Cor. 18.29]):
Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times b}$. The following conditions are equivalent:

(i) (A, B) is stabilizable.

(ii) If x is a left-eigenvector of A corresponding to an eigenvalue λ with $\Re(\lambda) \geq 0$, then $x^{\mathsf{H}}B \neq 0$.

(iii) $\operatorname{rank}\big(\begin{bmatrix} A - \lambda I & B \end{bmatrix}\big) = n$ for all $\lambda \in \mathbb{C}$ with $\Re(\lambda) \geq 0$.

(iv) $E^{c}(A) \oplus E^{u}(A) \subseteq \operatorname{im}(\mathcal{K}(A, B))$.

(v) $E^{s}(A) + \operatorname{im}(\mathcal{K}(A, B)) = \mathbb{R}^{n}$. $\Diamond$

Theorem 5.7 (Existence of Stabilizing Solutions, [73, Thm. 3]):
The ARE (5.1) has a stabilizing solution if and only if the following conditions both hold:

(i) (A, B) is stabilizable.

(ii) No eigenvalue of the Hamiltonian matrix $\mathcal{H}$ has zero real part, i.e., $\Lambda(\mathcal{H}) \cap \imath\mathbb{R} = \varnothing$. $\Diamond$

5.2.3 Symmetric Positive Semidefinite Solutions

The stabilizing solution of the ARE (5.1) is spsd; cf. Theorem 5.4. In this subsection, we focus on the uniqueness of spsd solutions of the ARE (5.1). We introduce the notion of a detectable matrix pair and give equivalent conditions. Finally, we review an existence and uniqueness result for spsd solutions of the ARE (5.1).

Definition 5.8 (Detectable, [1, Def. A.1.5]):
Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{c\times n}$. The pair (A, C) is called *detectable*, if there exists a matrix $K \in \mathbb{R}^{n\times c}$ such that $A - KC$ is stable. $\Diamond$

Theorem 5.9 (Detectable, [1, Thm. A.1.6], [100, Thm. 18.31]):
Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{c\times n}$. The following conditions are equivalent:

(i) (A, C) is detectable.

(ii) $(A^\mathsf{T}, C^\mathsf{T})$ is stabilizable.

(iii) If x is an eigenvector of A corresponding to an eigenvalue λ with $\Re(\lambda) \geq 0$, then $Cx \neq 0$.

(iv) $\operatorname{rank}\left(\begin{bmatrix} A-\lambda I \\ C \end{bmatrix}\right) = n$ for all $\lambda \in \mathbb{C}$ with $\Re(\lambda) \geq 0$.

(v) $\ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big) \subseteq E^s(A)$.

(vi) $(E^c(A) \oplus E^u(A)) \cap \ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big) = \{0\}$. ◊

Theorem 5.10 (Uniqueness of spsd Solutions, [84, Thm. 5.8]):
Assume that (A, C) is detectable and that $X \in \mathbb{C}^{n\times n}$ is a hpsd solution of the ARE (5.1). Then X is a stabilizing solution. ◊

If the pair (A, C) is detectable, then Theorem 5.4 and Theorem 5.10 yield that the ARE (5.1) has at most one spsd solution $X \in \mathbb{R}^{n\times n}$. Next, we review an important Lemma about the Hamiltonian Matrix $\mathcal{H}$ to prepare for Theorem 5.12, the final result in this section.

Lemma 5.11 ([84, Lem. 5.2]):
Assume that (A, B) is stabilizable and (A, C) is detectable. Then $\Lambda(\mathcal{H}) \cap \imath\mathbb{R} = \varnothing$. ◊

Finally, Theorem 5.12 follows by Theorems 5.4, 5.7, and 5.10 and Lemma 5.11.

Theorem 5.12 ([84, Thm. 5.10]):
Assume that (A, B) is stabilizable and (A, C) is detectable, then the ARE (5.1) has a unique spsd solution $X \in \mathbb{R}^{n\times n}$. Moreover, the solution $X \in \mathbb{R}^{n\times n}$ is stabilizing. ◊

5.2.4 Image of the Solution

Here, we study the image or, equivalently, the kernel of the stabilizing solution of the ARE (5.1). First, there is a natural motivation to study the kernel/image of the stabilizing solution of the ARE (5.1), because the kernel/image is an important feature of a matrix. Second, knowledge about the image of the stabilizing solution of the ARE can be used to develop numerical methods for the solution approximation.

Theorem 5.13 is the main result in this section. Theorems 5.14 and 5.15 are just specializations of Theorem 5.13. The proof of Theorem 5.13 given by [86, Thm. 7] is relatively concise, hence we provide a more detailed proof.

Theorem 5.13 (Kernel/Image of the Stabilizing Solution I, [86, Thm. 7]): Assume that (A, B) is stabilizable and that $\Lambda(\mathcal{H}) \cap \imath\mathbb{R} = \varnothing$. Then for the stabilizing solution $X \in \mathbb{R}^{n\times n}$ of the ARE (5.1) it holds

$$\ker(X) = \ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big) \cap E^s(A) \tag{5.6}$$

or equivalently

$$\operatorname{im}(X) = \operatorname{im}\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)\big) + \big(E^c\big(A^\mathsf{T}\big) \oplus E^u\big(A^\mathsf{T}\big)\big). \tag{5.7}$$

◊

Proof. According to Theorem 5.4 (i), the stabilizing solution $X \in \mathbb{R}^{n\times n}$ of the ARE (5.1) is symmetric. Hence, using the orthogonal complement, it is easy to see that Equation (5.6) and Equation (5.7) are equivalent; cf. Theorem 2.47 (iv).

We divide the proof into three parts:

(i) $\ker(X) \subseteq \ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big)$.

(ii) $\ker(X) \subseteq E^s(A)$.

(iii) $\ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big) \cap E^s(A) \subseteq \ker(X)$.

(i) Let $x \in \ker(X)$. We multiply the ARE (5.1) with x^T from the left and with x from the right and get $x^\mathsf{T}C^\mathsf{T}Cx = 0$. It follows that $Cx = 0$. Again, we multiply the ARE (5.1) with x from the right, use that $Cx = 0$, and get $XAx = 0$. This means that $\ker(X)$ is an A-invariant subspace. Consequently, we have shown $\ker(X) \subseteq \ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big)$.

(ii) Here, we can assume that $\ker(X)$ is nontrivial. Let the columns of the matrix N be a basis of $\ker(X)$. We have already seen in (i) that $\ker(X)$ is an A-invariant subspace. Therefore, there is a real matrix $\tilde{A}$ such that

$$AN = N\tilde{A}. \tag{5.8}$$

With that, the characteristic polynomial $\chi_{\tilde{A}}$ of $\tilde{A}$ divides the characteristic polynomial χ_A of A. Now, the main observation is

$$\big(A - BB^\mathsf{T}X\big)N = AN = N\tilde{A}.$$

This means that $\ker(X)$ is $\big(A - BB^\mathsf{T}X\big)$-invariant, and, therefore, the characteristic polynomial $\chi_{\tilde{A}}$ divides the characteristic polynomial of $A - BB^\mathsf{T}X$ as well. As $A - BB^\mathsf{T}X$ is stable, it follows that the roots of $\chi_{\tilde{A}}$ are a subset of $\mathbb{C}_-$. With the notation of Definition 2.46, we have that $\chi_{\tilde{A}}$ divides χ_s. This yields

$$\ker(\chi_{\tilde{A}}(A)) \subseteq \ker(\chi_s(A)) \overset{\text{def.}}{=} E^s(A). \tag{5.9}$$

Next, the Cayley–Hamilton theorem and Equation (5.8) yield

$$\chi_{\tilde{A}}(A)N = N\chi_{\tilde{A}}(\tilde{A}) = 0.$$

This means $\operatorname{im}(N) \subseteq \ker(\chi_{\tilde{A}}(A))$. The columns of N were chosen to be a basis of $\ker(X)$, thus,

$$\ker(X) \subseteq \ker(\chi_{\tilde{A}}(A)). \tag{5.10}$$

Finally, we combine Equations (5.9) and (5.10).

(iii) Again, we assume that $\ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big) \cap E^s(A)$ is nontrivial and that the columns of N are a basis of this space. We have $CN = 0$. Because $\ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big)$ and $E^s(A)$ are A-invariant, the intersection of these spaces is A-invariant as well. Therefore, there is a real matrix $\tilde{A}$ such that $AN = N\tilde{A}$. Moreover, the matrix $\tilde{A}$ is stable.[1] Next, if X is the stabilizing solution of the ARE (5.1), then

$$\big(A - BB^\mathsf{T}X\big)^\mathsf{T}X + XA + C^\mathsf{T}C = 0.$$

We multiply the latter equation with N from the right, use $CN = 0$ and $AN = N\tilde{A}$ and observe that the product XN satisfies the ASE

$$\big(A - BB^\mathsf{T}X\big)^\mathsf{T}XN + XN\tilde{A} = 0.$$

As $A - BB^\mathsf{T}X$ and $\tilde{A}$ are stable, the ASE has the unique solution $XN = 0$; cf. Theorem 3.1. Consequently,

$$\ker\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)^\mathsf{T}\big) \cap E^s(A) = \operatorname{im}(N) \subseteq \ker(X),$$

and, the proof is complete. □

Theorem 5.14 (Algebraic Bernoulli Equation):
Assume that (A, B) is stabilizable and $\Lambda(A) \cap \imath\mathbb{R} = \varnothing$. Then for the stabilizing solution $X \in \mathbb{R}^{n \times n}$ of the *algebraic Bernoulli equation* (ABE)

$$A^\mathsf{T}X + XA - XBB^\mathsf{T}X = 0 \tag{5.11}$$

it holds that $\ker(X) = E^s(A)$ or equivalently $\operatorname{im}(X) = E^u\big(A^\mathsf{T}\big)$. ◊

Proof. The associated Hamiltonian matrix $\mathcal{H}$ to the ABE (5.11) is given by

$$\mathcal{H} = \begin{bmatrix} A & -BB^\mathsf{T} \\ 0 & -A^\mathsf{T} \end{bmatrix}.$$

The Hamiltonian matrix $\mathcal{H}$ is block upper triangular, therefore, $\Lambda(\mathcal{H}) = \Lambda(A) \cup \Lambda(-A)$. As A has no eigenvalue with zero real part, the associated Hamiltonian matrix $\mathcal{H}$ also

[1] $0 = \chi_s(A)N$, $AN = N\tilde{A}$, $\tilde{A}x = \lambda x$ gives $0 = \chi_s(A)Nx = N\chi_s(\tilde{A})x = \chi_s(\lambda)Nx$, hence, $\chi_s(\lambda) = 0$.

not. Theorems 5.4 and 5.7 yield the existence of a unique stabilizing solution $X \in \mathbb{R}^{n\times n}$ of the ABE (5.11). Moreover, the solution X is spsd. Utilizing Theorem 5.13, we have $\ker(X) = E^s(A)$ or equivalently $\operatorname{im}(X) = E^s(A)^\perp$. Because of A has no eigenvalues with zero real part, the center subspace of A is trivial, e.g., $E^c(A) = \{0\}$, and Theorem 2.47 (iv) completes the proof. □

Theorem 5.15 (Kernel/Image of the Stabilizing Solution II): Assume that (A, B) is stabilizable and that (A, C) is detectable. Then for the stabilizing solution $X \in \mathbb{R}^{n\times n}$ of the ARE (5.1) it holds

$$\ker(X) = \ker\bigl(\mathcal{K}\bigl(A^\mathsf{T}, C^\mathsf{T}\bigr)^\mathsf{T}\bigr) \text{ or equivalently } \operatorname{im}(X) = \operatorname{im}\bigl(\mathcal{K}\bigl(A^\mathsf{T}, C^\mathsf{T}\bigr)\bigr). \qquad \diamondsuit$$

Proof. Combine Theorems 5.9 (v), 5.11 and 5.13. □

Theorems 5.14 and 5.15 are special cases of Theorem 5.13 or [86, Thm. 7]. Both results have been rediscovered several times in the literature. In particular, Theorem 5.14 can be found in a similar form in [10, Prop. 1]. Furthermore, Theorem 5.15 has been proven in [14, Thm. 3.2] and [4, Sec. 3.3].

Theorem 5.15 is of great importance for the numerical approximation of the solution of large-scale AREs because the explicit characterization of the image justifies the choice of the trial space of Krylov subspace methods for the numerical approximation of the solution of the ARE (5.1); cf. [105,106] and [18, Sec. 4.1]. Theorem 5.14 finds application in the numerical solution of large-scale ABEs; cf. [54, Sec. 7.4.3, Alg. 7.4.3].

CHAPTER 6

DIFFERENTIAL RICCATI EQUATION

Contents

In this chapter, we consider the *nonsymmetric differential Riccati equation* (NDRE), *symmetric differential Riccati equation* (SDRE), and the *differential Riccati equation* (DRE). Similar to the ARE, these equations find applications in control and system theory. Throughout this chapter, we focus on the autonomous version of the NDRE, SDRE, and DRE.

We organize Chapter 6 as follows. In Section 6.1, we introduce the NDRE and review the existence and uniqueness of solutions. Section 6.2 focuses on the SDRE and reviews the monotonicity and comparison theorem. Section 6.3 deals with the DRE. Section 6.4 contrasts the Davison–Maki and modified Davison–Maki method for the numerical solution of the NDRE. In Section 6.5, we develop a Galerkin approach for the numerical solution of the large-scale DRE. Finally, Section 6.6 reviews the splitting schemes for the DRE.

6.1 Nonsymmetric Differential Riccati Equation

We consider the NDRE

$$\dot{X}(t) = M_{22}X(t) - X(t)M_{11} - X(t)M_{12}X(t) + M_{21}, \tag{6.1a}$$
$$X(0) = M_0 \tag{6.1b}$$

for matrices $M_{11} \in \mathbb{R}^{n\times n},\ M_{12} \in \mathbb{R}^{n\times m},\ M_{21} \in \mathbb{R}^{m\times n},\ M_{22} \in \mathbb{R}^{m\times m}$ and an initial value $M_0 \in \mathbb{R}^{m\times n}$.

6.1.1 Existence and Uniqueness of Solutions

We review the existence and uniqueness of solutions of the NDRE (6.1).

Theorem 6.1 (Existence and Uniqueness, [1, Sec. 3.1]):
The NDRE (6.1) has a unique solution $X\colon (t_-, t_+) \to \mathbb{R}^{m\times n}$. ◊

Proof. We utilize Lemma 2.37. We define $x(t) := \operatorname{vec}(X(t))$ and consider the equivalent representation of the NDRE (6.1)

$$\dot{x}(t) = \big(I_n \otimes M_{22} - M_{11}^{\mathsf{T}} \otimes I_m\big)x(t) - \big(X(t)^{\mathsf{T}} \otimes X(t)\big)\operatorname{vec}(M_{12}) + \operatorname{vec}(M_{21}), \tag{6.2a}$$
$$x(0) = \operatorname{vec}(M_0). \tag{6.2b}$$

We note that $X(t) = \operatorname{vec}^{-1}(x(t))$. The right-hand side associated to Equation (6.2a) is given by

$$\begin{aligned} &f\colon \mathbb{R} \times \mathbb{R}^{mn} \to \mathbb{R}^{mn} \\ &f(t,x) = \big(I_n \otimes M_{22} - M_{11}^{\mathsf{T}} \otimes I_m\big)x - \big(\operatorname{vec}^{-1}(x)^{\mathsf{T}} \otimes \operatorname{vec}^{-1}(x)\big)\operatorname{vec}(M_{12}) + \operatorname{vec}(M_{21}). \end{aligned}$$

Each component of $f(t,x)$ is a (multivariate) polynomial in the variables $x_1, \ldots, x_{mn}$. Consequently, f is locally Lipschitz continuous with respect to x; cf. Lemma 2.30. Theorem 2.29 completes the proof. □

6.1.2 Radon's Lemma

We review Radon's Lemma. Radon's Lemma is a fundamental tool for the NDRE, as it relates the solution of the nonlinear NDRE to the solution of a linear system of size $(m+n) \times (m+n)$.

Theorem 6.2 (Radon's Lemma, [1, Thm. 3.1.1]):
Let $\mathbb{I} \subseteq \mathbb{R}$ be an open interval with $0 \in \mathbb{I}$. The following holds:

(i) Let $X \colon \mathbb{I} \to \mathbb{R}^{m\times n}$ be the solution of the NDRE (6.1) and $U \colon \mathbb{I} \to \mathbb{R}^{n\times n}$ be the solution of the linear IVP

$$\begin{aligned} \dot{U}(t) &= (M_{11} + M_{12}X(t))U(t), \\ U(0) &= I. \end{aligned} \tag{6.3}$$

Moreover, let $V(t) := X(t)U(t)$. Then $U \colon \mathbb{I} \to \mathbb{R}^{n\times n}$ and $V \colon \mathbb{I} \to \mathbb{R}^{m\times n}$ define the solution of

$$\begin{bmatrix} \dot{U}(t) \\ \dot{V}(t) \end{bmatrix} = M \begin{bmatrix} U(t) \\ V(t) \end{bmatrix} := \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} U(t) \\ V(t) \end{bmatrix}, \quad \begin{bmatrix} U(0) \\ V(0) \end{bmatrix} = \begin{bmatrix} I \\ M_0 \end{bmatrix}. \tag{6.4}$$

(ii) If $\begin{bmatrix} U \\ V \end{bmatrix} \colon \mathbb{I} \to \mathbb{R}^{(m+n)\times n}$ is a solution of Equation (6.4) and the matrix $U(t)$ is nonsingular for all $t \in \mathbb{I}$, then $X \colon \mathbb{I} \to \mathbb{R}^{m\times n},\ X(t) = V(t)U(t)^{-1}$ solves the NDRE (6.1).
◊

Proof.

(i) Utilizing Theorem 2.31, the linear IVP (6.3) has a unique solution U defined on the interval $\mathbb{I}$. It can be verified that U and V satisfy Equation (6.4).

(ii) We differentiate $X(t) = V(t)U(t)^{-1}$ and get

$$\dot{X}(t) = \dot{V}(t)U(t)^{-1} - V(t)U(t)^{-1}\dot{U}(t)U(t)^{-1}.$$

The claim follows from Equation (6.4) and the identity $X(t) = V(t)U(t)^{-1}$. □

Radon's Lemma (Theorem 6.2) also holds for time-dependent continuous matrix-valued functions as coefficients; cf. [1, Thm. 3.1.1]. Usually, the solution of the NDRE (6.1) has finite-time escape, while the solution of the system (6.4) exists for all $t \in \mathbb{R}$; cf. Theorem 2.31. As the function U is a solution of the linear IVP (6.3) and $U(0) = I$ is nonsingular, the determinant of $U(t)$ can not vanish on the interval $\mathbb{I}$. It follows that the matrix $U(t)$ is nonsingular for all $t \in \mathbb{I}$; cf. Theorem 2.31 Equation (2.5). Therefore, as long as the solution of the NDRE (6.1) exists, it can be recovered from the solution of the system (6.4) with the transformation $X(t) = V(t)U(t)^{-1}$.

6.2 Symmetric Differential Riccati Equation

Section 6.2 focuses on the SDRE

$$\dot{X}(t) = A^{\mathsf{T}} X(t) + X(t)A - X(t)SX(t) + Q, \tag{6.5a}$$
$$X(0) = X_0, \tag{6.5b}$$

for symmetric matrices $S, Q, X_0 \in \mathbb{R}^{n\times n}$ and coefficient matrix $A \in \mathbb{R}^{n\times n}$. We recall the existence and uniqueness of a symmetric solution. Then, we present the monotonicity theorem and the comparison theorem for the SDRE (6.5) (Theorems 6.5 and 6.7). This section is based on [1, Sec. 4.1.1] and [66, Sec. 10.1–10.2].

6.2.1 Existence and Uniqueness of Solutions

We review the existence and uniqueness of symmetric solutions of the SDRE (6.5).

Theorem 6.3 (Existence and Uniqueness, [1, Sec. 4.1]):
The SDRE (6.5) has a unique solution $X\colon (t_-, t_+) \to \mathbb{R}^{n\times n}$, and $X(t)$ is symmetric for all $t \in (t_-, t_+)$. ◊

Proof. Theorem 6.1 ensures the existence and uniqueness of solutions. As the matrices S, Q, and X_0 are symmetric, $X(t)^{\mathsf{T}}$ satisfies the SDRE (6.5) as well. Therefore, $X(t)$ is symmetric for all $t \in (t_-, t_+)$. □

6.2.2 Monotonicity Theorem

In preparation of Theorem 6.5, we show that the number of negative, zero, and positive eigenvalues of the derivative $\dot{X}(t)$ is constant as long as $\dot{X}(t)$ exists. The reference [66, Hilfssatz 10.4] states a slightly weaker result, hence we provide a proof of Lemma 6.4.

Lemma 6.4 (Inertia of the Derivative, [66, Hilfssatz 10.4]):
Let $X\colon (t_-, t_+) \to \mathbb{R}^{n\times n}$ be the unique solution of the SDRE (6.5). Then the inertia of the derivative $\dot{X}(t)$ is constant for all $t \in (t_-, t_+)$. ◊

Proof. We differentiate the SDRE (6.5a), use the symmetry of S and $X(t)$, and get

$$\ddot{X}(t) = (A - SX(t))^{\mathsf{T}} \dot{X}(t) + \dot{X}(t)(A - SX(t))$$
$$\dot{X}(0) = A^{\mathsf{T}} X(0) + X(0)A - X(0)SX(0) + Q.$$

Therefore, the derivative $\dot{X}$ satisfies a DLE. Theorem 4.1 yields

$$\dot{X}(t) = \Phi(t)\dot{X}(0)\Phi(t)^{\mathsf{T}}$$

for all $t \in (t_-, t_+)$, where $\Phi(t)$ is the fundamental matrix of the system

$$\dot{x}(t) = (A - SX(t))^{\mathsf{T}} x(t)$$

with $\Phi(0) = I$. Consequently, the matrices $\dot{X}(t)$ and $\dot{X}(0)$ are congruent and Sylvester's law of inertia completes the proof. □

Theorem 6.5 (Monotonicity Theorem, [1, Lem. 4.1.12]):
Let $X\colon (t_-, t_+) \to \mathbb{R}^{n\times n}$ be the unique solution of the SDRE (6.5). It holds:

- If $\dot{X}(0) \succcurlyeq 0$, then $X(t_1) \preccurlyeq X(t_2)$ for all $t_1, t_2 \in (t_-, t_+)$ such that $t_1 \leq t_2$, i.e., X is monotonic non-decreasing on (t_-, t_+).
- If $\dot{X}(0) \preccurlyeq 0$, then $X(t_1) \succcurlyeq X(t_2)$ for all $t_1, t_2 \in (t_-, t_+)$ such that $t_1 \leq t_2$, i.e., X is monotonic non-increasing on (t_-, t_+). ♢

Proof. We assume that $\dot{X}(0) \succcurlyeq 0$. Lemma 6.4 yields that $\dot{X}(t) \succcurlyeq 0$ for all $t \in (t_-, t_+)$. It holds that

$$X(t_2) - X(t_1) = \int_{t_1}^{t_2} \dot{X}(s)\,\mathrm{d}s \succcurlyeq 0.$$

If $\dot{X}(0) \preccurlyeq 0$, then the claim can be shown by similar arguments. □

The monotonicity property of the solution of the SDRE is remarkable, because the Loewner ordering is only a partial ordering on the set of symmetric matrices, and, hence not all symmetric matrices are comparable.

6.2.3 Comparison Theorem

The comparison theorem (Theorem 6.7) is the main result of this subsection. The next lemma states that, if a matrix function $X(t)$ satisfies the differential inequality (6.6a) and is spsd at time t_0, then $X(t)$ remains spsd at all subsequent times. To avoid confusion, we provide a proof of Lemma 6.6 and Theorem 6.7 because the given references consider the SDRE backward in time.

Lemma 6.6 (Linear Differential Inequality, [1, Thm. 4.1.2]):
Let $\mathbb{I} \subseteq \mathbb{R}$ be an open interval with $t_0 \in \mathbb{I}$, $A\colon \mathbb{I} \to \mathbb{R}^{n\times n}$ be continuous, and $X\colon \mathbb{I} \to \mathbb{R}^{n\times n}$ be continuously differentiable. If X satisfies the inequalities

$$\dot{X}(t) \succcurlyeq A(t)^{\mathsf{T}} X(t) + X(t)A(t), \tag{6.6a}$$
$$X(t_0) \succcurlyeq 0, \tag{6.6b}$$

then $X(t) \succcurlyeq 0$ for all $t \in \mathbb{I}$ such that $t \geq t_0$. ♢

Proof. Let $\Phi(t)$ be the fundamental matrix of the system

$$\dot{x}(t) = -A(t)x(t)$$

with $\Phi(t_0) = I$. We differentiate the matrix

$$Y(t) := \Phi(t)^{\mathsf{T}} X(t)\Phi(t)$$

and get

$$\dot{Y}(t) = -\Phi(t)^\mathsf{T}\big(A(t)^\mathsf{T}X(t) + X(t)A(t)\big)\Phi(t) + \Phi(t)^\mathsf{T}\dot{X}(t)\Phi(t) \succcurlyeq 0.$$

Therefore, $Y(t)$ is monotonically non-decreasing on $\mathbb{I}$. We have

$$\Phi(t)^\mathsf{T}X(t)\Phi(t) = Y(t) \succcurlyeq Y(t_0) = X(t_0) \succcurlyeq 0. \tag{6.7}$$

for all $t \in \mathbb{I}$ such that $t \geq t_0$. The matrix $\Phi(t)$ is nonsingular. Multiplication of Equation (6.7) with $\Phi(t)^{-\mathsf{T}}$ and $\Phi(t)^{-1}$ from the left and the right, respectively, completes the proof. □

In order to formulate the comparison theorem (Theorem 6.7), it is useful to associate the real Hamiltonian matrix

$$\mathcal{H} = \begin{bmatrix} A & -S \\ -Q & -A^\mathsf{T} \end{bmatrix}$$

to the SDRE (6.5)

$$\begin{aligned} \dot{X}(t) &= A^\mathsf{T}X(t) + X(t)A - X(t)SX(t) + Q, \\ X(0) &= X_0. \end{aligned}$$

With that, we can write the SDRE (6.5) as

$$\begin{aligned} \dot{X}(t) &= \begin{bmatrix} I \\ X(t) \end{bmatrix}^\mathsf{T} \begin{bmatrix} Q & A^\mathsf{T} \\ A & -S \end{bmatrix} \begin{bmatrix} I \\ X(t) \end{bmatrix} = -\begin{bmatrix} I \\ X(t) \end{bmatrix}^\mathsf{T} \mathcal{J}\mathcal{H} \begin{bmatrix} I \\ X(t) \end{bmatrix}, \\ X(0) &= X_0, \end{aligned}$$

where

$$\mathcal{J} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix};$$

cf. Definition 2.40. The comparison theorem yields that solution $X(t)$ depends monotonically on the real symmetric matrix $-\mathcal{J}\mathcal{H}$ and on the initial value X_0.

Theorem 6.7 (Comparison Theorem, [1, Thm. 4.1.4]):
Let $\mathbb{I}_i \subseteq \mathbb{R}$ be open intervals with $0 \in \mathbb{I}_i$ and $X_i \colon \mathbb{I}_i \to \mathbb{R}^{n\times n}$ $(i = 1, 2)$ be the solutions of the SDREs

$$\begin{aligned} \dot{X}_i(t) &= A_i^\mathsf{T}X_i(t) + X_i(t)A_i - X_i(t)S_iX_i(t) + Q_i, \\ X_i(0) &= X_{0,i}. \end{aligned}$$

If $X_{0,1} \preccurlyeq X_{0,2}$ and $-\mathcal{J}\mathcal{H}_1 \preccurlyeq -\mathcal{J}\mathcal{H}_2$, then $X_1(t) \preccurlyeq X_2(t)$ for all $t \geq 0$ such that $t \in \mathbb{I}_1 \cap \mathbb{I}_2$.$\lozenge$

Proof. We consider the difference $X(t) := X_2(t) - X_1(t)$ and get

$$\begin{aligned}\dot{X}(t) =& - \begin{bmatrix} I \\ X_2(t) \end{bmatrix}^{\mathsf{T}} \mathcal{J}\mathcal{H}_2 \begin{bmatrix} I \\ X_2(t) \end{bmatrix} + \begin{bmatrix} I \\ X_1(t) \end{bmatrix}^{\mathsf{T}} \mathcal{J}\mathcal{H}_1 \begin{bmatrix} I \\ X_1(t) \end{bmatrix} \\ =& - \begin{bmatrix} I \\ X_2(t) \end{bmatrix}^{\mathsf{T}} \mathcal{J}\mathcal{H}_2 \begin{bmatrix} I \\ X_2(t) \end{bmatrix} + \begin{bmatrix} I \\ X_1(t) \end{bmatrix}^{\mathsf{T}} \mathcal{J}\mathcal{H}_2 \begin{bmatrix} I \\ X_1(t) \end{bmatrix} \\ &+ \begin{bmatrix} I \\ X_1(t) \end{bmatrix}^{\mathsf{T}} (\mathcal{J}\mathcal{H}_1 - \mathcal{J}\mathcal{H}_2) \begin{bmatrix} I \\ X_1(t) \end{bmatrix} \\ \succcurlyeq& - \begin{bmatrix} I \\ X_2(t) \end{bmatrix}^{\mathsf{T}} \mathcal{J}\mathcal{H}_2 \begin{bmatrix} I \\ X_2(t) \end{bmatrix} + \begin{bmatrix} I \\ X_1(t) \end{bmatrix}^{\mathsf{T}} \mathcal{J}\mathcal{H}_2 \begin{bmatrix} I \\ X_1(t) \end{bmatrix} \\ =& A_2^{\mathsf{T}} X(t) + X(t) A_2 - X_2(t) S_2 X_2(t) + X_1(t) S_2 X_1(t).\end{aligned}$$

Hence, X satisfies the inequalities

$$\begin{aligned}\dot{X}(t) &\succcurlyeq A(t)^{\mathsf{T}} X(t) + X(t) A(t), \\ X(0) &= X_{0,2} - X_{0,1} \succcurlyeq 0,\end{aligned}$$

where $A(t) := A_2 - \frac{1}{2} S_2 (X_2(t) + X_1(t))$. The claim follows from Lemma 6.6. □

6.3 Differential Riccati Equation

In this section, we consider the DRE

$$\begin{aligned}\dot{X}(t) &= A^{\mathsf{T}} X(t) + X(t) A - X(t) B B^{\mathsf{T}} X(t) + C^{\mathsf{T}} C, && (6.8a) \\ X(0) &= Z_0 Z_0^{\mathsf{T}}, && (6.8b)\end{aligned}$$

for matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times b}$, $C \in \mathbb{R}^{c \times n}$, and $Z_0 \in \mathbb{R}^{n \times z_0}$. The DRE (6.8) is a special case of the SDRE (6.5).

We organize this section as follows. In Section 6.3.1, we discuss the existence and uniqueness, the definiteness, and the long-time existence of solutions of the DRE (6.8). In Section 6.3.2, we study the image or equivalently the kernel of the solution of the DRE (6.8). We review the convergence of the solution of the DRE (6.8) to the stationary point in Section 6.3.3. Finally, in Section 6.3.4, we review an explicit solution formula based on the solution of the ARE.

6.3.1 Existence and Uniqueness of Solutions

Here, we review the existence and uniqueness of solutions of the DRE (6.8). The initial condition (6.8b) yields that $X(0)$ is spsd and this is essential to establish the definiteness and long-time existence of the solution $X(t)$. We provide a proof of Theorem 6.8, because the given references consider the DRE backward in time and hence the terminology slightly differs.

Theorem 6.8 (Existence and Uniqueness,[1, Thm. 4.1.6], [66, Sec. 10.2]):
The DRE (6.8) admits a unique solution $X\colon (t_{-},\infty) \to \mathbb{R}^{n\times n}$, and $X(t)$ is symmetric for all $t \in (t_{-},\infty)$. Furthermore, $X(t)$ is spsd for all $t \geq 0$. $\Diamond$

Proof.
Existence, Uniqueness, and Symmetry:
Theorem 6.3 ensures the existences, uniqueness and symmetry of the solution.

$X(t)$ is spsd:
The solution X satisfies

$$\begin{aligned}
&\dot{X}(t) \succcurlyeq \left(A - \tfrac{1}{2}BB^{\mathsf{T}}X(t)\right)^{\mathsf{T}}X(t) + X(t)\left(A - \tfrac{1}{2}BB^{\mathsf{T}}X(t)\right),\\
&X(0) = Z_0Z_0^{\mathsf{T}} \succcurlyeq 0.
\end{aligned}$$

Hence, Lemma 6.6 yields that $X(t)$ is spsd for all $t \geq 0$.

Long-time Existence:
It remains to show that the maximal interval of existence is (t_{-},∞). For this, let $X_{\mathrm{u}}\colon \mathbb{R} \to \mathbb{R}^{n\times n}$ be the unique solution of the DLE

$$\begin{aligned}
&\dot{X}(t) = A^{\mathsf{T}}X(t) + X(t)A + C^{\mathsf{T}}C,\\
&X(0) = Z_0Z_0^{\mathsf{T}},
\end{aligned}$$

and assume that $t_{+} < \infty$ holds. We make use of the comparison theorem (Theorem 6.7) and get $X(t) \preccurlyeq X_{\mathrm{u}}(t)$ for all $t \in [0,t_{+})$. Therefore,

$$0 \preccurlyeq X(t) \preccurlyeq X_{\mathrm{u}}(t)$$

for all $t \in [0,t_{+})$. Theorem 2.4 (ii) gives $\|X(t)\|_2 \preccurlyeq \|X_u(t)\|_2$. Hence, the solution X does not escape in finite-time (in positive direction of time), and Theorem 2.29 completes the proof. $\square$

In general, the left-endpoint t_{-} of the maximal interval of existence (t_{-},∞) of the solution of the DRE (6.8) is finite. We illustrate this by an example.

Example 6.1 (Finite-Time Escape):
We consider the scalar DRE

$$\begin{aligned}
&\dot{x}(t) = -x(t) - x(t)^2,\\
&x(0) = 1.
\end{aligned}$$

$\Diamond$

The solution is given by $x(t) = \frac{1}{2e^{t}-1}$. The maximal interval of existence is $(-\ln(2),\infty)$. Figure 6.1 shows the graph of the solution $x(t)$ and the finite-time escape in negative direction of time.

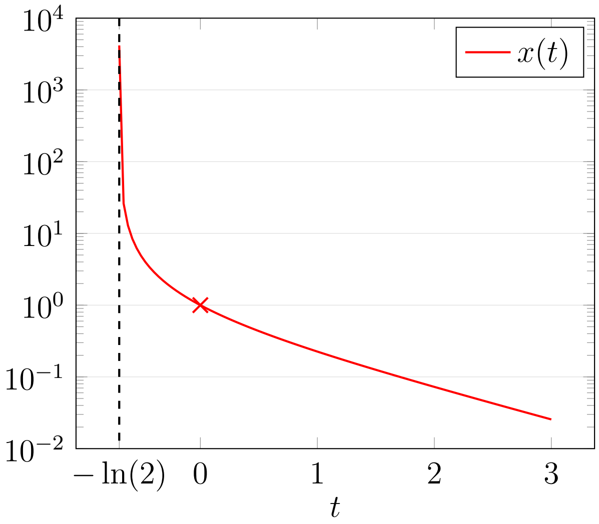


Fig. 6.1: Finite-time Escape of the Solution in Negative Direction of Time.

6.3.2 Image of the Solution

We focus on the image or equivalently the kernel of the solution of the DRE (6.8). Knowledge of the image of the solution of the DRE (6.8) finds application in Krylov subspace methods for the numerical solution approximation of large-scale DREs; cf. [6, 49, 65, 70]. Theorem 6.9 gives an upper bound on the image of the solutions. Theorem 6.10 focuses on the solution with zero initial conditions. Theorem 6.11 yields a results for positive times and is based on the latter two theorems and the comparison theorem Theorem 6.7.

Theorem 6.9 (Image of the Solution I):
Let $X\colon (t_{-}, \infty) \to \mathbb{R}^{n\times n}$ be the unique solution of the DRE (6.8). Then

$$\operatorname{im}(X(t)) \subseteq \operatorname{im}\big(\mathcal{K}\big(A^{\mathsf{T}}, \big[C^{\mathsf{T}}, Z_0\big]\big)\big)$$

for all $t \in (t_{-}, \infty)$. $\lozenge$

Proof. If X satisfies the DRE (6.8), then X also satisfies the DSE

$$\begin{aligned}\dot{X}(t) &= A^{\mathsf{T}} X(t) + X(t)\big(A - BB^{\mathsf{T}} X(t)\big) + C^{\mathsf{T}} C,\\ X(0) &= Z_0 Z_0^{\mathsf{T}},\end{aligned}$$

with coefficient matrices A^{T} and $A - BB^{\mathsf{T}}X(t)$. Theorem 4.1 yields

$$X(t) = e^{tA^{\mathsf{T}}} Z_0 Z_0^{\mathsf{T}} \Psi(t)^{\mathsf{T}} + \int\limits_0^t e^{(t-s)A^{\mathsf{T}}} C^{\mathsf{T}} C \Psi(s)^{-1} \Psi(t)^{\mathsf{T}} \, \mathrm{d}s, \tag{6.9}$$

where $\Psi(t)$ is the fundamental matrix of the system

$$\dot{x}(t) = \big(A - BB^\mathsf{T}X(t)\big)^\mathsf{T}x(t)$$

with $\Psi(0) = I$. We assume that $C \neq 0$ or $Z_0 \neq 0$. Let the columns of the matrix Q be an orthonormal basis of the linear space

$$\operatorname{im}\big(\mathcal{K}\big(A^\mathsf{T}, \big[C^\mathsf{T}, Z_0\big]\big)\big) \subseteq \mathbb{R}^n.$$

The space is A^T-invariant, and the matrix QQ^T represents its orthogonal projection. Therefore,

$$A^\mathsf{T}Q = QQ^\mathsf{T}A^\mathsf{T}Q,$$

from which it follows that

$$e^{tA^\mathsf{T}}Q = Qe^{tQ^\mathsf{T}A^\mathsf{T}Q}. \tag{6.10}$$

Moreover,

$$C^\mathsf{T} = QQ^\mathsf{T}C^\mathsf{T} \text{ and } Z_0 = QQ^\mathsf{T}Z_0. \tag{6.11}$$

We combine Equations (6.9)–(6.11). The identity

$$\begin{aligned} X(t) &= e^{tA^\mathsf{T}}QQ^\mathsf{T}Z_0Z_0^\mathsf{T}\Psi(t)^\mathsf{T} + \int\limits_0^t e^{(t-s)A^\mathsf{T}}QQ^\mathsf{T}C^\mathsf{T}C\Psi(s)^{-1}\Psi(t)^\mathsf{T}\,\mathrm{d}s \\ &= Q\left(e^{tQ^\mathsf{T}A^\mathsf{T}Q}Q^\mathsf{T}Z_0Z_0^\mathsf{T}\Psi(t)^\mathsf{T} + \int\limits_0^t e^{(t-s)Q^\mathsf{T}A^\mathsf{T}Q}Q^\mathsf{T}C^\mathsf{T}C\Psi(s)^{-1}\Psi(t)^\mathsf{T}\,\mathrm{d}s\right) \end{aligned}$$

completes the proof. □

Theorem 6.10 (Image of the Solution II):
Let $X\colon (t_-, \infty) \to \mathbb{R}^{n\times n}$ be the unique solution of the DRE (6.8a) such that $X(0) = 0$. Then

$$\operatorname{im}(X(t)) = \operatorname{im}\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)\big)$$

for all $t \in (t_-, \infty) \setminus \{0\}$. ◊

Proof. First, we show that $\operatorname{im}\big(\mathcal{K}\big(A^\mathsf{T}, C^\mathsf{T}\big)\big) \subseteq \operatorname{im}(X(t))$. We divide the proof into three parts:

(i) $X(t)x = 0$ implies $Cx = 0$,

(ii) $X(t)x = 0$ implies $\dot{X}(t)x = 0$,

(iii) $\ker(X(t))$ is A-invariant,

for all $t \in (t_-, \infty) \setminus \{0\}$.

(i) If X satisfies the DRE (6.8a), then

$$\begin{aligned}\dot{X}(t) &= \big(A - \tfrac{1}{2}BB^{\mathsf{T}}X(t)\big)^{\mathsf{T}}X(t) + X(t)\big(A - \tfrac{1}{2}BB^{\mathsf{T}}X(t)\big) + C^{\mathsf{T}}C,\\ X(0) &= 0.\end{aligned}$$

Theorem 4.1 yields

$$X(t) = \int\limits_0^t \Phi(t)\Phi(s)^{-1}C^{\mathsf{T}}C\Phi(s)^{-\mathsf{T}}\Phi(t)^{\mathsf{T}}\,\mathrm{d}s,$$

where $\Phi(t)$ is the fundamental matrix of the system

$$\dot{x}(t) = \big(A - \tfrac{1}{2}BB^{\mathsf{T}}X(t)\big)^{\mathsf{T}}x(t)$$

with $\Phi(0) = I$. Let $x \in \ker(X(t))$ be and assume that $t > 0$. We have

$$0 = x^{\mathsf{T}}X(t)x = \int\limits_0^t \big\|C\Phi(s)^{-\mathsf{T}}\Phi(t)^{\mathsf{T}}x\big\|_2^2\,\mathrm{d}s$$

For reasons of continuity, it follows that

$$C\Phi(s)^{-\mathsf{T}}\Phi(t)^{\mathsf{T}}x = 0$$

for all $s \in [0,t]$. Therefore, for $s = t$, it follows that $Cx = 0$. The argumentation for $t < 0$ is analoguous.

(ii) The assumption $X(0) = 0$ gives $\dot{X}(0) = C^{\mathsf{T}}C \succcurlyeq 0$. Lemma 6.4 yields $\dot{X}(t) \succcurlyeq 0$. We know that $X(t)x = 0$ implies $Cx = 0$. We multiply the DRE with x^{T} from the left and with x from the right,

$$0 \leq x^{\mathsf{T}}\dot{X}(t)x = x^{\mathsf{T}}\big(A^{\mathsf{T}}X(t) + X(t)A - X(t)BB^{\mathsf{T}}X(t) + C^{\mathsf{T}}C\big)x = 0.$$

Lemma 2.5 gives $\dot{X}(t)x = 0$.

(iii) We have $X(t)x = 0$ implies $Cx = 0$ and $\dot{X}(t)x = 0$ for all $t \neq 0$. We multiply the DRE (6.8a) from the right with x and get

$$0 = \dot{X}(t)x = A^{\mathsf{T}}X(t)x + X(t)Ax - X(t)BB^{\mathsf{T}}X(t)x + C^{\mathsf{T}}Cx = X(t)Ax.$$

Consequently, $X(t)x = 0$ implies $X(t)Ax = 0$. Hence, $\ker(X(t))$ is A-invariant.

Parts (i), (ii), and (iii) give

$$\ker(X(t)) \subseteq \ker\big(\mathcal{K}\big(A^{\mathsf{T}}, C^{\mathsf{T}}\big)^{\mathsf{T}}\big).$$

We consider the orthogonal complements of the spaces and get

$$\operatorname{im}\big(\mathcal{K}\big(A^{\mathsf{T}}, C^{\mathsf{T}}\big)\big) \subseteq \operatorname{im}(X(t)).$$

Theorem 6.9 provides the other inclusion. □

Theorem 6.11 (Image of the Solution III):
Let $X\colon (t_{-},\infty) \to \mathbb{R}^{n\times n}$ be the unique solution of the DRE (6.8). Then

$$\operatorname{im}\bigl(\mathcal{K}\bigl(A^\mathsf{T}, C^\mathsf{T}\bigr)\bigr) \subseteq \operatorname{im}(X(t)) \subseteq \operatorname{im}\bigl(\mathcal{K}\bigl(A^\mathsf{T}, \bigl[C^\mathsf{T}, Z_0\bigr]\bigr)\bigr)$$

for all $t > 0$. ◊

Proof. Let X_l be the unique solution of the DRE (6.8a) such that $X_l(0) = 0$. We have

$$0 = X_l(0) \preccurlyeq Z_0 Z_0^\mathsf{T} = X(0).$$

The comparison theorem (Theorem 6.7) yields $X_l(t) \preccurlyeq X(t)$ for all $t \geq 0$. The matrix $X(t)$ is spsd for all $t \geq 0$. Hence, $0 \preccurlyeq X_l(t) \preccurlyeq X(t)$ for all $t \geq 0$. Theorems 6.9 and 6.10 and Lemma 2.6 yield

$$\operatorname{im}\bigl(\mathcal{K}\bigl(A^\mathsf{T}, C^\mathsf{T}\bigr)\bigr) = \operatorname{im}(X_l(t)) \subseteq \operatorname{im}(X(t)) \subseteq \operatorname{im}\bigl(\mathcal{K}\bigl(A^\mathsf{T}, \bigl[C^\mathsf{T}, Z_0\bigr]\bigr)\bigr).$$

for all $t > 0$. □

6.3.3 Convergence to a Stationary Point

Here, we focus on the convergence of the solution of the DRE (6.8). First, we prove that the solution is bounded if the pair (A, B) is stabilizable.

The proof of Theorem 6.12 provided by both references make use of a cost functional of a linear-quadratic regulator problem associated with the DRE. Alternatively, it is sufficient to utilize the comparison theorem (Theorem 6.7).

Theorem 6.12 (Bounded Solutions, [66, Proof of Satz 10.5], [34, Thm. 8.8]):
Assume that the pair (A, B) is stabilizable. Then the solution of the DRE (6.8) is bounded on the interval $[0, \infty)$. ◊

Proof. With (A, B) stabilizable, there exists a matrix $K \in \mathbb{R}^{b\times n}$ such that $A - BK$ is stable. Let $X_\mathrm{u}\colon \mathbb{R} \to \mathbb{R}^{n\times n}$ be the solution of the DLE

$$\begin{aligned}\dot{X}_\mathrm{u}(t) &= (A - BK)^\mathsf{T} X_\mathrm{u}(t) + X_\mathrm{u}(t)(A - BK) + K^\mathsf{T}K + C^\mathsf{T}C,\\ X_\mathrm{u}(0) &= Z_0 Z_0^\mathsf{T}.\end{aligned}$$

As the matrix $A - BK$ is stable, the solution X_u converges for $t \to \infty$. In particular, X_u is bounded on the interval $[0, \infty)$. We consider the corresponding Hamiltonian matrices and get

$$\begin{bmatrix} C^\mathsf{T}C & A^\mathsf{T} \\ A & -BB^\mathsf{T} \end{bmatrix} = -\mathcal{J}\mathcal{H} \preccurlyeq -\mathcal{J}\mathcal{H}_u = \begin{bmatrix} K^\mathsf{T}K + C^\mathsf{T}C & (A - BK)^\mathsf{T} \\ A - BK & 0 \end{bmatrix}.$$

Hence, the comparison theorem (Theorem 6.7) and the definiteness of the solution X of the DRE (6.8) (Theorem 6.8) give $0 \preccurlyeq X(t) \preccurlyeq X_\mathrm{u}(t)$ for all $t \geq 0$. Finally, Theorem 2.4 (ii) yields $\|X(t)\|_2 \leq \|X_\mathrm{u}(t)\|_2$ for all $t \geq 0$. □

If $X(0) = 0$, then $\dot{X}(0) = C^\mathsf{T} C \succeq 0$. Therefore, the monotonicity theorem (Theorem 6.5) yields that X is monotonically non-decreasing. If in addition the pair (A, B) is stabilizable, then the solution is bounded. Hence, we can prove that X converges to a stationary point and obtain an existence results of a real spsd solution of the ARE (5.1) under weak assumptions. The second part of Theorem 6.13 is not provided in the references. Furthermore, we use some arguments in Theorem 6.14, therefore, we provide a proof.

Theorem 6.13 (ARE: spsd Solutions, [66, Satz 10.5], [79, Thm. 7.9.3]): Assume that (A, B) is stabilizable. Then the ARE (5.1) has a spsd solution $X_m \in \mathbb{R}^{n\times n}$. Moreover, X_m is minimal, i.e., $X_m \preccurlyeq X_s$ for all real spsd solutions of the ARE (5.1). ♢

Proof.
Existence:
Let X be the solution of the DRE (6.8a) such that $X(0) = 0$. According to Theorems 6.5 and 6.12 the solution X is monotonically non-decreasing and bounded. We consider the real-valued scalar functions

$$h_i(t) = e_i^\mathsf{T} X(t) e_i \text{ and } h_{i,j}(t) = \big(e_i + e_j\big)^\mathsf{T} X(t) \big(e_i + e_j\big).$$

The functions are bounded and monotonically non-decreasing on $[0, \infty)$, and, therefore, convergent. The convergence of $h_i(t)$ implies that every diagonal entry of $X(t)$ is convergent. The off-diagonal entries of $X(t)$ can be written as

$$e_i^\mathsf{T} X(t) e_j = \tfrac{1}{2}\big(h_{i,j}(t) - h_i(t) - h_j(t)\big).$$

Therefore, the off-diagonal entries converge as well. Hence, there is a real matrix X_m such that

$$\lim_{t\to\infty} X(t) = X_m.$$

The limit X_m is real spsd because the set of real spsd matrices is closed. Finally, as X_m is the limit of the solution of the DRE (6.8), it must be a stationary point; cf. [125, §10 XI. (h)].

Minimality:
If $X_s \in \mathbb{R}^{n\times n}$ is a spsd solution of the ARE (5.1), then X_s is also a constant solution of the DRE. Because of $X(0) = 0 \preccurlyeq X_s$, the comparison theorem (Theorem 6.7) yields $X(t) \preccurlyeq X_s$ for all $t \geq 0$, and, hence, $X_m \preccurlyeq X_s$. □

Here, we review the convergence of the solution of the DRE (6.8) under the assumption that (A, B) is stabilizable and (A, C) is detectable. The proof makes use of the comparison theorem (Theorem 6.7).

Theorem 6.14 (Convergence to the Stationary Point, [66, Satz 10.3]): Assume that (A, B) is stabilizable and (A, C) is detectable. Then the solution of the DRE (6.8) converges to the unique real spsd stabilizing solution of the ARE (5.1) for $t \to \infty$. ♢

Proof. Let X be the unique solution of the DRE (6.8). We will consider a monotonically non-decreasing lower solution and a monotonically non-increasing upper solution.

Lower Solution:

Let X_{l} be the unique solution of the DRE (6.8a) such that $X_{\mathrm{l}}(0) = 0$. Then, it follows from $X_{\mathrm{l}}(0) \preccurlyeq Z_0 Z_0^{\mathsf{T}} = X(0)$ that $X_{\mathrm{l}}(t) \preccurlyeq X(t)$ for all $t \geq 0$; cf. Theorem 6.7. Moreover, X_{l} is monotonically non-decreasing.

Upper Solution:

The construction of a monotonically non-increasing upper solution is more involved. Let $Q \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix such that $C^{\mathsf{T}} C \preccurlyeq Q$. Theorem 6.13 yields that the ARE

$$A^{\mathsf{T}} X + XA - XBB^{\mathsf{T}} X + Q = 0$$

has a real spsd solution X_m. Because of Q is positive definite, the solution X_m is positive definite as well. Therefore, there is real number $\alpha > 1$ such that $Z_0 Z_0^{\mathsf{T}} \preccurlyeq \alpha X_m$. Let X_{u} be the unique solution of the DRE

$$\begin{aligned} \dot{X}(t) &= A^{\mathsf{T}} X(t) + X(t)A - X(t)BB^{\mathsf{T}} X(t) + C^{\mathsf{T}} C, \\ X(0) &= \alpha X_m. \end{aligned}$$

Because of

$$\begin{aligned} \dot{X}_{\mathrm{u}}(0) &= \alpha\big(A^{\mathsf{T}} X_m + X_m A\big) - \alpha^2 X_m BB^{\mathsf{T}} X_m + C^{\mathsf{T}} C \\ &= \big(\alpha - \alpha^2\big) X_m BB^{\mathsf{T}} X_m + C^{\mathsf{T}} C - Q \preccurlyeq 0, \end{aligned}$$

the solution X_{u} is monotonically non-increasing; cf. Theorem 6.5. By construction, it holds that $X(0) \preccurlyeq X_{\mathrm{u}}(0)$. Hence, Theorem 6.7 gives $X(t) \preccurlyeq X_{\mathrm{u}}(t)$ for all $t \geq 0$.

Squeezing Argument:

With similar arguments as in the proof of Theorem 6.13, we infer that X_{l} and X_{u} converge to real spsd solutions of the ARE. Theorem 5.10 yields that there is a unique real spsd solution of the ARE (5.1). Hence,

$$\lim_{t \to \infty} X_{\mathrm{l}}(t) = X_{\infty} = \lim_{t \to \infty} X_{\mathrm{u}}(t).$$

Finally, the inequality $X_{\mathrm{l}}(t) \preccurlyeq X(t) \preccurlyeq X_{\mathrm{u}}(t)$ gives $\lim\limits_{t \to \infty} X(t) = X_{\infty}$. □

The convergence of the solution to a stationary point has been extensively studied in the literature. Convergence criteria under weaker assumptions as in Theorem 6.14 have been established; cf. [30–33].

6.3.4 Solution Representation

Radon's Lemma (Theorem 6.2) enables certain solution representation for the DRE (6.8): Theorem 6.8 ensures that the DRE (6.8) has a unique solution defined on the interval

(t^{-}, ∞). By Radon's Lemma (Theorem 6.2), we have that $U(t)$ is nonsingular on the same interval.

The matrices $U(t)$ and $V(t)$ are determined by the linear initial value problem

$$\begin{bmatrix} \dot{U}(t) \\ \dot{V}(t) \end{bmatrix} = -\mathcal{H} \begin{bmatrix} U(t) \\ V(t) \end{bmatrix}, \quad \begin{bmatrix} U(0) \\ V(0) \end{bmatrix} = \begin{bmatrix} I \\ Z_0 Z_0^{\mathsf{T}} \end{bmatrix}. \tag{6.12}$$

We obtain

$$\begin{bmatrix} U(t) \\ V(t) \end{bmatrix} = e^{-t\mathcal{H}} \begin{bmatrix} I \\ Z_0 Z_0^{\mathsf{T}} \end{bmatrix}.$$

The strategy is to decompose the Hamiltonian matrix $\mathcal{H}$, such that Equation (6.12) decouples. The solution formula of Theorem 6.15 was presented in the references without proof. Since the existence of the involved inverse is not trivially established, we provide a proof.

Theorem 6.15 (Solution Representation, [32, Thm. 1], [96]):
Let (A, B) be stabilizable, (A, C) be detectable, and $X_\infty \in \mathbb{R}^{n\times n}$ be the unique spsd stabilizing solution of the ARE (5.1). Moreover, let $\hat{A} := A - BB^{\mathsf{T}} X_\infty$ be and $X_{\mathrm{L}} \in \mathbb{R}^{n\times n}$ be the unique spsd solution of the ALE

$$\hat{A} X_{\mathrm{L}} + X_{\mathrm{L}} \hat{A}^{\mathsf{T}} + BB^{\mathsf{T}} = 0. \tag{6.13}$$

Then the solution of the DRE (6.8) is represented by

$$X(t) = X_\infty - e^{t\hat{A}^{\mathsf{T}}} (X_\infty - X_0) \Big(I - \Big(X_{\mathrm{L}} - e^{t\hat{A}} X_{\mathrm{L}} e^{t\hat{A}^{\mathsf{T}}} \Big) (X_\infty - X_0) \Big)^{-1} e^{t\hat{A}},$$

where $X_0 = Z_0 Z_0^{\mathsf{T}}$. ◊

Proof. Similar to the proof of Theorem 5.7, we use similarity transformations to decompose the Hamiltonian matrix $\mathcal{H}$. This is also known as a Riccati–Lyapunov transformation [1, Ch. 3.1.1]. We obtain

$$S^{-1}\mathcal{H}S := \begin{bmatrix} I & 0 \\ -X_\infty & I \end{bmatrix} \begin{bmatrix} A & -BB^{\mathsf{T}} \\ -C^{\mathsf{T}}C & -A^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I & 0 \\ X_\infty & I \end{bmatrix} = \begin{bmatrix} \hat{A} & -BB^{\mathsf{T}} \\ 0 & -\hat{A}^{\mathsf{T}} \end{bmatrix},$$

and

$$T^{-1}S^{-1}\mathcal{H}ST := \begin{bmatrix} I & X_{\mathrm{L}} \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{A} & -BB^{\mathsf{T}} \\ 0 & -\hat{A}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I & -X_{\mathrm{L}} \\ 0 & I \end{bmatrix} = \begin{bmatrix} \hat{A} & 0 \\ 0 & -\hat{A}^{\mathsf{T}} \end{bmatrix} =: \hat{\mathcal{H}}.$$

We use Theorem 2.21 and get

$$\begin{aligned} e^{-t\mathcal{H}} &= e^{-tST\hat{\mathcal{H}}T^{-1}S^{-1}} = STe^{-t\hat{\mathcal{H}}}T^{-1}S^{-1} \\ &= \begin{bmatrix} I & -X_{\mathrm{L}} \\ X_\infty & I - X_\infty X_{\mathrm{L}} \end{bmatrix} \begin{bmatrix} e^{-t\hat{A}} & 0 \\ 0 & e^{t\hat{A}^{\mathsf{T}}} \end{bmatrix} \begin{bmatrix} I - X_{\mathrm{L}}X_\infty & X_{\mathrm{L}} \\ -X_\infty & I \end{bmatrix} \\ &= \begin{bmatrix} e^{-t\hat{A}}(I - X_{\mathrm{L}}X_\infty) + X_{\mathrm{L}}e^{t\hat{A}}X_\infty & e^{-t\hat{A}}X_{\mathrm{L}} - X_{\mathrm{L}}e^{t\hat{A}^{\mathsf{T}}} \\ X_\infty e^{-t\hat{A}}(I - X_{\mathrm{L}}X_\infty) - (I - X_\infty X_{\mathrm{L}})e^{t\hat{A}}X_\infty & X_\infty e^{-t\hat{A}}X_{\mathrm{L}} + (I - X_\infty X_{\mathrm{L}})e^{t\hat{A}} \end{bmatrix}. \end{aligned} \tag{6.14}$$

With that, we get

$$\begin{bmatrix} U(t) \\ V(t) \end{bmatrix} = e^{-t\mathcal{H}} \begin{bmatrix} I \\ X_0 \end{bmatrix} = \begin{bmatrix} e^{-t\hat{A}}(I - X_{\mathrm{L}}(X_\infty - X_0)) + X_{\mathrm{L}} e^{t\hat{A}^\mathsf{T}}(X_\infty - X_0) \\ X_\infty e^{-t\hat{A}}(I - X_{\mathrm{L}}(X_\infty - X_0)) - (X_\infty X_{\mathrm{L}} + I) e^{t\hat{A}^\mathsf{T}}(X_\infty - X_0) \end{bmatrix}. \tag{6.15}$$

Now, we observe that

$$\begin{aligned} U(t) &= e^{-t\hat{A}}\Big(I - \Big(X_{\mathrm{L}} - e^{t\hat{A}} X_{\mathrm{L}} e^{t\hat{A}^\mathsf{T}}\Big)(X_\infty - X_0)\Big), \\ V(t) &= X_\infty e^{-t\hat{A}}\Big(I - \Big(X_{\mathrm{L}} - e^{t\hat{A}} X_{\mathrm{L}} e^{t\hat{A}^\mathsf{T}}\Big)(X_\infty - X_0)\Big) - e^{t\hat{A}^\mathsf{T}}(X_\infty - X_0) \\ &= X_\infty U(t) - e^{t\hat{A}^\mathsf{T}}(X_\infty - X_0). \end{aligned} \tag{6.16}$$

As the matrices $U(t)$ and $e^{-t\hat{A}}$ are nonsingular, the matrix in brackets (Equation (6.16)) is nonsingular as well. Therefore,

$$\begin{aligned} X(t) &= V(t)U(t)^{-1} \\ &= X_\infty - e^{t\hat{A}^\mathsf{T}}(X_\infty - X_0)\Big(I - \Big(X_{\mathrm{L}} - e^{t\hat{A}} X_{\mathrm{L}} e^{t\hat{A}^\mathsf{T}}\Big)(X_\infty - X_0)\Big)^{-1} e^{t\hat{A}}. \qquad \square \end{aligned}$$

From the proof of Theorem 6.15, it is clear that the property that X_∞ is a stabilizing solution of the ARE (5.1) can be weakened to

$$\Lambda(\hat{A}) \cap \Lambda(-\hat{A}^\mathsf{T}) = \varnothing.$$

With this condition, the ALE (6.13) has a unique solution. Therefore, it is sufficient to assume that $X_\infty \in \mathbb{R}^{n\times n}$ is a symmetric and *strictly unmixed* solution [1, Thm. 4.2.1]. In [5, Ch. 15.4], one can find another solution formula which holds under more restrictive assumptions. A solution formula based on the Jordan canonical form of the Hamiltonian matrix $\mathcal{H}$ is given in [1, Thm. 3.2.1]. Other solution representations can be found in [101, 109].

6.4 Davison–Maki and Modified Davison–Maki Method

In this section we review the Davison–Maki and modified Davison–Maki method for the numerical solution of the NDRE (6.1)

$$\begin{aligned} \dot{X}(t) &= M_{22}X(t) - X(t)M_{11} - X(t)M_{12}X(t) + M_{21}, \\ X(0) &= M_0. \end{aligned}$$

Both methods are based on Radon's Lemma (Theorem 6.2) and on the linear IVP (6.4)

$$\begin{bmatrix} \dot{U}(t) \\ \dot{V}(t) \end{bmatrix} = M \begin{bmatrix} U(t) \\ V(t) \end{bmatrix} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} U(t) \\ V(t) \end{bmatrix}, \quad \begin{bmatrix} U(0) \\ V(0) \end{bmatrix} = \begin{bmatrix} I_n \\ M_0 \end{bmatrix}.$$

6.4.1 Davison–Maki Method

The Davison–Maki method for the NDRE (6.1) was proposed in [37]. The method is based on first computing the matrix exponential e^{hM} for a given step size $h > 0$. According to Radon's Lemma (Theorem 6.2), we have that

$$\begin{bmatrix} U(h) \\ V(h) \end{bmatrix} = e^{hM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix}, \quad X(h) = V(h)U(h)^{-1}.$$

The next step is then to make use of the semigroup property of the matrix exponential (Theorem 2.21 (ii)):

$$\begin{bmatrix} U(kh) \\ V(kh) \end{bmatrix} = e^{khM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix} = \left(e^{hM}\right)^k \begin{bmatrix} I_n \\ M_0 \end{bmatrix}, \quad X(kh) = V(kh)U(kh)^{-1}.$$

Another variant of the Davison–Maki method updates U and V instead of the matrix exponential. The variant follows from

$$\begin{bmatrix} U(kh) \\ V(kh) \end{bmatrix} = e^{khM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix} = e^{hM} e^{(k-1)hM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix} = e^{hM} \begin{bmatrix} U((k-1)h) \\ V((k-1)h) \end{bmatrix}.$$

Both variants of the method are given in Algorithms 6.1 and 6.2.

6.4.2 Modified Davison–Maki Method

When the Davison–Maki method (Algorithms 6.1 and 6.2) is applied to the DRE (6.8), usually numerical instabilities occur because each block of the matrix exponential of the Hamiltonian matrix $e^{-t\mathcal{H}}$ as well as $U(t)$ and $V(t)$ contains the matrix $e^{-t\hat{A}}$; cf. Equations (6.14) and (6.15). Since $\hat{A} = A - BB^{\mathsf{T}}X_\infty$ is stable, the matrix exponential of $-t\hat{A}$ exhibits exponential growth, which becomes problematic for large t. The occurrence of these numerical problems with the Davison–Maki method was also pointed out in [35, 64, 81, 122]. Another reason is that the spectrum of a real Hamiltonian matrix comes in quadruples; cf. Lemma 2.44 (ii). Therefore, the spectrum of the Hamiltonian matrix usually contains eigenvalues with positive real part, and its matrix exponential grows for large times.

A suitable modification of the Davison–Maki method was proposed in [64]. Nevertheless, the modified method originates back to [63, p. 9].

By Radon's Lemma (Theorem 6.2), we have the identities

$$\begin{bmatrix} \tilde{U} \\ \tilde{V} \end{bmatrix} := e^{-hM} \begin{bmatrix} I_n \\ X((k-1)h) \end{bmatrix}, \quad X(kh) = \tilde{V}\tilde{U}^{-1}.$$

The modified Davison–Maki method is given in Algorithm 6.3.

Any computationally efficient norm can be used for the matrix exponential in Algorithm 6.3 line 3. The modified Davison–Maki method is also more efficient than

Algorithm 6.1: Davison–Maki Method I for the NDRE (6.1) [37].

```
   Input: real matrices $M_0$ and $M_{ij}$ as in Equation (6.1), a step size $h > 0$, and a
          final time $t_f > 0$
   Assumptions: The solution $X$ of the NDRE (6.1) exists on the interval $[0, t_f)$.
   Output: matrices $X_0, \dots, X_k$ such that $X(kh) = X_k$ for $k \in \mathbb{N}_0$ and $kh < t_f$
   % initialization:
 1 $X_0 := M_0; \quad k := 1;$
   % compute matrix exponential:
 2 $\Theta_h := \exp\left( h \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \right);$
   variant with matrix exponential update:
 3 $\Theta := \Theta_h;$
 4 while $kh < t_f$ do
 5     partition $\begin{array}{c|cc} & m & n \\ \hline m & \Theta_{11} & \Theta_{12} \\ n & \Theta_{12} & \Theta_{22} \end{array} = \Theta;$
 6     $U_{\mathtt{dm}} := \Theta_{11} + \Theta_{12} M_0;$
 7     $V_{\mathtt{dm}} := \Theta_{21} + \Theta_{22} M_0;$
 8     $X_k := V_{\mathtt{dm}} U_{\mathtt{dm}}^{-1};$
 9     $\Theta := \Theta \Theta_h;$
10     $k := k + 1;$
11 return $X_0, \dots, X_k$;
```

the Davison–Maki method in both variants because fewer matrix-matrix products are needed per time step; cf. Algorithm 6.1 lines 4–10 and Algorithm 6.2 lines 6–10 with Algorithm 6.3 lines 6–10.

A decrease of the step size $h > 0$ does not improve the accuracy in general because, in theory, the exact values of $U(kh)$ and $V(kh)$ are computed with the matrix exponential. In practice, the accuracy is determined by the accuracy of the matrix exponentiation, the repeated multiplication by the exponential, and the involved matrix inversion.

For the realization in a simulation, the following considerations can be made. The step size should not be chosen arbitrarily large as the matrix exponential may become too large in the norm, leading to cancellation errors. Thus, we suggest computing the norm of the matrix exponential before the iteration starts. If the norm is too large, then the step size has to be decreased. On the other hand, a small time step means more multiplications with the matrix exponential and possibly accumulating rounding errors.

Algorithm 6.2: Davison–Maki Method II for the NDRE (6.1) [37].

```
Input: real matrices $M_0$ and $M_{ij}$ as in Equation (6.1), a step size $h > 0$, and a
       final time $t_f > 0$
Assumptions: The solution $X$ of the NDRE (6.1) exists on the interval $[0, t_f)$.
Output: matrices $X_0, \dots, X_k$ such that $X(kh) = X_k$ for $k \in \mathbb{N}_0$ and $kh < t_f$
% initialization:
$X_0 := M_0$;   $k := 1$;
% compute matrix exponential:
$\Theta_h := \exp\left(h \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}\right)$;
variant with updating $U$ and $V$:
$U_{\mathtt{dm}} := I_m$;
$V_{\mathtt{dm}} := M_0$;
partition $\begin{array}{c|cc} & m & n \\ \hline m & \Theta_{11} & \Theta_{12} \\ n & \Theta_{12} & \Theta_{22} \end{array} := \Theta$;
while $kh < t_f$ do
    $U_{\mathtt{dm}} := \Theta_{11} U_{\mathtt{dm}} + \Theta_{12} V_{\mathtt{dm}}$;
    $V_{\mathtt{dm}} := \Theta_{21} U_{\mathtt{dm}} + \Theta_{22} V_{\mathtt{dm}}$;
    $X_k := V_{\mathtt{dm}} U_{\mathtt{dm}}^{-1}$;
    $k := k + 1$;
return $X_0, \dots, X_k$;
```

In the k-th iteration of modified Davison–Maki method (Algorithm 6.3), we have

$$\begin{bmatrix} U_{\mathtt{mod\_dm}} \\ V_{\mathtt{mod\_dm}} \end{bmatrix} = e^{hM} \begin{bmatrix} I_n \\ X((k-1)h) \end{bmatrix} = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} \begin{bmatrix} I_n \\ X((k-1)h) \end{bmatrix},$$

and the norm of the iterates can be bounded by

$$\begin{aligned} \|U_{\mathtt{mod\_dm}}\|_2 &\le \|\Theta_{11}\|_2 + \|\Theta_{12}\|_2 \, \|X((k-1)h)\|_2 \,, \\ \|V_{\mathtt{mod\_dm}}\|_2 &\le \|\Theta_{21}\|_2 + \|\Theta_{22}\|_2 \, \|X((k-1)h)\|_2 \,. \end{aligned}$$

For small step sizes of $h > 0$, it holds $e^{hM} \approx I_{n+m} + hM$ and

$$\Theta_{11} \approx I_n + hM_{11}, \; \Theta_{12} \approx hM_{12}, \; \Theta_{21} \approx hM_{21}, \text{ and } \Theta_{22} \approx I_m + hM_{22}.$$

Therefore, for small enough step size and moderate norm of the solution $\|X(t)\|_2$, the norm of the iterates cannot grow heavily in contrast to the Davison–Maki method (Algorithms 6.1 and 6.2). Assuming that the matrix exponential in line 2 of Algorithm 6.3

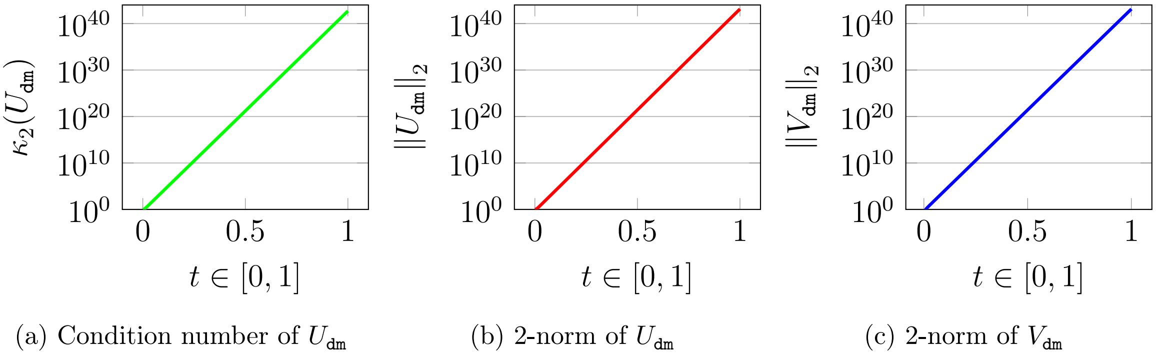
Algorithm 6.3: Modified Davison–Maki Method for the NDRE (6.1) [63, 64].

```
Input: real matrices $M_0$ and $M_{ij}$ as in Equation (6.1), a step size $h > 0$, a final
       time $t_f > 0$, and a moderate large number $tol_{\texttt{exp}} > 0$
Assumptions: The solution $X$ of the NDRE (6.1) exists on the interval $[0, t_f)$.
Output: matrices $X_0, \ldots, X_k$ such that $X(kh) = X_k$ for $k \in \mathbb{N}_0$ and $kh < t_f$

% initialization:
$X_0 := M_0; \quad k := 1;$
% compute matrix exponential:
$\Theta := \exp\left( h \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \right);$
% check the norm of the matrix exponential:
if $\|\Theta\|_1 > tol_{exp}$ then
    error (“1-norm of the matrix exponential is too large.”);
partition $\begin{array}{cc} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12} & \Theta_{22} \end{bmatrix} \end{array} = \Theta;$
while $kh < t_f$ do
    $U_{\texttt{mod\_dm}} := \Theta_{11} + \Theta_{12} X_{k-1};$
    $V_{\texttt{mod\_dm}} := \Theta_{21} + \Theta_{22} X_{k-1};$
    $X_k := V_{\texttt{mod\_dm}} U_{\texttt{mod\_dm}}^{-1};$
    $k := k + 1;$
return $X_0, \ldots, X_k;$
```

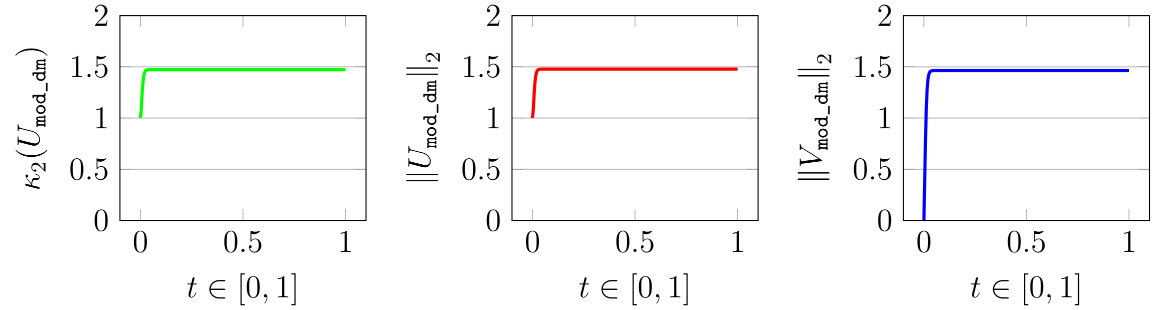
was approximated using the scaling and squaring method, then the intermediates of the squaring phase can be used, so that the matrix exponential is not recomputed from scratch.

Example 6.2 (Davison–Maki Method: Exponential Growth):
We applied the Davison–Maki method (Algorithm 6.1) with step size $h = 2^{-8}$ to a DRE defined by the matrices of the `TRIDIAG`(α) example with $\alpha = 5$ (Table A.1). We have used the variable-precision arithmetic `vpa` of MATLAB with 512 significant digits. We plot the 2-norm of the iterates $U_{\texttt{dm}}$ and $V_{\texttt{dm}}$ as well as the 2-norm condition number of $U_{\texttt{dm}}$ on the interval $[0, 1]$. Figure 6.2 shows that all quantities grow exponentially over time. Therefore, eventually, either a floating-point overflow will occur, or the matrix inversion ceases to be executed accurately. Figure 6.3 shows the same quantities for the iterates $U_{\texttt{mod\_dm}}$ and $V_{\texttt{mod\_dm}}$ of the modified Davison–Maki method (Algorithm 6.3).



(a) Condition number of $U_{\texttt{dm}}$ (b) 2-norm of $U_{\texttt{dm}}$ (c) 2-norm of $V_{\texttt{dm}}$

Fig. 6.2: Davison–Maki Method and the Growth of $U_{\texttt{dm}}$ and $V_{\texttt{dm}}$.



(a) Condition number of $U_{\texttt{mod\_dm}}$ (b) 2-norm of $U_{\texttt{mod\_dm}}$ (c) 2-norm of $V_{\texttt{mod\_dm}}$

Fig. 6.3: Modified Davison–Maki Method and the Growth of $U_{\texttt{mod\_dm}}$ and $V_{\texttt{mod\_dm}}$.

If a symmetric solution is expected, then line 9 of Algorithm 6.3 should be altered to

$$X_k := \tfrac{1}{2}\left(X_k + X_k^{\mathsf{T}}\right)$$

because due to numerical errors, the symmetry will be lost after some iterations; cf. Theorem 2.16. Similar, if a spsd solution is expected, one may compute the nearest spsd matrix. However, usually, this involves the computation an expensive polar decomposition; cf. Section 2.4.3 and Theorem 2.18.

6.5 Galerkin Approach

In this section, we develop a Galerkin approach approach for the numerical approximation of the solution of the DRE (6.8a)

$$\begin{aligned}\dot{X}(t) &= A^{\mathsf{T}} X(t) + X(t)A - X(t)BB^{\mathsf{T}}X(t) + C^{\mathsf{T}}C,\\ X(0) &= 0\end{aligned}$$

with zero initial conditions.

According to Theorem 6.5, the solution X is monotonically non-decreasing for all $t \in (t_{-}, \infty)$ and spsd for all $t \geq 0$. Moreover, we assume that the corresponding ARE has a unique stabilizing solution $X_{\infty} \in \mathbb{R}^{n\times n}$, and that $X(t)$ converges to X_{∞}; cf. Theorem 6.14. By combining these results, we get the inequality

$$0 \preccurlyeq X(t) \preccurlyeq X_{\infty} \tag{6.17}$$

for all $t \geq 0$. Theorem 2.4 (iv) gives that the k-largest eigenvalue $t \mapsto \lambda_k^{\downarrow}(X(t))$ is monotonically non-decreasing function, and that the inequality

$$0 \leq \lambda_k^{\downarrow}(X(t)) \leq \lambda_k^{\downarrow}(X_{\infty}),$$

holds for all $t \geq 0$. Therefore, the number of eigenvalues of $X(t)$ greater than or equal to a given threshold is monotonically non-decreasing over time.

Example 6.3 (Eigenvalue Decay):
We illustrate this by an example in Figure 6.4. We have chosen the matrices of the `TRIDIAG`(α) benchmark problem with $\alpha = 5$ (Table A.1). We have used the variable-precision arithmetic of MATLAB with 512 significant digits and the modified Davison–Maki method with step size $h = 2^{-5}$ for the numerical approximation. The eigenvalues of $X(t)$ are plotted for $t \in \{0.5, 1, \ldots, 15\}$. The functions $t \mapsto \lambda_k^{\downarrow}(X(t))$ are highlighted in green for $k \in \{10, 20, 30, 40, 50\}$. All data below 10^{-60} were truncated. The shadowed green plane is drawn at the level of machine precision $\varepsilon_{\mathtt{mach}}$.

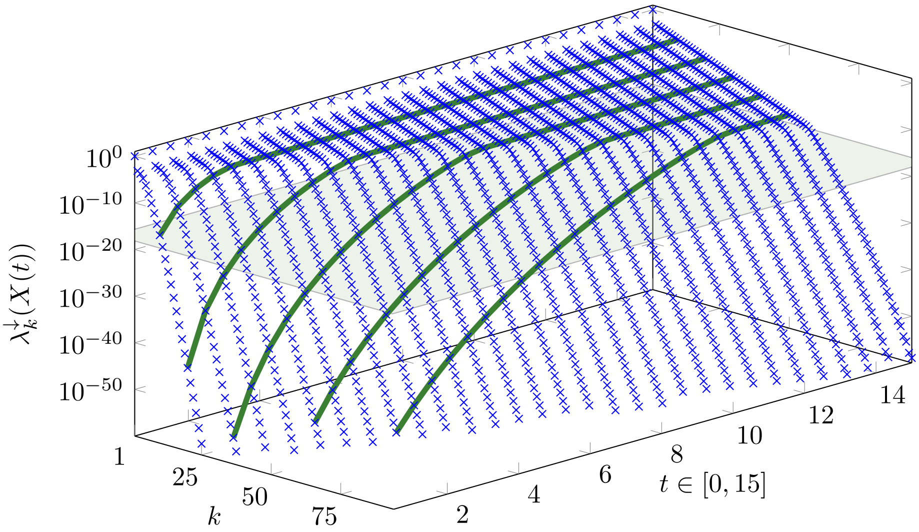


Fig. 6.4: Eigenvalues $\lambda_k^{\downarrow}(X(t))$ of the Numerical Approximation of $X(t)$.

Furthermore, if $X(t)$ solves the DRE (6.8) then $X(t)$ solves the ALE

$$A^{\mathsf{T}}X(t) + X(t)A = \dot{X}(t) + X(t)BB^{\mathsf{T}}X(t) - C^{\mathsf{T}}C.$$

According to Lemma 6.4, the derivative $\dot{X}(t)$ has constant rank, i.e.,

$$\operatorname{rank}(\dot{X}(t)) = \operatorname{rank}(\dot{X}(0)) = \operatorname{rank}(A^{\mathsf{T}}Z_0Z_0^{\mathsf{T}} + Z_0Z_0^{\mathsf{T}}A - Z_0Z_0^{\mathsf{T}}BB^{\mathsf{T}}Z_0Z_0^{\mathsf{T}} + C^{\mathsf{T}}C).$$

Therefore, if Z_0 and B_0 have only a few columns and C has only a few rows then the rank of the right-hand side

$$\dot{X}(t) + X(t)BB^{\mathsf{T}}X(t) - C^{\mathsf{T}}C$$

is much smaller than the state-space dimension n and the bounds on the singular value decay or eigenvalue decay presented in Section 3.5 apply.

6.5.1 Projection Error and Eigenvalue Decay

We consider a spectral decomposition of the stabilizing solution of the ARE (5.1) of the form

$$X_\infty = Q_\infty D_\infty Q_\infty^{\mathsf{T}}, \tag{6.18a}$$
$$Q_\infty = \begin{bmatrix} q_1, \ldots, q_n \end{bmatrix} \in \mathbb{R}^{n\times n}, \tag{6.18b}$$
$$D_\infty = \operatorname{diag}(\lambda_1^{\downarrow}(X_\infty), \ldots, \lambda_n^{\downarrow}(X_\infty)) \in \mathbb{R}^{n\times n}, \tag{6.18c}$$

where $q_1 \ldots, q_n \in \mathbb{R}^n$ is a system of orthonormal eigenvectors of X_∞ corresponding to the eigenvalues $\lambda_1^{\downarrow}(X_\infty), \ldots, \lambda_n^{\downarrow}(X_\infty)$. We represent the solution of the DRE as

$$X(t) = Q_\infty Q_\infty^{\mathsf{T}} X(t) Q_\infty Q_\infty^{\mathsf{T}}.$$

This representation has the advantage that the absolute value of the entries of the matrix

$$Q_\infty^\mathsf{T} X(t) Q_\infty = \left(q_i^\mathsf{T} X(t) q_j\right)_{\substack{i=1,\ldots,n\\ j=1,\ldots,n}}$$

can be bounded in terms of the eigenvalues $\lambda_1^\downarrow(X_\infty), \ldots, \lambda_n^\downarrow(X_\infty)$.

Lemma 6.16:
Assume that (A, B) is stabilizable and (A, C) is detectable. Moreover, let $X_\infty \in \mathbb{R}^{n\times n}$ be the stabilizing solution of the ARE (5.1) and $q_1, \ldots, q_n \in \mathbb{R}^n$ be a system of orthonormal eigenvectors of X_∞ corresponding to the eigenvalues $\lambda_1^\downarrow(X_\infty), \ldots, \lambda_n^\downarrow(X_\infty)$ as in (6.18). Then the bounds

$$\left|q_i^\mathsf{T} X(t) q_j\right| \leq \begin{cases} \sqrt{\left(\lambda_i^\downarrow(X_\infty) - q_i^\mathsf{T} X(t) q_i\right)\left(\lambda_j^\downarrow(X_\infty) - q_j^\mathsf{T} X(t) q_j\right)} & \text{if } i \neq j, \\ \lambda_i^\downarrow(X_\infty) & \text{if } i = j, \end{cases} \tag{6.19}$$

$$\left|q_i^\mathsf{T} X(t) q_j\right| \leq \sqrt{\lambda_i^\downarrow(X_\infty)\lambda_j^\downarrow(X_\infty)} \tag{6.20}$$

hold for all $t \geq 0$, where X is the unique solution of the DRE (6.8a) with $X(0) = 0$. ◊

Proof. As already discussed, Inequality (6.17)

$$0 \preccurlyeq X(t) \preccurlyeq X_\infty$$

holds for all $t \geq 0$; cf. Theorems 6.5 and 6.14. This implies $0 \preccurlyeq X_\infty - X(t)$. Multiplying the latter inequality with Q_∞^T from the left and with Q_∞ from the right gives

$$0 \preccurlyeq Q_\infty^\mathsf{T}(X_\infty - X(t))Q_\infty = D_\infty - Q_\infty^\mathsf{T} X(t) Q_\infty.$$

The matrix D_∞ is diagonal and contains the eigenvalues on its diagonal. Theorem 2.4 (vi) yields

$$\left|q_i^\mathsf{T} X(t) q_j\right| \leq \sqrt{\left(\lambda_i^\downarrow(X_\infty) - q_i^\mathsf{T} X(t) q_i\right)\left(\lambda_j^\downarrow(X_\infty) - q_j^\mathsf{T} X(t) q_j\right)}.$$

for all $i \neq j$. Again, Inequality (6.17) implies

$$0 \leq q_i^\mathsf{T} X(t) q_i \leq q_i^\mathsf{T} X_\infty q_i = \lambda_i^\downarrow(X_\infty)$$

and the claim follows. □

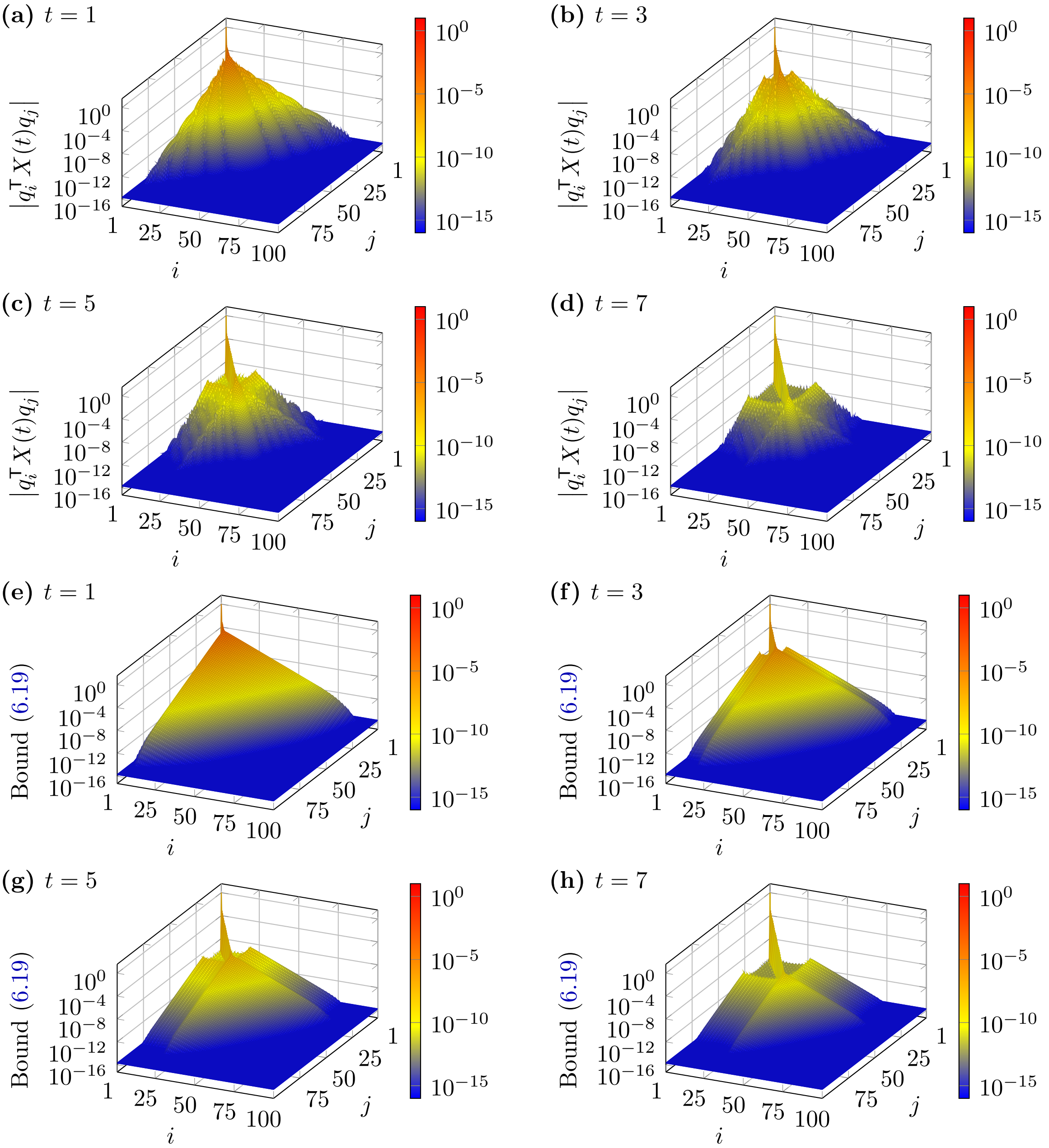
Example 6.4 (Decay of Absolute Values and Eigenvalue Decay):
We illustrate the decay of $\left|q_i^\mathsf{T} X(t) q_j\right|$ and the Bounds (6.19) and (6.20) of Lemma 6.16. We have chosen the matrices of the `TRIDIAG(`α`)` example with $\alpha = 5$ (Table A.1). To improve the visualization, all values in Figures 6.5a–6.5h and 6.5j below machine precision were set to machine precision $\varepsilon_{\texttt{mach}}$. Figures 6.5a–6.5d show the absolute value of the entries of

$$Q_\infty^\mathsf{T} X(t) Q_\infty = \left(q_i^\mathsf{T} X(t) q_j\right)_{\substack{i=1,\ldots,n\\ j=1,\ldots,n}}$$

for $t \in \{1, 3, 5, 7\}$. Figures 6.5e–6.5h show the time-dependent Bound (6.19). The eigenvalue decay of the stabilizing solution X_∞ of the corresponding ARE is shown in Figure 6.5i. Figure 6.5j visualizes the time-independent Bound (6.20). The eigenvalues of X_∞ are sorted in a non-increasing fashion, therefore, the quantity

$$\sqrt{\lambda_i^{\downarrow}(X_\infty)\lambda_j^{\downarrow}(X_\infty)} \qquad \diamondsuit$$

is monotonically non-increasing in i and j. From a geometric point of view, the surface of Figure 6.5j is on top of the surfaces of Figures 6.5a–6.5d. The same holds true for Figures 6.5e–6.5h and Figures 6.5a–6.5d, respectively.



(a) $t = 1$

(b) $t = 3$

(c) $t = 5$

(d) $t = 7$

(e) $t = 1$

(f) $t = 3$

(g) $t = 5$

(h) $t = 7$

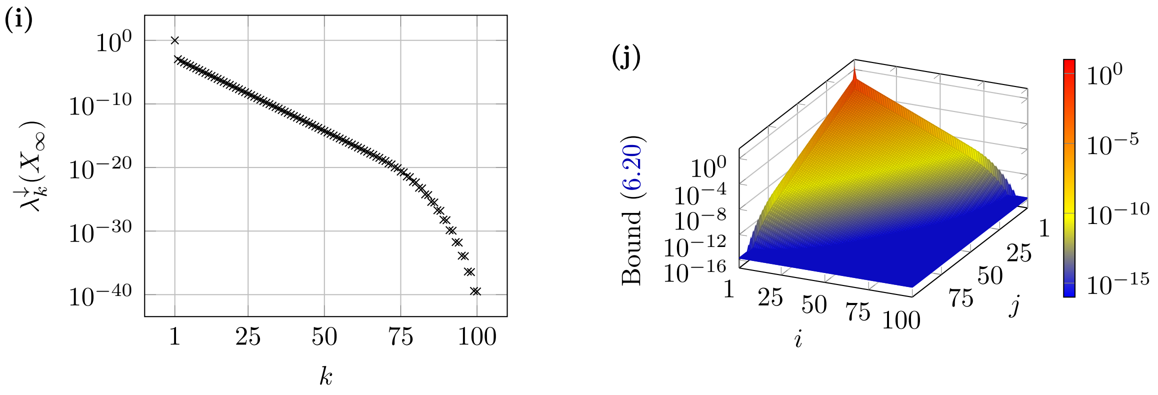


Fig. 6.5: **(a)–(d)** and **(e)–(h)** Decay of $\left|q_i^{\mathsf{T}} X(t) q_j\right|$ and Bound (6.19) for $t \in \{1,3,5,7\}$. **(i)** The Eigenvalue Decay of X_∞. **(j)** Bound (6.20).

Projection Error in the Frobenius Norm

For quick enough eigenvalue decay, we expect that

$$\left|q_i^{\mathsf{T}} X(t) q_j\right| \leq \sqrt{\lambda_i^{\downarrow}(X_\infty)\lambda_j^{\downarrow}(X_\infty)} \approx 0$$

for i or j large enough. We have

$$X(t) = Q_\infty Q_\infty^{\mathsf{T}} X(t) Q_\infty Q_\infty^{\mathsf{T}} = Q_\infty \left(q_i^{\mathsf{T}} X(t) q_j\right)_{\substack{i=1,\ldots,n\\ j=1,\ldots,n}} Q_\infty^{\mathsf{T}} = \sum_{i,j=1}^{n} \left(q_i^{\mathsf{T}} X(t) q_j\right) q_i q_j^{\mathsf{T}}.$$

Next, we truncate the parts of the sum with possibly small contribution to the solution and get

$$X(t) \approx \sum_{i,j=1}^{k} \left(q_i^{\mathsf{T}} X(t) q_j\right) q_i q_j^{\mathsf{T}} = Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}} X(t) Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}},$$

where $Q_{\infty,k} := \left[q_1, \ldots, q_k\right] \in \mathbb{R}^{n\times k}$. We consider the corresponding linear space

$$\mathcal{Q}_k := \left\{ Q_{\infty,k} Y Q_{\infty,k}^{\mathsf{T}} \mid Y \in \mathbb{R}^{k\times k} \right\}$$

together with

$$\mathcal{P}_k \colon \mathbb{R}^{n\times n} \to \mathcal{Q}_k,\ \mathcal{P}_k X = Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}} X Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}}. \tag{6.21}$$

Using that the matrix $Q_{\infty,k}$ has orthonormal columns, it can be verified that

$$\mathcal{P}_k^2 X = \mathcal{P}_k X \text{ and } \left\langle X - \mathcal{P}_k X, Q_{\infty,k} Y Q_{\infty,k}^{\mathsf{T}} \right\rangle_{\mathrm{F}} = 0$$

holds. Hence, $\mathcal{P}_k$ is the orthogonal projection onto $\mathcal{Q}_k$ and $\mathcal{P}_k(X(t))$ is the best approximation of $X(t)$ in $\mathcal{Q}_k$; cf. Section 2.3.3. The matrices $\left(q_i q_j^{\mathsf{T}}\right)_{i,j=1,\ldots,n}$ form an orthonormal basis of the Hilbert space $(\mathbb{R}^{n\times n}, \langle\cdot,\cdot\rangle_{\mathrm{F}})$. In this setting, it is more convenient to express the coefficients using the inner product, e.g.,

$$q_i^{\mathsf{T}} X(t) q_j = \left\langle X(t), q_i q_j^{\mathsf{T}} \right\rangle_{\mathrm{F}}.$$

The projection error in Frobenius norm can be bounded using Parseval's identity and Bound (6.20)

$$\begin{aligned}
\|X(t) - \mathcal{P}_k X(t)\|_{\mathrm{F}} &= \left\| \sum_{i,j=1}^{n} \left(q_i^{\mathsf{T}} X(t) q_j\right) q_i q_j^{\mathsf{T}} - \sum_{i,j=1}^{k} \left(q_i^{\mathsf{T}} X(t) q_j\right) q_i q_j^{\mathsf{T}} \right\|_{\mathrm{F}} \\
&= \left\| \sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{n} \left(q_i^{\mathsf{T}} X(t) q_j\right) q_i q_j^{\mathsf{T}} \right\|_{\mathrm{F}} = \left\| \sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{n} \left\langle X(t), q_i q_j^{\mathsf{T}} \right\rangle_{\mathrm{F}} q_i q_j^{\mathsf{T}} \right\|_{\mathrm{F}} \\
&= \sqrt{\sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{n} \left| q_i^{\mathsf{T}} X(t) q_j \right|^2} \overset{(6.20)}{\leq} \sqrt{\sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{n} \lambda_i^{\downarrow}(X_\infty) \lambda_j^{\downarrow}(X_\infty)}.
\end{aligned} \tag{6.22}$$

Projection Error in the 2-Norm

We measure the projection error in the 2-norm to simplify Bound (6.22).

Theorem 6.17 (Projection Error in the 2-norm):
Assume that (A, B) is stabilizable and (A, C) is detectable. Moreover, let $X_\infty \in \mathbb{R}^{n\times n}$ be the stabilizing solution of the ARE (5.1). Then for all $k = 1, \ldots, n-1$ and all $t \geq 0$, the projection error is bounded by

$$\|X(t) - \mathcal{P}_k X(t)\|_2 \leq 2\sqrt{\lambda_{k+1}^{\downarrow}(X_\infty)\, \|X(t)\|_2} \leq 2\sqrt{\lambda_{k+1}^{\downarrow}(X_\infty) \lambda_1^{\downarrow}(X_\infty)}, \tag{6.23}$$

where X is the unique solution of the DRE (6.8a) with $X(0) = 0$, and $\mathcal{P}_k$ is the orthogonal projection (6.21). ◊

Proof. Again, we utilize Inequality (6.17) $0 \preccurlyeq X(t) \preccurlyeq X_\infty$. We use Theorem 2.4 (ii) and get

$$X_\infty = \mathcal{P}_k X_\infty + X_\infty - \mathcal{P}_k X_\infty \preccurlyeq \mathcal{P}_k X_\infty + \|X_\infty - \mathcal{P}_k X_\infty\|_2 I = \mathcal{P}_k X_\infty + \lambda_{k+1}^{\downarrow}(X_\infty) I.$$

We combine the inequalities and get

$$0 \preccurlyeq X(t) \preccurlyeq \mathcal{P}_k X_\infty + \lambda_{k+1}^{\downarrow}(X_\infty) I = Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}} X_\infty Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}} + \lambda_{k+1}^{\downarrow}(X_\infty) I. \tag{6.24}$$

The projection matrix $I - Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}}$ is symmetric, and $\left(I - Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}}\right) Q_{\infty,k}$ vanishes. We multiply Inequality (6.24) from the left and right with $I - Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}}$ and get

$$0 \preccurlyeq \left(I - Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}}\right) X(t) \left(I - Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}}\right) \preccurlyeq \lambda_{k+1}^{\downarrow}(X_\infty) \left(I - Q_{\infty,k} Q_{\infty,k}^{\mathsf{T}}\right); \tag{6.25}$$

cf. Theorem 2.4 (iii). Inequality (6.25) and Theorem 2.4 (v) give

$$\begin{aligned}\left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)X(t)^{1/2}\right\|_2^2 &= \left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)X(t)\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)\right\|_2 \\ &\leq \left\|\lambda_{k+1}^{\downarrow}(X_\infty)\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)\right\|_2 \\ &= \lambda_{k+1}^{\downarrow}(X_\infty).\end{aligned}$$

We get

$$\begin{aligned}\|X(t) - \mathcal{P}_k X(t)\|_2 &= \left\|X(t) - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}X(t)Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right\|_2 \\ &= \left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)X(t) + Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}X(t)\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)\right\|_2 \\ &\leq 2\left\|\left(I - Q_{\infty,k}Q_{\infty,k}^{\mathsf{T}}\right)X(t)^{1/2}\right\|_2\left\|X(t)^{1/2}\right\|_2 \\ &\leq 2\sqrt{\lambda_{k+1}^{\downarrow}(X_\infty)\,\|X(t)\|_2}.\end{aligned}$$

Inequality (6.17) and Theorem 2.4 (ii) yield $\|X(t)\|_2 \leq \|X_\infty\|_2 = \lambda_1^{\downarrow}(X_\infty)$, and the proof is complete. □

Example 6.5 (Projection Error in the 2-norm):
We illustrate the projection error and Bound (6.23) of Theorem 6.17 by an example. We have chosen the matrices of the TRIDIAG(α) example with $\alpha = 5$ (Table A.1). Figures 6.6a–6.6c show the projection error and Bound (6.23) for $k \in \{50, 60, 70\}$.

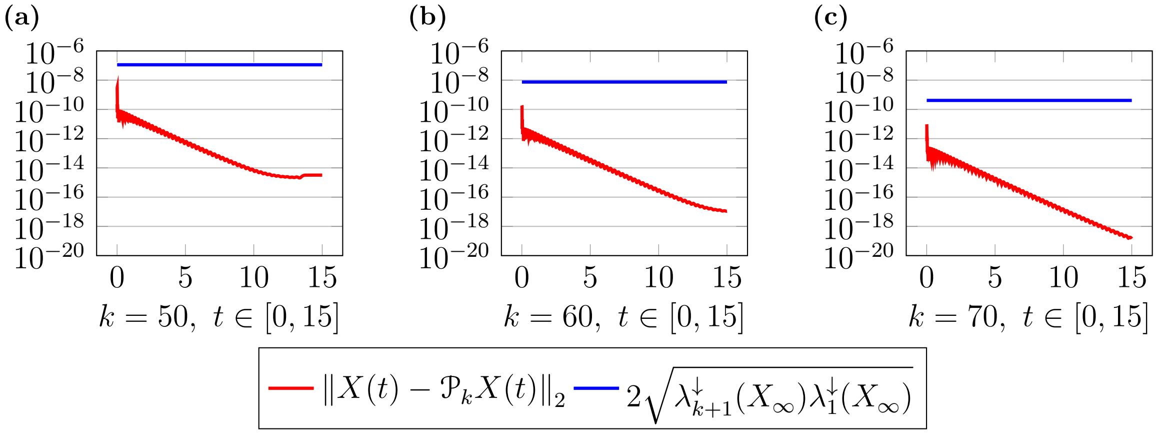


Fig. 6.6: **(a)–(c)** The Projection Error and Bound (6.23).

6.5.2 Solution Formula and Matrix Exponential

According to Theorem 6.15, the solution of the DRE (6.8a)

$$\begin{aligned}&\dot{X}(t) = A^{\mathsf{T}}X(t) + X(t)A - X(t)BB^{\mathsf{T}}X(t) + C^{\mathsf{T}}C, \\ &X(0) = 0\end{aligned}$$

with zero initial conditions is given by

$$X(t) = X_\infty - e^{t\hat{A}^\mathsf{T}} X_\infty \Big(I - \Big(X_\mathrm{L} - e^{t\hat{A}} X_\mathrm{L} e^{t\hat{A}^\mathsf{T}} \Big) X_\infty \Big)^{-1} e^{t\hat{A}},$$

where X_∞ is the stabilizing solution of the ARE, $\hat{A} = A - BB^\mathsf{T} X_\infty$, and X_L is the solution of the ALE (6.13). The identity $(I - P(t))^{-1} = I + (I - P(t))^{-1} P(t)$ leads to

$$\begin{aligned} X(t) &= X_\infty - e^{t\hat{A}^\mathsf{T}} X_\infty e^{t\hat{A}} \\ &\quad - e^{t\hat{A}^\mathsf{T}} X_\infty \Big(I - \Big(X_\mathrm{L} - e^{t\hat{A}} X_\mathrm{L} e^{t\hat{A}^\mathsf{T}} \Big) X_\infty \Big)^{-1} \Big(X_\mathrm{L} - e^{t\hat{A}} X_\mathrm{L} e^{t\hat{A}^\mathsf{T}} \Big) X_\infty e^{t\hat{A}}, \end{aligned} \tag{6.26}$$

We focus on the action of the matrix exponential on X_∞.

Lemma 6.18 (Action of the Matrix Exponential on X_∞):
Assume that (A, B) is stabilizable and (A, C) is detectable. Moreover, let $X_\infty \in \mathbb{R}^{n \times n}$ be the stabilizing solution of the ARE (5.1). Then for all $k = 1, \ldots, n-1$ and all $t \geq 0$, it holds

$$\left\| e^{t\hat{A}^\mathsf{T}} X_\infty - Q_{\infty,k} Q_{\infty,k}^\mathsf{T} e^{t\hat{A}^\mathsf{T}} X_\infty \right\|_2 \leq \sqrt{\lambda_{k+1}^{\downarrow}(X_\infty) \lambda_1^{\downarrow}(X_\infty)}. \tag{6.27}$$

◊

Proof. If X_∞ is the stabilizing solution of the ARE (5.1), then X_∞ satisfies the ALE

$$\hat{A}^\mathsf{T} X_\infty + X_\infty \hat{A} = -X_\infty BB^\mathsf{T} X_\infty - C^\mathsf{T} C.$$

We apply Theorem 3.7 and get

$$X_\infty = \int_0^\infty e^{s\hat{A}^\mathsf{T}} \big(X_\infty BB^\mathsf{T} X_\infty + C^\mathsf{T} C \big) e^{s\hat{A}} \, \mathrm{d}s.$$

We multiply the latter equation with the matrix exponential and get

$$\begin{aligned} 0 \preccurlyeq e^{t\hat{A}^\mathsf{T}} X_\infty e^{t\hat{A}} &= \int_0^\infty e^{(t+s)\hat{A}^\mathsf{T}} \big(X_\infty BB^\mathsf{T} X_\infty + C^\mathsf{T} C \big) e^{(t+s)\hat{A}} \, \mathrm{d}s \\ &\preccurlyeq \int_0^\infty e^{s\hat{A}^\mathsf{T}} \big(X_\infty BB^\mathsf{T} X_\infty + C^\mathsf{T} C \big) e^{s\hat{A}} \, \mathrm{d}s = X_\infty. \end{aligned}$$

The remaining arguments are similar as in the proof of Theorem 6.17. We get

$$0 \preccurlyeq e^{t\hat{A}^\mathsf{T}} X_\infty e^{t\hat{A}} \preccurlyeq X_\infty \preccurlyeq Q_{\infty,k} Q_{\infty,k}^\mathsf{T} X_\infty Q_{\infty,k} Q_{\infty,k}^\mathsf{T} + \lambda_{k+1}^{\downarrow}(X_\infty) I.$$

We multiply with the projection matrix $I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T}$ from the left and right and get

$$0 \preccurlyeq \big(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T} \big) e^{t\hat{A}^\mathsf{T}} X_\infty e^{t\hat{A}} \big(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T} \big) \preccurlyeq \lambda_{k+1}^{\downarrow}(X_\infty) \big(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T} \big).$$

It follows that

$$\begin{aligned} \left\| \big(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T} \big) e^{t\hat{A}^\mathsf{T}} X_\infty \right\|_2^2 &\leq \left\| \big(I - Q_{\infty,k} Q_{\infty,k}^\mathsf{T} \big) e^{t\hat{A}^\mathsf{T}} X_\infty^{1/2} \right\|_2^2 \left\| X_\infty^{1/2} \right\|_2^2 \\ &\leq \lambda_{k+1}^{\downarrow}(X_\infty) \lambda_1^{\downarrow}(X_\infty). \end{aligned}$$

□

For sufficiently quick eigenvalue decay, we expect Bound (6.27) of Lemma 6.18 to be small. This motivates to approximate the action of the matrix exponential by

$$e^{t\hat{A}^\mathsf{T}} X_\infty \approx Q_{\infty,k} Q_{\infty,k}^\mathsf{T} e^{t\hat{A}^\mathsf{T}} X_\infty.$$

Moreover, we approximate the stabilizing solution X_∞ of the ARE by its best approximation of rank at most k

$$X_\infty \approx Q_{\infty,k} Q_{\infty,k}^\mathsf{T} X_\infty Q_{\infty,k} Q_{\infty,k}^\mathsf{T} = Q_{\infty,k} D_{\infty,k} Q_{\infty,k}^\mathsf{T},$$

where $D_{\infty,k} := \operatorname{diag}(\lambda_1^{\downarrow}(X_\infty), \ldots, \lambda_k^{\downarrow}(X_\infty)) \in \mathbb{R}^{k\times k}$. We combine this with Equation (6.26) and get an approximation of the form

$$X(t) \approx Q_{\infty,k} D_{\infty,k} Q_{\infty,k}^\mathsf{T} - Q_{\infty,k} \tilde{X}(t) Q_{\infty,k}^\mathsf{T}. \tag{6.28}$$

6.5.3 Galerkin System

Because of the results in Sections 6.5.1 and 6.5.2, we propose to set up a trial space for the Galerkin approach using a system of orthonormal eigenvectors corresponding to the largest eigenvalues. This can be obtained by using a low-rank method to obtain a numerical approximation of the solution of the ARE. A compact SVD of the numerical low-rank approximation of X_∞ can be used to approximate the eigenvectors corresponding to the largest eigenvalues. By virtue of Bounds (6.22) and (6.23), we remove the small singular values from the compact SVD. This also reduces the dimension of the trial space. Let $Z = QSV^\mathsf{T}$ be the truncated compact SVD of the low-rank approximation. With that, the trial space for the Galerkin approach is given by

$$\left\{ QYQ^\mathsf{T} \mid Y \in \mathbb{R}^{k\times k} \right\}.$$

As $X(t)$ converges to X_∞ and $X_\infty \approx ZZ^\mathsf{T}$, we propose the Galerkin approach

$$X(t) \approx ZZ^\mathsf{T} - Q\tilde{X}(t)Q^\mathsf{T} =: \hat{X}(t), \tag{6.29}$$

for the numerical approximation similar as in Equation (6.28). We insert the Approach (6.29) into the DRE (6.8a) and consider the defect

$$D(t) := \dot{\hat{X}}(t) - \left(A^\mathsf{T}\hat{X}(t) + \hat{X}(t)A - \hat{X}(t)BB^\mathsf{T}\hat{X}(t) + C^\mathsf{T}C \right).$$

We require the defect $D(t)$ to be orthogonal to the trial space, e.g.,

$$\left\langle D(t), QYQ^\mathsf{T} \right\rangle_\mathrm{F} = 0 \text{ for all } Y \in \mathbb{R}^{k\times k}.$$

This gives

$$\begin{aligned} 0 &= Q^\mathsf{T} D(t) Q \\ &= Q^\mathsf{T}\left(\dot{\hat{X}}(t) - \left(A^\mathsf{T}\hat{X}(t) + \hat{X}(t)A - \hat{X}(t)BB^\mathsf{T}\hat{X}(t) + C^\mathsf{T}C \right)\right) Q. \end{aligned}$$

Combining the latter equation with Equation (6.29) yields a SDRE for the unknown $\tilde{X}(t) \in \mathbb{R}^{k \times k}$

$$\begin{aligned}\dot{\tilde{X}}(t) &= Q^{\mathsf{T}}\big(A - BB^{\mathsf{T}}ZZ^{\mathsf{T}}\big)^{\mathsf{T}} Q\tilde{X}(t) + \tilde{X}(t)Q^{\mathsf{T}}\big(A - BB^{\mathsf{T}}ZZ^{\mathsf{T}}\big)Q \\ &\quad + \tilde{X}(t)Q^{\mathsf{T}}BB^{\mathsf{T}}Q\tilde{X}(t) - Q^{\mathsf{T}}RQ,\end{aligned}$$

where R is the residual of the low-rank approximation ZZ^{T}, i.e.,

$$R = A^{\mathsf{T}}ZZ^{\mathsf{T}} + ZZ^{\mathsf{T}}A - ZZ^{\mathsf{T}}BB^{\mathsf{T}}ZZ^{\mathsf{T}} + C^{\mathsf{T}}C.$$

We assume that the numerical low-rank approximation is accurate enough such that

$$\left\|Q^{\mathsf{T}}RQ\right\|_2 \leq \|R\|_2 \ll 1.$$

This means that the projected residual $Q^{\mathsf{T}}RQ$ is even smaller than the residual of the low-rank approximation R, therefore, we neglect the residual. Using the initial condition $X(0) = 0$ and the truncated compact SVD ($Z = QSV^{\mathsf{T}}$) of the low-rank factor Z, we obtain

$$0 = ZZ^{\mathsf{T}} - Q\tilde{X}(0)Q^{\mathsf{T}}.$$

This yields $\tilde{X}(0) = S^2$ and the Galerkin system for $\tilde{X}$ is given by

$$\begin{aligned}\dot{\tilde{X}}(t) &= Q^{\mathsf{T}}\big(A - BB^{\mathsf{T}}ZZ^{\mathsf{T}}\big)^{\mathsf{T}} Q\tilde{X}(t) + \tilde{X}(t)Q^{\mathsf{T}}\big(A - BB^{\mathsf{T}}ZZ^{\mathsf{T}}\big)Q \\ &\quad + \tilde{X}(t)Q^{\mathsf{T}}BB^{\mathsf{T}}Q\tilde{X}(t), \\ \tilde{X}(0) &= S^2.\end{aligned}$$

6.5.4 Algorithm

With minor adjustments, all arguments also hold for the generalized DRE

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C, \tag{6.30a}$$

$$X(0) = 0, \tag{6.30b}$$

with $M \in \mathbb{R}^{n \times n}$ nonsingular that can accommodate, e.g., a mass matrix from a finite element discretization. In summary, the proposed Galerkin approach can be implemented based on Algorithm 6.4.

6.5.5 Numerical Experiments

Setup

To quantify the performance of the `ARE-Galerkin-DRE` method (Algorithm 6.4), we consider several (generalized) DREs arising from the benchmark problems `RAIL`, `CONV_DIFF`, `FLOW`, and `COOKIE`; cf. Table A.1.

Algorithm 6.4: Galerkin Approach for the DRE (6.30) (`ARE-Galerkin-DRE`).

```
Input: $M \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times b}$, $C \in \mathbb{R}^{c \times n}$, truncation tolerance $tol_{\texttt{trunc}} > 0$.
Assumptions: $(AM^{-1}, B)$ is stabilizable and $(AM^{-1}, CM^{-1})$ is detectable.
Output: $X(t) \approx ZZ^{\mathsf{T}} - Q\tilde{X}(t)Q^{\mathsf{T}}$ that approximates the solution to the
        generalized DRE.

% solve the ARE:
$A^{\mathsf{T}}X_\infty M + M^{\mathsf{T}}X_\infty A - M^{\mathsf{T}}X_\infty BB^{\mathsf{T}}X_\infty M + C^{\mathsf{T}}C = 0$ for $X_\infty \approx ZZ^{\mathsf{T}}$;
% compute compact SVD and truncate:
use Algorithm B.1 with $(Z, tol_{\texttt{trunc}})$ as input and obtain the truncated
  compact SVD $(Z, Q, S)$ as output;
% compute matrices:
$\tilde{A} := Q^{\mathsf{T}}\left(AM^{-1} - BB^{\mathsf{T}}ZZ^{\mathsf{T}}\right)Q$;
$\tilde{B} := Q^{\mathsf{T}}B$;
% solve the SDRE using the modified Davison–Maki method (Algorithm 6.3):
$\dot{\tilde{X}}(t) = \tilde{A}^{\mathsf{T}}\tilde{X}(t) + \tilde{X}(t)\tilde{A} + \tilde{X}(t)\tilde{B}\tilde{B}^{\mathsf{T}}\tilde{X}(t),\ \tilde{X}(0) = S^2$;
```

We compare the Galerkin method (Algorithm 6.4) with the splitting methods (Section 6.6). The action of the matrix exponential was approximated utilizing the Gauss–Legendre Runge–Kutta method. We employed the Lie and Strang splittings of order 1 and 2, respectively, and the symmetric splittings of order $2, 4, 6$, and 8. We abbreviate the methods by `LIE`, `STRANG`, `SYMMETRIC2`, `SYMMETRIC4`, `SYMMETRIC6`, and `SYMMETRIC8`.

To evaluate the error, we computed a reference solution $X_{\texttt{ref}}$ using `SYMMETRIC8` with a constant time step size h. Table 6.1 gives basic information about the setup of the benchmark problems.

Instance	n	Interval	Reference Solution
`RAIL`	5177	$[0, 4512]$	`SYMMETRIC8`, $h = 2^{-5}$
`CONV_DIFF`	6400	$[0, 0.125]$	`SYMMETRIC8`, $h = 2^{-20}$
`FLOW`	9669	$[0, 0.25]$	`SYMMETRIC8`, $h = 2^{-21}$
`COOKIE`	7488	$[0, 4]$	`SYMMETRIC8`, $h = 2^{-16}$

Tab. 6.1: Information about the Benchmark Problems.

The reference solution $X_{\texttt{ref}}$ was computed on the Gold Node (Appendix A). All other computations were carried out on the Silver Node (Appendix A).

We report the absolute and relative errors

$$\|X(t) - X_{\texttt{ref}}(t)\| \quad \text{and} \quad \frac{\|X(t) - X_{\texttt{ref}}(t)\|}{\|X_{\texttt{ref}}(t)\|},$$

in the 2-norm and Frobenius norm, where $X(t)$ is the numerical approximation, and $X_{\texttt{ref}}$ is the reference solution. We give the norm of the reference solution $X_{\texttt{ref}}$ and the convergence to the stationary point $\|X_{\texttt{ref}}(t) - X_{\infty}\|_2$.

Furthermore, we evaluate the best approximation in the trial space of the reference solution, which is given by

$$X_{\texttt{best}}(t) := QQ^{\mathsf{T}} X_{\texttt{ref}}(t) QQ^{\mathsf{T}} = \operatorname*{arg\,min}_{X \in \left\{ QYQ^{\mathsf{T}} | Y \in \mathbb{R}^{k \times k} \right\}} \|X - X_{\texttt{ref}}(t)\|_{\mathrm{F}},$$

where Q is the matrix of Algorithm 6.4 line 2.

Numerical results for the Galerkin approximation of Algorithm 6.4 and for the splitting scheme solvers can be found in Appendices D.2.2 and D.2.3. The computational times for both methods are given in Appendix D.2.1.

Galerkin Approach and Splitting Schemes

The initial step of Algorithm 6.4 requires the solution of the associated (generalized) ARE. For this task, we use the RADI algorithm that iteratively computes the numerical solution to the following absolute and relative residuals

$$\left\|A^{\mathsf{T}} ZZ^{\mathsf{T}} M + M^{\mathsf{T}} ZZ^{\mathsf{T}} A - M^{\mathsf{T}} ZZ^{\mathsf{T}} BB^{\mathsf{T}} ZZ^{\mathsf{T}} M + C^{\mathsf{T}} C\right\|_2$$

and

$$\frac{\left\|A^{\mathsf{T}} ZZ^{\mathsf{T}} M + M^{\mathsf{T}} ZZ^{\mathsf{T}} A - M^{\mathsf{T}} ZZ^{\mathsf{T}} BB^{\mathsf{T}} ZZ^{\mathsf{T}} M + C^{\mathsf{T}} C\right\|_2}{\|C^{\mathsf{T}} C\|_2}.$$

The achieved values for the different test setups and the number of columns of the corresponding low-rank factor Z after truncation (Algorithm 6.4 line 2), that define the dimension of the reduced model, are listed in Table 6.4.

The 1-norm bound for the matrix exponential $tol_{\texttt{exp}}$ of the modified Davison–Maki method (Algorithm 6.3) was set to 10^{10}. The resulting step sizes are given in Table 6.3. We used two values for the truncation threshold $tol_{\texttt{trunc}}$, namely the machine precision $\varepsilon_{\texttt{mach}}$ and the rougher value $\sqrt{\varepsilon_{\texttt{mach}}}$.

We plot the numerical errors for the `ARE-Galerkin-DRE` method in Figures D.17, D.20, D.23, and D.26. The Figures D.18, D.19, D.21, D.22, D.24, D.25, D.27, and D.28 show the norm of the reference solution and the convergence to the stationary point. In view of the performance, we can interpret the presented numbers and plots as follows: As discussed in Section 6.4, the accuracy of the modified Davison–Maki method is nearly independent of the step size; cf. Figures D.17b/D.17d, D.20b/D.20d, D.23b/D.23d, and D.26b/D.26d.

The computational times for `ARE-Galerkin-DRE` include the numerical solution of the corresponding ARE and the subsequent integration of the projected dense DRE. Since the efforts for the time integration exactly doubles with a bisection of the step size, from the timings for the `RAIL` problem, with, e.g., 79s ($h = 2^{-3}$) and 148s ($h = 2^{-4}$)

(see Figure D.13a, $tol_{\texttt{trunc}} = \varepsilon_{\texttt{mach}}$), one infers that most of the time is spent solving the dense DRE.

The reference solution for the `RAIL` and `FLOW` problem is large in norm which makes the absolute error comparatively large; cf. Figures D.18 and D.24 in Appendix D.2.2.

The `LIE`, `STRANG`, and `SYMMETRIC2` splitting schemes gave an absolute and relative error nearly at the same level. Therefore, the Figures D.29–D.35 only show the error of the `SYMMETRIC2` splitting scheme.

In all examples, in terms of accuracy, the `ARE-Galerkin-DRE` ($tol_{\texttt{trunc}} = \varepsilon_{\texttt{mach}}$) approximation is nearly at the same level as the high-order splitting schemes; cf. Figures D.17b/D.29f, D.20b/D.31f, D.23b/D.33f,and D.26b/D.35f. However, we note that the `ARE-Galerkin-DRE` method does not give the best possible approximation in the trial space; compare the error levels for $X_{\texttt{best}}$.

In any case, the `ARE-Galerkin-DRE` method outperforms the splitting methods in computational time versus accuracy in all test examples. The performance can be further improved by adapting the truncation threshold $tol_{\texttt{trunc}}$; cf. line 2 of Algorithm 6.4. Apart from the savings in the timings, the reduced memory requirements can be significant; cf. Figures D.13a, D.14a, D.15a, and D.16a. For the `RAIL` example, the rougher tolerance $\sqrt{tol_{\texttt{trunc}}}$ instead of machine precision reducing the storage by a factor of ${}^{279^2}/{}_{193^2} \approx 2$; cf. Table 6.4. Indeed, these savings come at the expense of accuracy. For the `RAIL` example, this means a relative error level of about 10^{-9} versus 10^{-11} if truncation has happened to machine precision; cf. Figure D.17b. For the other examples, the approximation accuracy was only slightly affected by the larger truncation tolerance. The most favorable example is the `FLOW` example, where the relaxed truncation threshold led to savings of a factor ${}^{407\text{s}}/{}_{107\text{s}} \approx 4$ ($h = 2^{-20}$) in the timings (Figure D.15a) and a factor of ${}^{252^2}/{}_{115^2} \approx 5$ in memory requirements (Table 6.4) while, except for a short initial phase, maintaining the same approximation accuracy (Figure D.23d).

Instance	Time to solve ARE (s)
`RAIL`	0.85
`CONV_DIFF`	1.59
`FLOW`	2.06
`COOKIE`	2.51

Tab. 6.2: Time to solve the ARE.

Instance	Step sizes h
`RAIL`	$\{2^{0}, 2^{-1}, \ldots, 2^{-5}\}$
`CONV_DIFF`	$\{2^{-12}, 2^{-13}, \ldots, 2^{-16}\}$
`FLOW`	$\{2^{-15}, 2^{-16}, \ldots, 2^{-20}\}$
`COOKIE`	$\{2^{-15}, 2^{-16}, \ldots, 2^{-20}\}$

Tab. 6.3: Step sizes h for the modified Davison–Maki Method.

Instance	n	$tol_{\mathtt{trunc}}$	Galerkin System	Abs. Residual	Rel. Residual
RAIL	5177	$\sqrt{\varepsilon_{\mathtt{mach}}}$	193	$2.91 \cdot 10^{-14}$	$2.43 \cdot 10^{-15}$
		$\varepsilon_{\mathtt{mach}}$	279	$3.25 \cdot 10^{-14}$	$2.71 \cdot 10^{-15}$
CONV_DIFF	6400	$\sqrt{\varepsilon_{\mathtt{mach}}}$	36	$1.37 \cdot 10^{-10}$	$3.06 \cdot 10^{-14}$
		$\varepsilon_{\mathtt{mach}}$	54	$1.39 \cdot 10^{-10}$	$3.11 \cdot 10^{-14}$
FLOW	9669	$\sqrt{\varepsilon_{\mathtt{mach}}}$	115	$2.85 \cdot 10^{-8}$	$2.85 \cdot 10^{-8}$
		$\varepsilon_{\mathtt{mach}}$	252	$1.06 \cdot 10^{-11}$	$1.06 \cdot 10^{-11}$
COOKIE	7488	$\sqrt{\varepsilon_{\mathtt{mach}}}$	122	$1.29 \cdot 10^{-14}$	$3.07 \cdot 10^{-12}$
		$\varepsilon_{\mathtt{mach}}$	169	$1.33 \cdot 10^{-14}$	$3.16 \cdot 10^{-12}$

Tab. 6.4: Residuals for the generalized ARE $0 = A^{\mathsf{T}}XM + M^{\mathsf{T}}XA - M^{\mathsf{T}}XBB^{\mathsf{T}}XM + C^{\mathsf{T}}C$.

Large-scale Examples

We consider the benchmark problems RAIL, CONV_DIFF, and COOKIE for finer space discretization resulting in a larger state-space dimension n. Moreover, we consider the CHIP model; cf. Table A.1. As the computation of reference solutions for large-order systems by the high-order splitting methods easily exceeds computational resources, we only report residuals and the computational timings. Table 6.5 reports the state-space dimension n of the models, the size of the resulting Galerkin system, and the absolute and relative residual of the numerical approximation of the ARE. Detailed information about computational timings of the ARE-Galerkin-DRE method Algorithm 6.4 are given in Table 6.6. We report the time for the numerical approximation of the ARE ($t_{\mathtt{ARE}}$), the computation of the singular value decomposition ($t_{\mathtt{svd}}$), the assembly of the system matrices of the Galerkin system ($t_{\mathtt{gal}}$), the approximation of the matrix exponential and the norm computation ($t_{\mathtt{expm}}$); cf. Algorithm 6.4 line 1, line 2, lines 3–4, and Algorithm 6.3 lines 2–3. The computational time for the time-stepping of the modified Davison–Maki method (Algorithm 6.3 lines 6–10) is excluded. All timings are given in seconds.

As the timings suggest, for similar setups, increasing system sizes almost exclusively affect the time needed to solve the ARE, and to some extent, to compute the compact SVD for truncating the basis. As the truncation extracts the relevant directions, the resulting sizes of the projected systems only show a moderate increase. Accordingly, the efforts for solving the projected equations increase slightly.

Instance	n	$tol_{\mathtt{trunc}}$	Galerkin System	Abs. Residual	Rel. Residual
RAIL_20K	20 209	$\sqrt{\varepsilon_{\mathtt{mach}}}$	224	$6.75 \cdot 10^{-14}$	$5.63 \cdot 10^{-15}$
		$\varepsilon_{\mathtt{mach}}$	323	$6.37 \cdot 10^{-14}$	$5.31 \cdot 10^{-15}$
RAIL_79K	79 841	$\sqrt{\varepsilon_{\mathtt{mach}}}$	254	$6.13 \cdot 10^{-14}$	$5.11 \cdot 10^{-15}$
		$\varepsilon_{\mathtt{mach}}$	353	$5.90 \cdot 10^{-14}$	$4.92 \cdot 10^{-15}$
CONV_DIFF_160K	160 000	$\sqrt{\varepsilon_{\mathtt{mach}}}$	48	$2.16 \cdot 10^{-9}$	$1.93 \cdot 10^{-14}$
		$\varepsilon_{\mathtt{mach}}$	79	$2.16 \cdot 10^{-9}$	$1.93 \cdot 10^{-14}$
CONV_DIFF_1M	1 000 000	$\sqrt{\varepsilon_{\mathtt{mach}}}$	52	$1.94 \cdot 10^{-8}$	$2.77 \cdot 10^{-14}$
		$\varepsilon_{\mathtt{mach}}$	82	$1.93 \cdot 10^{-8}$	$2.76 \cdot 10^{-14}$
CONV_DIFF_9M	9 000 000	$\sqrt{\varepsilon_{\mathtt{mach}}}$	57	$3.72 \cdot 10^{-7}$	$5.91 \cdot 10^{-14}$
		$\varepsilon_{\mathtt{mach}}$	101	$3.02 \cdot 10^{-7}$	$4.80 \cdot 10^{-14}$
COOKIE_425K	425 272	$\sqrt{\varepsilon_{\mathtt{mach}}}$	156	$5.09 \cdot 10^{-14}$	$6.43 \cdot 10^{-10}$
		$\varepsilon_{\mathtt{mach}}$	238	$5.06 \cdot 10^{-14}$	$6.40 \cdot 10^{-10}$
COOKIE_1185K	1 185 586	$\sqrt{\varepsilon_{\mathtt{mach}}}$	163	$1.58 \cdot 10^{-13}$	$5.59 \cdot 10^{-9}$
		$\varepsilon_{\mathtt{mach}}$	309	$1.60 \cdot 10^{-13}$	$5.65 \cdot 10^{-9}$
COOKIE_2656K	2 656 643	$\sqrt{\varepsilon_{\mathtt{mach}}}$	168	$4.51 \cdot 10^{-13}$	$3.58 \cdot 10^{-8}$
		$\varepsilon_{\mathtt{mach}}$	306	$1.05 \cdot 10^{-13}$	$8.35 \cdot 10^{-9}$
CHIP	20 082	$\sqrt{\varepsilon_{\mathtt{mach}}}$	103	$6.81 \cdot 10^{-10}$	$6.81 \cdot 10^{-10}$
		$\varepsilon_{\mathtt{mach}}$	198	$1.79 \cdot 10^{-12}$	$1.79 \cdot 10^{-12}$

Tab. 6.5: Residuals for the generalized ARE $0 = A^{\mathsf{T}}XM + M^{\mathsf{T}}XA - M^{\mathsf{T}}XBB^{\mathsf{T}}XM + C^{\mathsf{T}}C$ and the Size of the Galerkin System.

6.5.6 Nonzero Initial Conditions

In this section, we extend our discussion to the case of spsd nonzero initial conditions $X(0) = Z_0Z_0^{\mathsf{T}}$. Here, the Inequality (6.17) ($0 \preccurlyeq X(t) \preccurlyeq X_\infty$) needs to be established by other means than Theorem 6.5, which requires $\dot{X}(0) \succcurlyeq 0$.

We distinguish the cases of $\dot{X}(0)$ symmetric positive semidefinite, negative semidefinite, and indefinite.

If $\dot{X}(0) \succcurlyeq 0$, then, as for $X_0 = 0$, the solution is monotonically non-decreasing and Inequality (6.17) holds; cf. Theorem 6.5.

Interestingly, the case of $\dot{X}(0) \preccurlyeq 0$ fits the framework without relying on the solution of the ARE. In fact, in this case, the solution $X(t)$ is monotonically non-increasing; cf. Theorem 6.5. Still, the solution $X(t)$ is spsd for all $t \geq 0$, so that the inequality

$$0 \preccurlyeq X(t) \preccurlyeq Z_0Z_0^{\mathsf{T}}$$

holds for all $t \geq 0$; cf. Theorem 6.8. In particular, it follows that $\operatorname{im}(X(t)) \subseteq \operatorname{im}(Z_0)$ for all $t \geq 0$ (Lemma 2.6), and a trial space is readily defined by a basis of $\operatorname{im}(Z_0)$. Thus, a

Instance	n	$tol_{\mathtt{trunc}}$	$t_{\mathtt{ARE}}$	$t_{\mathtt{svd}}$	$t_{\mathtt{gal}}$	$t_{\mathtt{expm}}$	$t_{\mathtt{total}}$
`RAIL_20K`	20 209	$\sqrt{\varepsilon_{\mathtt{mach}}}$	3.23	0.41	0.17	$4.66 \cdot 10^{-2}$	**3.86**
		$\varepsilon_{\mathtt{mach}}$	3.48	0.43	0.25	$6.87 \cdot 10^{-2}$	**4.23**
`RAIL_79K`	79 841	$\sqrt{\varepsilon_{\mathtt{mach}}}$	11.13	1.76	0.67	$6.80 \cdot 10^{-2}$	**13.63**
		$\varepsilon_{\mathtt{mach}}$	11.96	1.92	0.94	$6.95 \cdot 10^{-2}$	**14.89**
`CONV_DIFF_160K`	160 000	$\sqrt{\varepsilon_{\mathtt{mach}}}$	22.56	0.75	0.08	$2.98 \cdot 10^{-2}$	**23.42**
		$\varepsilon_{\mathtt{mach}}$	22.35	0.80	0.14	$5.71 \cdot 10^{-3}$	**23.29**
`CONV_DIFF_1M`	1 000 000	$\sqrt{\varepsilon_{\mathtt{mach}}}$	159.53	5.10	0.48	$4.22 \cdot 10^{-3}$	**165.12**
		$\varepsilon_{\mathtt{mach}}$	160.22	5.82	0.85	$6.16 \cdot 10^{-3}$	**166.90**
`CONV_DIFF_9M`	9 000 000	$\sqrt{\varepsilon_{\mathtt{mach}}}$	2754.57	45.40	5.61	$8.75 \cdot 10^{-2}$	**2805.66**
		$\varepsilon_{\mathtt{mach}}$	2739.42	46.46	9.22	$4.31 \cdot 10^{-2}$	**2795.14**
`COOKIE_425K`	425 272	$\sqrt{\varepsilon_{\mathtt{mach}}}$	193.43	6.40	10.52	$1.58 \cdot 10^{-2}$	**210.36**
		$\varepsilon_{\mathtt{mach}}$	192.66	6.85	13.33	$2.03 \cdot 10^{-2}$	**212.86**
`COOKIE_1185K`	1 185 586	$\sqrt{\varepsilon_{\mathtt{mach}}}$	874.55	26.66	42.15	$1.45 \cdot 10^{-2}$	**943.38**
		$\varepsilon_{\mathtt{mach}}$	869.38	27.30	59.20	$2.47 \cdot 10^{-2}$	**955.90**
`COOKIE_2656K`	2 656 643	$\sqrt{\varepsilon_{\mathtt{mach}}}$	2364.17	56.35	109.58	$2.31 \cdot 10^{-2}$	**2530.12**
		$\varepsilon_{\mathtt{mach}}$	2376.80	57.15	155.41	$3.28 \cdot 10^{-2}$	**2589.40**
`CHIP`	20 082	$\sqrt{\varepsilon_{\mathtt{mach}}}$	4.97	0.25	0.07	$1.51 \cdot 10^{-2}$	**5.30**
		$\varepsilon_{\mathtt{mach}}$	5.33	0.21	0.10	$1.85 \cdot 10^{-2}$	**5.65**

Tab. 6.6: Timings for the Large-scale Examples.

basis can be computed by a QR factorization or a compact SVD of Z_0.

The symmetric but indefinite case that we write as

$$\dot{X}(0) = Z_+Z_+^{\mathsf{T}} - Z_-Z_-^{\mathsf{T}}$$

requires additional reasoning. One may compute a suitable upper bound $\tilde{X}_\infty$ that replaces X_∞ in Inequality (6.17) as follows. Consider the modified DRE

$$\begin{aligned}\dot{\tilde{X}}(t) &= A^{\mathsf{T}}\tilde{X}(t) + \tilde{X}(t)A - \tilde{X}(t)BB^{\mathsf{T}}\tilde{X}(t) + C^{\mathsf{T}}C + Z_-Z_-^{\mathsf{T}},\\ \tilde{X}(0) &= Z_0Z_0^{\mathsf{T}}.\end{aligned}$$

By construction, it holds $\dot{\tilde{X}}(0) = Z_+Z_+^{\mathsf{T}} \succcurlyeq 0$ so that $\tilde{X}(t)$ is monotonically non-decreasing for all $t \geq 0$; cf. Theorem 6.5. Moreover, with (A, B) stabilizable and (A, C) detectable, the solution $\tilde{X}(t)$ converges to the unique positive semidefinite solution $\tilde{X}_\infty$ of the ARE

$$0 = A^{\mathsf{T}}\tilde{X}_\infty + \tilde{X}_\infty A - \tilde{X}_\infty BB^{\mathsf{T}}\tilde{X}_\infty + C^{\mathsf{T}}C + Z_-Z_-^{\mathsf{T}};$$

cf. Theorem 6.14. This gives $0 \preccurlyeq \tilde{X}(t) \preccurlyeq \tilde{X}_\infty$ for all $t \geq 0$. The comparison theorem (Theorem 6.7) yields $X(t) \preccurlyeq \tilde{X}(t)$ for all $t \geq 0$. With that, the inequality

$$0 \preccurlyeq X(t) \preccurlyeq \tilde{X}(t) \preccurlyeq \tilde{X}_\infty$$

holds for all $t \geq 0$, and the bounds on the projection error Bound (6.22) and Theorem 6.17 can be established analogously.

6.6 Splitting Methods

This section is based on [112, 113]. We consider splitting methods for the DRE (6.8)

$$\begin{aligned} \dot{X}(t) &= A^\mathsf{T} X(t) + X(t)A - X(t)BB^\mathsf{T} X(t) + C^\mathsf{T} C, \\ X(0) &= Z_0 Z_0^\mathsf{T}. \end{aligned}$$

We focus on the large-scale case. This means that n is large and the matrices B, C^T, and Z_0 have only a few columns. The key idea of the splitting methods is to divide the problem into a linear subproblem

$$\begin{aligned} \dot{X}(t) &= A^\mathsf{T} X(t) + X(t)A + C^\mathsf{T} C, && (6.31a) \\ X(0) &= Z_0 Z_0^\mathsf{T}, && (6.31b) \end{aligned}$$

and nonlinear subproblem

$$\begin{aligned} \dot{X}(t) &= -X(t)BB^\mathsf{T} X(t), && (6.32a) \\ X(0) &= Z_0 Z_0^\mathsf{T}. && (6.32b) \end{aligned}$$

6.6.1 Lie and Strang Splitting Methods

The major motivation for the splitting is that the linear problem (6.31) and the nonlinear problem (6.32) are easier to solve than the original problem. Next, we introduce the solution operators or the flow maps for both equations. For this, we fix a positive step size $h > 0$. The solution operator for the linear problem (6.31) is

$$\mathcal{F}_h \colon \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n},\ X \mapsto e^{hA^\mathsf{T}} X e^{hA} + \int\limits_0^h e^{sA^\mathsf{T}} C^\mathsf{T} C e^{sA} \,\mathrm{d}s. \tag{6.33}$$

For the nonlinear subproblem (6.32), we restrict the domain of definition to the spsd matrices,

$$\begin{aligned} &\mathcal{G}_h \colon \left\{X \in \mathbb{R}^{n\times n} \mid X \succcurlyeq 0\right\} \to \left\{X \in \mathbb{R}^{n\times n} \mid X \succcurlyeq 0\right\}, \\ &X \mapsto \left(I + hXBB^\mathsf{T}\right)^{-1} X. \end{aligned} \tag{6.34}$$

As the matrices X and BB^T are real spsd, it follows that all eigenvalues of the product XBB^T are nonnegative; cf. [59, Cor. 7.6.2 (b)]. Therefore, the matrix $I + hXBB^\mathsf{T}$ is nonsingular. Moreover, the identity

$$\big(I + hXBB^\mathsf{T}\big)X = X\big(I + hBB^\mathsf{T}X\big)$$

yields that $\big(I + hXBB^\mathsf{T}\big)^{-1}X$ is symmetric. The matrix X is spsd. Therefore, we write $X = ZZ^\mathsf{T}$ for some matrix Z and verify that

$$\begin{aligned}\mathcal{G}_h(X) &= \big(I + hXBB^\mathsf{T}\big)^{-1}X = \big(I + hZZ^\mathsf{T}BB^\mathsf{T}\big)^{-1}ZZ^\mathsf{T} \\ &= Z\big(I + hZ^\mathsf{T}BB^\mathsf{T}Z\big)^{-1}Z^\mathsf{T} \succcurlyeq 0.\end{aligned} \tag{6.35}$$

Lemma 6.19 ([112]):
Assume that $h > 0$. The operators $\mathcal{F}_h$ and $\mathcal{G}_h$ map spsd matrices to spsd matrices. ◊

To generate an approximation of the solution of the DRE (6.8), we use the solution operators to alternate between the linear and nonlinear subproblem. One time step of the *Lie* and *Strang* splitting method are given by

$$\mathcal{G}_h\mathcal{F}_h \text{ and } \mathcal{G}_{h/2}\mathcal{F}_h\mathcal{G}_{h/2},$$

respectively. We apply the time-stepping k times and obtain the approximations

$$\begin{aligned}X(kh) &\approx (\mathcal{G}_h\mathcal{F}_h)^k Z_0Z_0^\mathsf{T}, \\ X(kh) &\approx \big(\mathcal{G}_{h/2}\mathcal{F}_h\mathcal{G}_{h/2}\big)^k Z_0Z_0^\mathsf{T},\end{aligned}$$

where X is the solution of the DRE (6.8). The *Lie* and the *Strang* splitting methods are of first and second-order accuracy, respectively. Both methods generate real spsd approximations; cf. Lemma 6.19.

6.6.2 Numerical Realization

We focus on the numerical realization of the splitting schemes. Here, the main problem is the application of the operators $\mathcal{F}_h$ and $\mathcal{G}_h$. Because of memory limitations, it is crucial to implement the method in a low-rank fashion.

Application of the Operator $\mathcal{G}_h$

Using Equation (6.34), the application of the operator $\mathcal{G}_h$ on a low-rank product ZZ^T can be implemented as in Algorithm 6.5.

Algorithm 6.5: Application of the Operator $\mathcal{G}_h$ [112, Sec. II B].

```
Input: matrix $B$ of the DRE (6.8), real low-rank factor $Z$, and step size $h > 0$
Output: matrix $\tilde{Z}$ such that $\tilde{Z}\tilde{Z}^\mathsf{T} = \mathcal{G}_h ZZ^\mathsf{T}$

% compute a Cholesky factorization:
$L := \texttt{chol}(I + hZ^\mathsf{T}BB^\mathsf{T}Z)$;

% solve the linear system:
$\tilde{Z} := ZL^{-1}$;
return $\tilde{Z}$;
```

Application of the Operator $\mathcal{F}_h$

The application of the operator $\mathcal{F}_h$ on a low-rank product ZZ^T is more involved. Here, the action of the matrix exponential on the low-rank factor Z

$$e^{hA^\mathsf{T}}Z$$

and the integral term

$$\int\limits_0^h e^{sA^\mathsf{T}}CC^\mathsf{T}e^{sA}\,\mathrm{d}s \tag{6.36}$$

have to be approximated.

Action of the Matrix Exponential

Plenty of methods are available to approximate the action of the matrix exponential on a vector or a matrix, e.g., Krylov subspace methods [50, 57, 102] or interpolation-based methods [29, 87]. Here, we focus on fully implicit Runge–Kutta methods applied to the corresponding linear IVP

$$\begin{aligned}\dot{\tilde{Z}}(t) &= A^\mathsf{T}\tilde{Z}(t),\\ \tilde{Z}(0) &= Z.\end{aligned}$$

Let

$$\begin{array}{c|c}\mathbf{c} & \mathbf{A}\\ \hline & \mathbf{b}^\mathsf{T}\end{array}$$

be the Butcher tableau of an s-stage Runge–Kutta method with weights $\mathbf{b} \in \mathbb{R}^s$, nodes $\mathbf{c} \in \mathbb{R}^s$, and coefficients $\mathbf{A} \in \mathbb{R}^{s\times s}$. We assume that $\mathbf{A}$ is nonsingular, then one step of the s-stage Runge–Kutta method with step size $h > 0$ requires the solution of the linear system

$$\left(\mathbf{A}^{-1} \otimes I - hI_s \otimes A^\mathsf{T}\right) \begin{bmatrix} \hat{Z}_1 \\ \vdots \\ \hat{Z}_s \end{bmatrix} = h \begin{bmatrix} A^\mathsf{T} Z \\ \vdots \\ A^\mathsf{T} Z \end{bmatrix} \tag{6.37}$$

for the unknowns $\hat{Z}_1, \ldots, \hat{Z}_s$. The approximate $\tilde{Z}_1 \approx \tilde{Z}(h)$ can be computed as

$$\tilde{Z}_1 = \tilde{Z}(0) + \sum_{i=1}^{s} \mathbf{d}_i \hat{Z}_i = Z + \sum_{i=1}^{s} \mathbf{d}_i \hat{Z}_i,$$

where $\mathbf{d} = \mathbf{b}^\mathsf{T} \mathbf{A}^{-1}$; cf. [52, IV.8]. The direct application of sparse-direct methods to System (6.37) should be avoided. The matrix

$$\mathbf{A}^{-1} \otimes I - hI_s \otimes A^\mathsf{T} = \begin{bmatrix} \mathbf{A}_{1,1}^{-1} I - hA^\mathsf{T} & \mathbf{A}_{1,2}^{-1} I & \cdots & \mathbf{A}_{1,s}^{-1} I \\ \mathbf{A}_{2,1}^{-1} I & \ddots & & \mathbf{A}_{2,s}^{-1} I \\ \vdots & & \ddots & \vdots \\ \mathbf{A}_{s,1}^{-1} I & \mathbf{A}_{s,2}^{-1} I & \cdots & \mathbf{A}_{s,s}^{-1} I - hA^\mathsf{T} \end{bmatrix}$$

is of size $ns \times ns$, and each of s^2 blocks represents a sparse matrix of size $n \times n$. We diagonalize the coefficient matrix ($\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$) to decouple System (6.37); cf. [27]. We use Lemma 2.39. System (6.37) transforms to a block diagonal system

$$\left(\mathbf{D}^{-1} \otimes I - hI_s \otimes A^\mathsf{T}\right) \mathbf{V}^{-1} \otimes I \begin{bmatrix} \hat{Z}_1 \\ \vdots \\ \hat{Z}_s \end{bmatrix} = h\mathbf{V}^{-1} \otimes I \begin{bmatrix} A^\mathsf{T} Z \\ \vdots \\ A^\mathsf{T} Z \end{bmatrix}. \tag{6.38}$$

As $\mathbf{A}$ is real, the eigenvalues and eigenvectors of $\mathbf{A}$ come in complex conjugated pairs. Therefore, we can assume that the matrices $\mathbf{D}$ and $\mathbf{V}$ have the forms

$$\begin{aligned} \mathbf{D} &= \operatorname{diag}(\alpha_1, \overline{\alpha}_1, \ldots, \alpha_m, \overline{\alpha}_m, \alpha_{m+1}, \ldots, \alpha_r), \\ \mathbf{V} &= \left[\mathbf{v}_1, \overline{\mathbf{v}}_1, \ldots, \mathbf{v}_m, \overline{\mathbf{v}}_m, \mathbf{v}_{m+1}, \ldots, \mathbf{v}_r\right]. \end{aligned} \tag{6.39}$$

Hence, also, the rows of $\mathbf{V}^{-1}$ come in complex conjugated pairs, i.e.,

$$\mathbf{V}^{-1} = \left[\mathbf{w}_1, \overline{\mathbf{w}}_1, \ldots, \mathbf{w}_m, \overline{\mathbf{w}}_m, \mathbf{w}_{m+1}, \ldots, \mathbf{w}_r\right]^\mathsf{T}. \tag{6.40}$$

We combine Equations (6.38)–(6.40) and notice that for each complex conjugated pair of eigenvalues only one linear system has to be solved. Algorithm 6.6 summarizes our considerations.

Algorithm 6.6: Approximation of $e^{hA^\mathsf{T}}Z$ with an Implicit Runge–Kutta Method.

```
   Input: matrix $A$ of the DRE (6.8), real low-rank factor $Z$, step size $h > 0$,
   coefficients $\mathbf{A}$ and weights $\mathbf{b}$ of the implicit Runge–Kutta method, and vector
   $\mathbf{d} = \mathbf{b}^\mathsf{T}\mathbf{A}^{-1}$
   Assumptions: $\mathbf{A}$ is nonsingular and is diagonalizable; $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$.
   The matrices $\mathbf{V}, \mathbf{D}$, and $\mathbf{V}^{-1}$ are sorted as in Equations (6.39) and (6.40).
   Output: matrix $\tilde{Z}$ such that $\tilde{Z} \approx e^{hA^\mathsf{T}}Z$

   % prepare the right-hand side:
 1 $Z_{\texttt{rhs}} := hA^\mathsf{T}Z$;
   % apply the implicit Runge–Kutta method:
 2 for $i = 1, \ldots, s$ do
       % sum entries of $i$-th row of $\mathbf{V}^{-1}$:
 3     $\gamma := \sum_{j=1}^{s} \mathbf{V}^{-1}_{i,j}$;
       % solve the linear system:
 4     $Z_{\texttt{temp},i} := \left(\frac{1}{\alpha_i}I - hA^\mathsf{T}\right)^{-1}\gamma Z_{\texttt{rhs}}$;
 5     if $\Im(\alpha_i) \neq 0$ then
 6         $Z_{\texttt{temp},i+1} := \overline{Z}_{\texttt{temp},i}$;
 7         $i := i + 1$;
   % back transform the solution and advance in time:
 8 $\tilde{Z} := Z$;
 9 for $i = 1, \ldots, s$ do
10     $\tilde{Z} := \tilde{Z} + \mathbf{d}_i \sum_{j=1}^{s} \mathbf{V}_{i,j} Z_{\texttt{temp},j}$;
11 return $\tilde{Z}$;
```

The action of the matrix exponential to a low-rank matrix has to be approximated in each time step of the splitting method. Algorithm 6.6 line 4 requires the solution of a sparse linear system. If sparse-direct methods are employed, and if the step size h does not change, then the factorizations can be reused.

Integral Term

For the approximation of the integral term (6.36), we focus on quadrature rules. Alternatively, Krylov subspace methods also apply; cf. [75]. The quadrature rule approximation is given by

$$\int_0^h e^{sA^\mathsf{T}}C^\mathsf{T}Ce^{sA}\,\mathrm{d}s \approx \sum_{i=0}^{k} w_i e^{\tau_i A^\mathsf{T}}CC^\mathsf{T}e^{\tau_i A} = \tilde{Z}\tilde{Z}^\mathsf{T},$$

where $w_i > 0$ are the weights, τ_i are the nodes, and

$$\tilde{Z} = \left[\sqrt{w_0} e^{\tau_0 A^\mathsf{T}} C, \ldots, \sqrt{w_k} e^{\tau_k A^\mathsf{T}} C\right].$$

Here, the action of the matrix exponential $e^{\tau_k A^\mathsf{T}} C$ can be approximated using Algorithm 6.6. To reduce the memory requirements, the product $\tilde{Z}\tilde{Z}^\mathsf{T}$ should be truncated using a compact SVD.

Algorithm 6.7: Approximation of the Integral Term (6.36) [112, Alg. 2].

```
Input: matrices $A$ and $C$ of the DRE (6.8), positive weights $w_0, \ldots, w_k$ and
       nodes $\tau_0, \ldots, \tau_k$ of the quadrature rule
Output: matrix $\tilde{Z}$ such that $\int_0^h e^{sA^\mathsf{T}} C^\mathsf{T} C e^{sA} \, \mathrm{d}s \approx \tilde{Z}\tilde{Z}^\mathsf{T}$

% apply the quadrature rule:
$\tilde{Z} := [\,]$;
for $i = 0, \ldots, k$ do
    use Algorithm 6.6 to approximate $e^{\tau_i A^\mathsf{T}} C \approx Z_{\mathtt{temp}}$;
    $\tilde{Z} := \left[\tilde{Z}, \sqrt{w_i} Z_{\mathtt{temp}}\right]$;
truncate $\tilde{Z}$ using Algorithm B.1;
return $\tilde{Z}$;
```

6.6.3 Higher-order Splitting Methods

Higher-order splitting methods can be obtained by additive combinations of the solution operators $\mathcal{F}_h$ and $\mathcal{G}_h$. We consider the *symmetric* splitting method

$$\sum_{k=1}^{s} \gamma_k \left(\mathcal{F}_{h/k} \mathcal{G}_{h/k}\right)^k + \gamma_k \left(\mathcal{G}_{h/k} \mathcal{F}_{h/k}\right)^k,$$

where the coefficients $\gamma_1, \ldots, \gamma_s \in \mathbb{R}$ are appropriately chosen; cf. Table 6.7. The symmetric splitting method has order $2s$. The main disadvantage is that some coefficients may be negative. Here, the definiteness preserving low-rank formulation ZZ^T for the approximations is not applicable. Similar as for the BDF methods (Section 4.3), an LDL^T-type low-rank formulation is proposed. In this setting, the application of the operators $\mathcal{F}_h$ and $\mathcal{G}_h$ can be implemented like in Algorithms 6.5 and 6.6.

$2s$	γ_1	γ_2	γ_3	γ_4
2	$\frac{1}{2}$			
4	$-\frac{1}{6}$	$\frac{2}{3}$		
6	$\frac{1}{48}$	$-\frac{8}{15}$	$\frac{81}{80}$	
8	$-\frac{1}{720}$	$\frac{8}{45}$	$-\frac{729}{560}$	$\frac{512}{315}$

Tab. 6.7: Coefficients γ of the Symmetric Splitting Methods of Orders $2, 4, 6$, and 8.

Remark 6.20:

The Lie, Strang, and the symmetric splitting methods also apply to the generalized DRE (M nonsingular)

$$\begin{aligned} M^{\mathsf{T}}\dot{X}(t)M &= A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C, \\ X(0) &= Z_0Z_0^{\mathsf{T}}. \end{aligned} \qquad \Diamond$$

CHAPTER 7

CONCLUSIONS

Contents

7.1 Conclusions and Future Research Perspectives

This work investigated representations and properties of solutions of the autonomous DLE and DRE. A numerical approach was derived that uses the space spanned by a system of orthonormal eigenvectors corresponding to the largest eigenvalues of the solution approximation of the ALE and ARE. The proposed Galerkin approach comes with error bounds related to the projection of the solution onto the trial space. Several numerical results have shown superior performance in terms of memory requirements and computational times over `BDF/ADI` and splitting methods of high-order for problems with stable coefficient matrix and homogeneous initial conditions.

The key result that ensures feasibility and performance of the approach bases on a lower and an upper bound of the solution with respect to the Loewner ordering and a sufficiently quick eigenvalue decay of the solution of their algebraic counterparts; cf. Inequality (6.17). While for the DLE, the bounds can be derived from the solution representation Equation (4.11), for the DRE, the ordering can be derived by monotonicity and comparison arguments.

The derivation of lower and upper bounds is also key to generalizing the obtained results to other problem classes; as it is discussed in Section 6.5.6 for nonzero initial values. The treatment of the particular cases has been fully characterized in theory, however, the numerical computation of the decompositions needed for setting up the auxiliary problems have not been investigated. In this respect, the performance of the algorithms for nonzero initial conditions has not been tested.

A further generalization would concern problems with time-varying matrices. Here, the upper bound on the solution may, in general, not be a priori available. However, as much as the bounds are concerned, one may investigate periodic coefficients and resort to theoretical results and numerical approaches on periodic DREs; cf. [1, 121].

Another open problem relates to the error and defect estimates for the Galerkin approximation.

Appendices

APPENDIX A

HARDWARE, SOFTWARE AND BENCHMARK PROBLEMS

Contents

A.1 Hardware and Software Specifications

The hardware and software specifications of the computing server nodes used for the numerical experiments are listed below.

Silver Node

- operating system: `CentOS Linux release 7.5.1804`
- kernel release: `3.10.0-862.9.1.el7.x86_64`
- CPU type: Intel® Xeon® Silver 4110
- number of physical CPUs: 2
- number of cores (virtual): 16
- RAM: 192GB
- MATLAB R2019b
- environment variable: `MKL_ENABLE_INSTRUCTIONS=AVX2`

Gold Node

- operating system: `CentOS Linux release 7.5.1804`
- kernel release: `3.10.0-862.9.1.el7.x86_64`
- CPU type: Intel® Xeon® Gold 6130

- number of physical CPUs: 2
- number of cores (virtual): 32
- RAM: 192GB
- MATLAB R2019b

A.2 Benchmark Problems

Information about the benchmark problems used for the numerical experiments are listed below and the running example `TRIDIAG(`α`)`.

Instance	n	Matrices	Reference
`CHIP`	20 082	M diagonal positive definite, A symmetric and stable, $M^{-1}A$ stable, $B \in \mathbb{R}^{n \times 1},\ C \in \mathbb{R}^{5 \times n}$	[115]
`CONV_DIFF`	6400	$M = I_n$, A nonsymmetric and stable, $B \in \mathbb{R}^{n \times 1},\ C \in \mathbb{R}^{1 \times n}$	[93]
`COOKIE`	7488	M nonsymmetric, A nonsymmetric and stable, $M^{-1}A$ stable, $B \in \mathbb{R}^{n \times 1},\ C \in \mathbb{R}^{4 \times n}$	[97]
`FLOW`	9669	M diagonal positive definite, A symmetric and stable, $M^{-1}A$ stable, $B \in \mathbb{R}^{n \times 1},\ C \in \mathbb{R}^{5 \times n}$	[115]
`RAIL`	5177	M symmetric positive definite, A symmetric, $M^{-1}A$ stable, $B \in \mathbb{R}^{n \times 6},\ C \in \mathbb{R}^{7 \times n}$	[90]
`TRIDIAG(`α`)`	100	parameter $\alpha \in \mathbb{R}$, $A = \operatorname{tridiag}(\alpha, -1, -\alpha) \in \mathbb{R}^{n \times n}$, $\Lambda(A) = \left\{-1 + 2\,\lvert\alpha\rvert \cos\left(k\frac{\pi}{n+1}\right) \boldsymbol{\imath} \mid k = 1, \ldots, n\right\}$, A is stable and normal, $C^{\mathsf{T}} = B = \left[1, \ldots, 1\right]^{\mathsf{T}} \in \mathbb{R}^{n \times 1}$	[40, p. 15]

Tab. A.1: Information about the Benchmark Problems and the Running Example.

APPENDIX B
TRUNCATION OF A LOW-RANK PRODUCT

Algorithm B.1: Truncation of a Low-rank Product ZZ^{T}.

```
Input: real low-rank factor $Z \in \mathbb{R}^{n \times m}$ and truncation tolerance $tol_{\texttt{trunc}} > 0$
Output: real matrices $\tilde{Z}$, $\tilde{Q}$, $\tilde{S}$ such that $\tilde{Z} = \tilde{Q}\tilde{S}$, $\tilde{Q}$ has orthonormal columns,
        $\tilde{S}$ is diagonal, and $\left\| ZZ^{\mathsf{T}} - \tilde{Z}\tilde{Z}^{\mathsf{T}} \right\|_2 \leq tol_{\texttt{trunc}}^2 \left\| ZZ^{\mathsf{T}} \right\|_2$

% compute a compact SVD:
$\left[Q,\, S,\, \sim\right] := \texttt{svd}(Z, 0)$;
% determine index $k$:
$k := m$;
$\tau = tol_{\texttt{trunc}} \cdot S_{1,1}$;
for $i = 1, \dots, m$ do
    if $S_{i,i} < \tau$ then
        $k := i - 1$;
        break;
$k := \max\{1, k\}$;
% select the first $k$ columns of $Q$:
$\tilde{Q} := Q_{*,\{1,\dots,k\}}$;
% select the first $k$ rows and columns of $S$:
$\tilde{S} := S_{\{1,\dots,k\},\{1,\dots,k\}}$;
% compute matrix $\tilde{Z}$:
$\tilde{Z} := \tilde{Q}\tilde{S}$;
return $\tilde{Z}$, $\tilde{Q}$, $\tilde{S}$;
```

APPENDIX C

RADI METHOD

We review the RADI method for the numerical low-rank solution of the generalized ARE

$$A^{\mathsf{T}}XM + M^{\mathsf{T}}XA - M^{\mathsf{T}}XBB^{\mathsf{T}}XM + C^{\mathsf{T}}C = 0. \tag{C.1}$$

The RADI method is an iterative method generating a monotonic non-decreasing sequence of positive-semidefinite low-rank approximations. The RADI method can be seen as an extension of the low-rank ADI method (Algorithm 3.1) and requires a shift parameter in each iteration as well. It can be shown that the RADI method is equivalent to the *quadratic ADI method* [126, 127] and the *Cayley subspace iteration* [83]. The RADI method and a shift parameter heuristic is given in Algorithms C.1 and C.2, respectively.

Algorithm C.1: RADI Method for the Generalized ARE (C.1) [19, Alg. 2].

```
Input: matrices $A, M \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times b}$, $C \in \mathbb{R}^{c \times n}$ defining Equation (C.1), and
       a tolerance $0 < \varepsilon_{\texttt{rel}} \ll 1$ for the relative residual
Assumptions: $M$ is nonsingular, $(AM^{-1}, B)$ is stabilizable, $(AM^{-1}, CM^{-1})$ is
             detectable, and $b, c \ll n$
Output: real matrix $Z$ such that $ZZ^{\mathsf{T}} \approx X$, the absolute and relative residual
        $r_{\texttt{abs}}$ and $r_{\texttt{rel}}$

% initialization:
$W := C^{\mathsf{T}};\quad K = 0;\quad Z := [\,];\quad r_{\texttt{abs}} := \|W\|_2^2;\quad \gamma_C := r_{\texttt{abs}};$
% iterate until relative residual is smaller than $\varepsilon_{\texttt{rel}}$:
while $r_{\texttt{abs}} \geq \varepsilon_{\texttt{rel}} \gamma_C$ do
    obtain new shift parameter $\alpha$;
    $V := \big(A - BK^{\mathsf{T}} + \alpha E\big)^{-\mathsf{T}} W;$
    if $\Im(\alpha) = 0$ then
        $V_B := V^{\mathsf{T}} B;\quad Y := I + V_B V_B^{\mathsf{T}};$
        $L := \texttt{chol}(Y);\quad V := V L^{-1};\quad Z := \big[Z, \sqrt{-2\alpha} V\big];\quad \hat{V} := E^{\mathsf{T}} V L^{-\mathsf{T}};$
        $K := K - 2\alpha \hat{V};\quad W := W - 2\alpha \hat{V} V_B;$
    else
        $\alpha_a := |\alpha|;\quad \alpha_r := \Re(\alpha);\quad \alpha_i := \Im(\alpha);$
        $V_1 := \sqrt{-2\alpha_r} \Re(V);\quad V_2 := \sqrt{-2\alpha_r} \Im(V);\quad V_r := V_1^{\mathsf{T}} B;\quad V_i := V_2^{\mathsf{T}} B;$
        $F_1 := \begin{bmatrix} -\alpha_r/\alpha_a V_r - \alpha_i/\alpha_a V_i \\ \alpha_i/\alpha_a V_r - \alpha_r/\alpha_a V_i \end{bmatrix};\quad F_2 := \begin{bmatrix} V_r \\ V_i \end{bmatrix};\quad F_3 := \begin{bmatrix} \alpha_i/\alpha_a I_c \\ \alpha_r/\alpha_a I_c \end{bmatrix};$
        $Y := \begin{bmatrix} I_c & 0 \\ 0 & 1/2 I_c \end{bmatrix} - \frac{1}{4\alpha_r} F_1 F_1^{\mathsf{T}} - \frac{1}{4\alpha_r} F_2 F_2^{\mathsf{T}} - \frac{1}{2} F_3 F_3^{\mathsf{T}};$
        $L := \texttt{chol}(Y);\quad V := \begin{bmatrix} V_1 & V_2 \end{bmatrix} L^{-1};\quad \hat{V} := E^{\mathsf{T}} V L^{-\mathsf{T}};$
        $K := K + \hat{V} F_2;\quad W := W + \sqrt{-2\alpha_r} \hat{V}_{*,\{1,\dots,c\}};$
    % update absolute residual:
    $r_{\texttt{abs}} := \|W\|_2^2;$
% set relative residual:
$r_{\texttt{rel}} := r_{\texttt{abs}} / \gamma_C;$
return $Z$, $r_{\texttt{abs}}$, $r_{\texttt{rel}}$;
```

Heuristic C.2: Shift Parameter Heuristic for Algorithm C.1 line 3; [19, Sec. 4.5.1], [76, Sec. 2.1.3].

```
Input: matrices $A, M \in \mathbb{R}^{n\times n}, W \in \mathbb{R}^{n\times c}, B \in \mathbb{R}^{n\times b}, K \in \mathbb{R}^{b\times n}, Z \in \mathbb{R}^{n\times z}$ as
       in Algorithm C.1 line 3 and a number $l \in \mathbb{N}$
Output: shift parameter $\alpha \in \mathbb{C}_-$
if $Z$ is empty then
    $l := c; \quad \widehat{Z} := W;$
else
    $l := \min\{l, z\};$
    set $\widehat{Z}$ to the last $l$ columns of $Z$;
compute a reduced QR decomposition of $\hat{Z}$:
  $QR := \widehat{Z};$
$B_Q := Q^{\mathsf{T}}B; \quad A_Q := Q^{\mathsf{T}}AQ - B_QK^{\mathsf{T}}Q; \quad M_Q := Q^{\mathsf{T}}MQ; \quad R_Q := Q^{\mathsf{T}}W;$
$H := \begin{bmatrix} A_Q^{\mathsf{T}} & B_QB_Q^{\mathsf{T}} \\ R_QR_Q^{\mathsf{T}} & -A_Q \end{bmatrix}; \quad E := \begin{bmatrix} M_Q^{\mathsf{T}} & 0 \\ 0 & M_Q \end{bmatrix};$
compute the generalized eigenvalues and eigenvectors of $(H, E)$:
  $HV = EVD$ with $V = \begin{bmatrix} v_1, \dots, v_{2l} \\ w_1, \dots, w_{2l} \end{bmatrix}$ and $D = \operatorname{diag}(\alpha_1, \dots, \alpha_{2l});$
$\alpha := -1; \quad \gamma := 0;$
for $i = 1, \dots, 2l$ do
    if $\Re(\alpha_i) < 0$ then
        $\tau := \|w_i\|_2^2 \,/\, \left|w_i^{\mathsf{H}} M_Q v_i\right|;$
        if $\tau \geq \gamma$ then
            $\gamma := \tau; \quad \alpha := \alpha_i;$
return $\alpha$;
```

APPENDIX D

NUMERICAL RESULTS

Contents

D.1 Differential Lyapunov Equation

D.1.1 Computational Timings

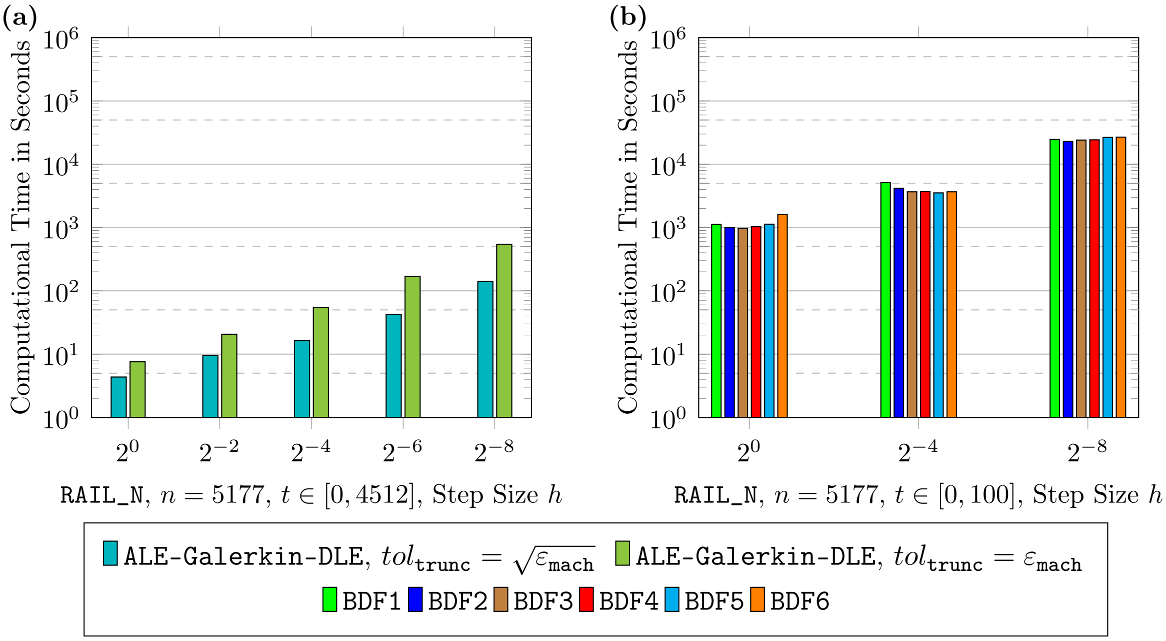


Fig. D.1: RAIL_N: **(a)** Timings for ALE-Galerkin-DLE. **(b)** Timings for BDF/ADI.

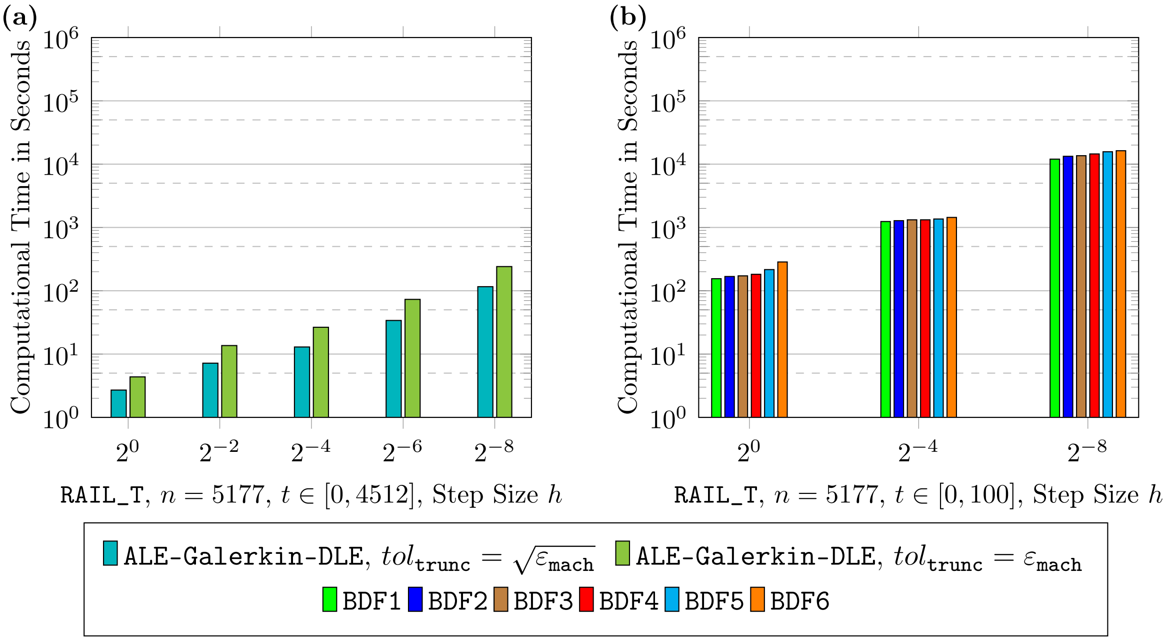


Fig. D.2: RAIL_T: **(a)** Timings for ALE-Galerkin-DLE. **(b)** Timings for BDF/ADI.

D.1.2 `ALE-Galerkin-DLE` (Algorithm 4.1)

D.1.2.1 `RAIL_N`

$$M\dot{X}(t)M^{\mathsf{T}} = AX(t)M^{\mathsf{T}} + MX(t)A^{\mathsf{T}} + BB^{\mathsf{T}},\ \ X(0) = 0.$$

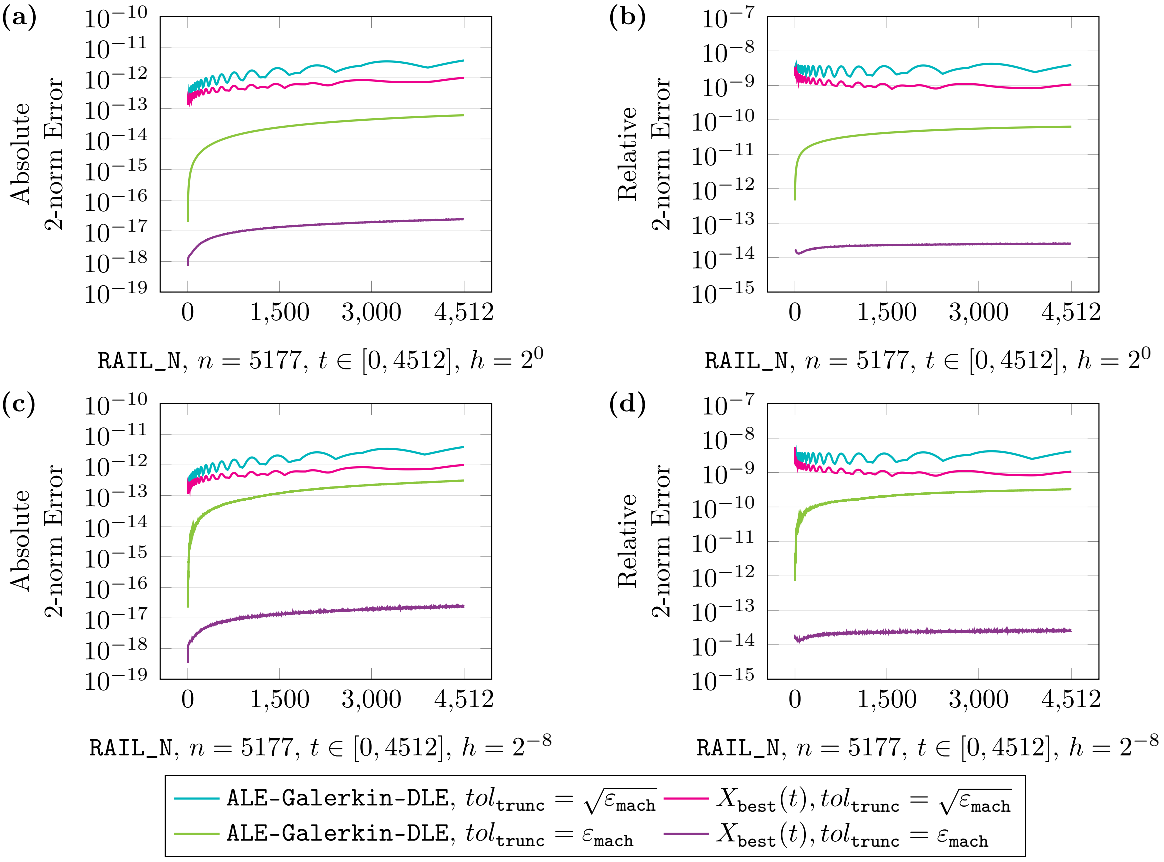


Fig. D.3: **(a), (c)** Absolute Error of the `ALE-Galerkin-DLE` and Best Approximation. **(b), (d)** Relative Error of the `ALE-Galerkin-DLE` and Best Approximation.

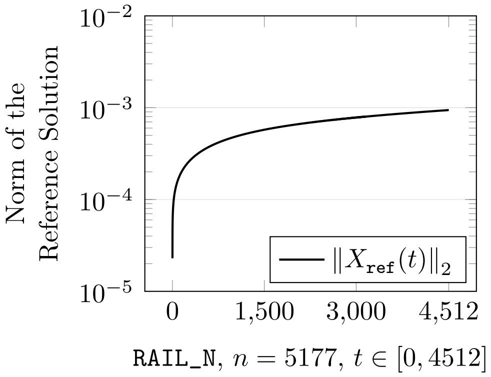


Fig. D.4: Norm of the Reference Solution.

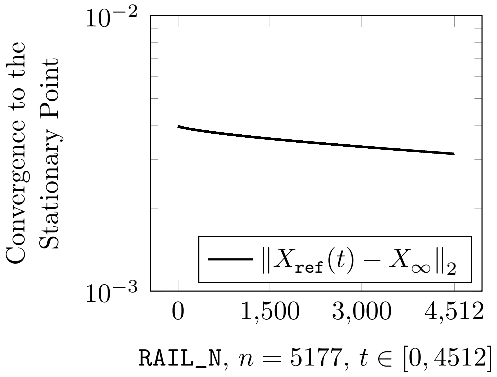


Fig. D.5: Convergence to the Stationary Point.

D.1.2.2 `RAIL_T`

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A + C^{\mathsf{T}}C,\ \ X(0) = 0.$$

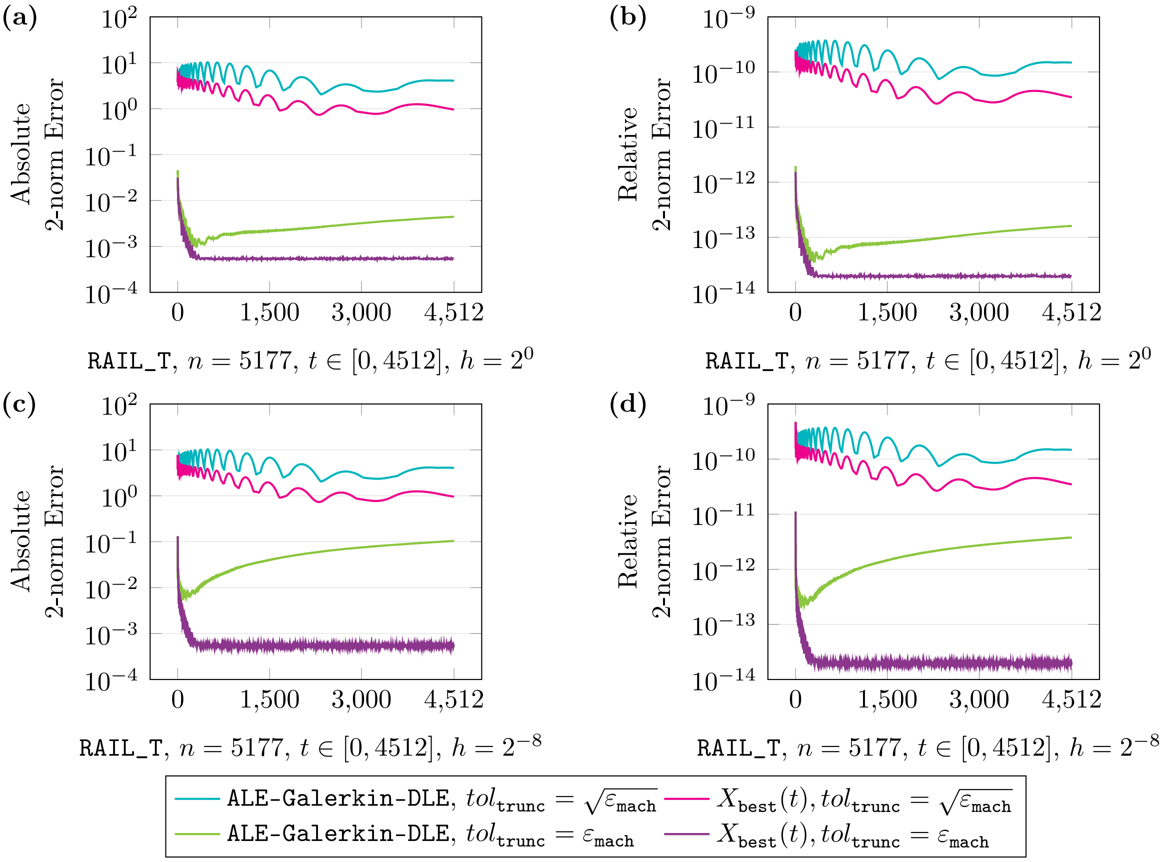


Fig. D.6: **(a), (c)** Absolute Error of the `ALE-Galerkin-DLE` and Best Approximation. **(b), (d)** Relative Error of the `ALE-Galerkin-DLE` and Best Approximation.

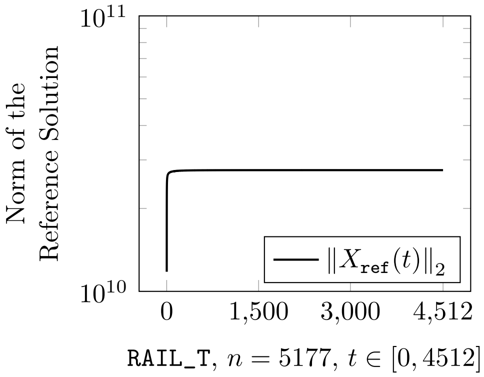


Fig. D.7: Norm of the Reference Solution.

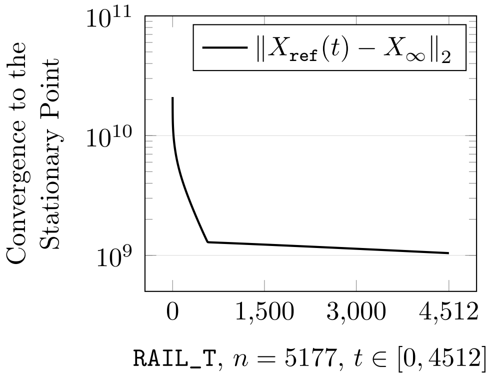
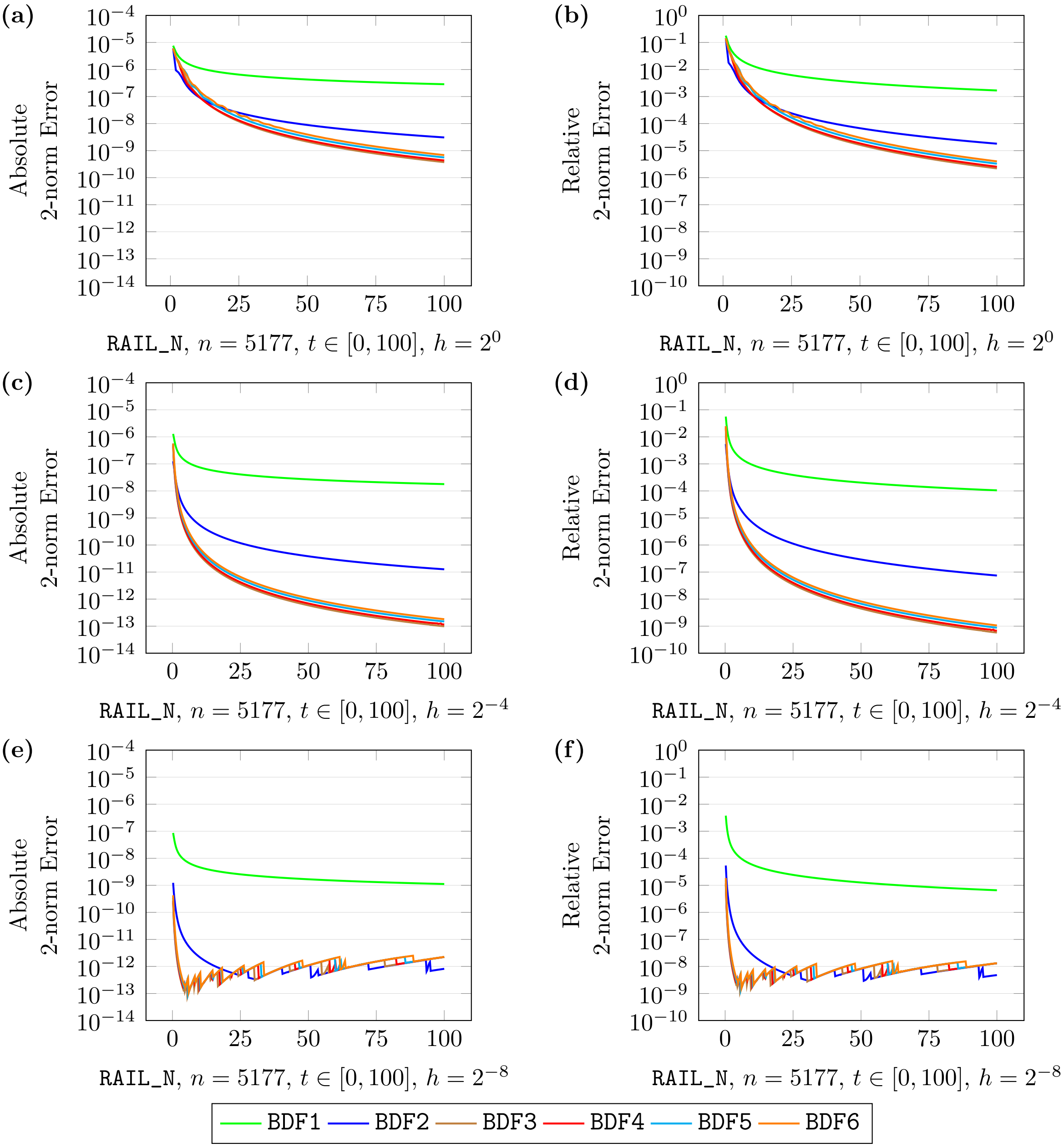


Fig. D.8: Convergence to the Stationary Point.

D.1.3 `BDF/ADI`

D.1.3.1 `RAIL_N`

$$M\dot{X}(t)M^{\mathsf{T}} = AX(t)M^{\mathsf{T}} + MX(t)A^{\mathsf{T}} + BB^{\mathsf{T}},\ X(0) = 0.$$



(a) `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{0}$

(b) `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{0}$

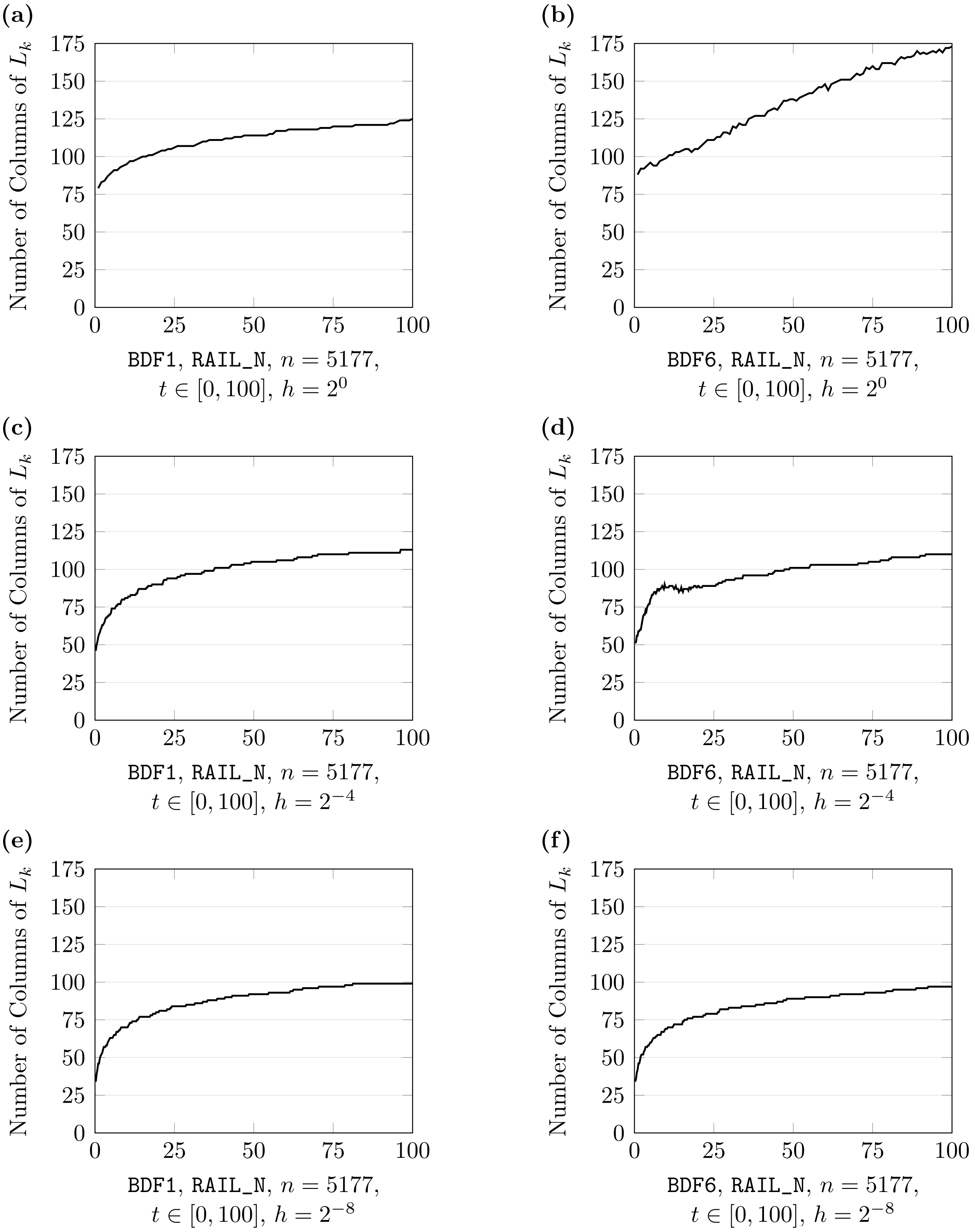
(c) `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

(d) `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

(e) `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

(f) `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

Fig. D.9: **(a), (c), (e)** Absolute Error of the `BDF/ADI` Approximation.
(b), (d), (f) Relative Error of the `BDF/ADI` Approximation.



(a) `BDF1`, `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^0$

(b) `BDF6`, `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^0$

(c) `BDF1`, `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

(d) `BDF6`, `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

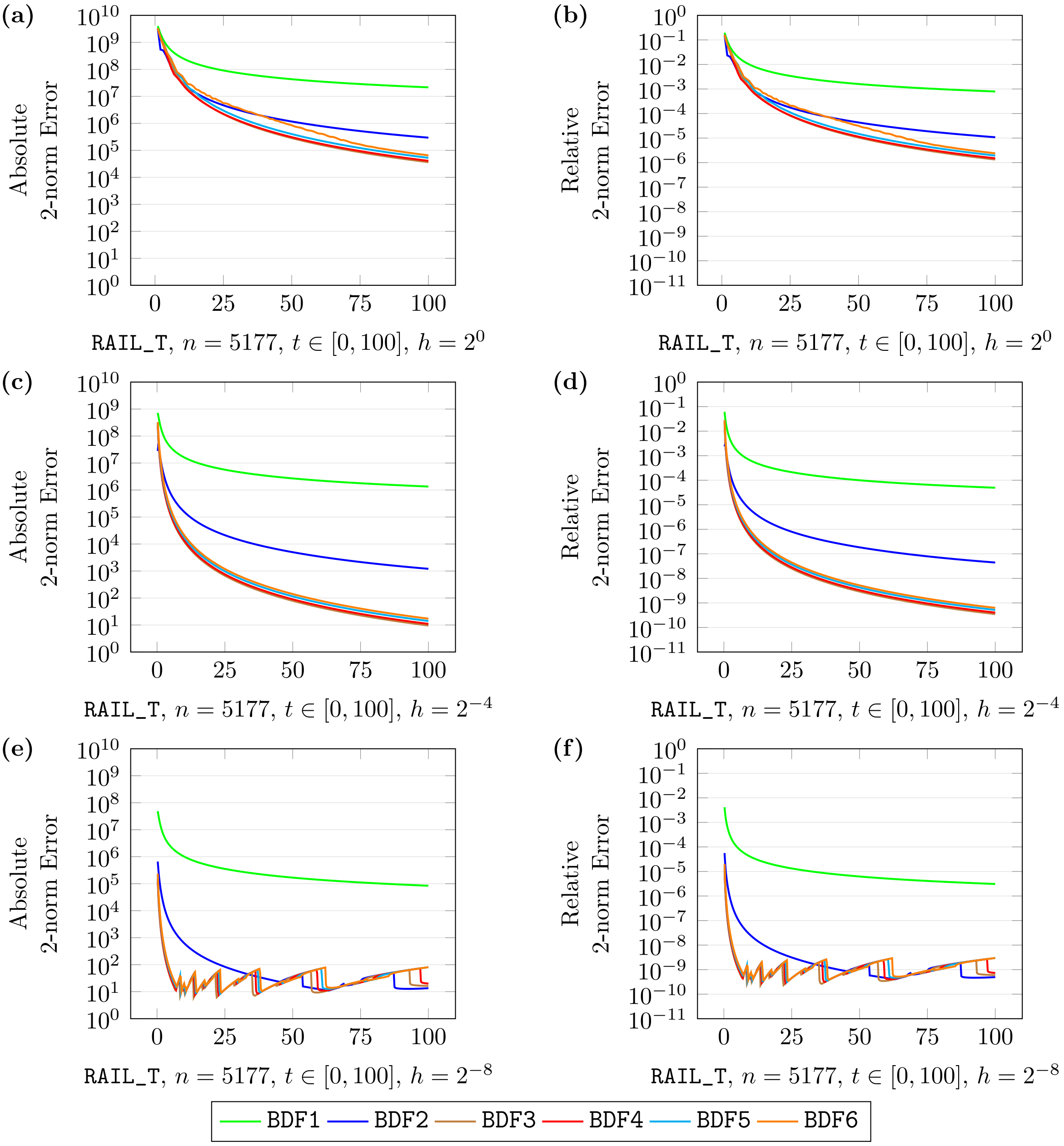
(e) `BDF1`, `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

(f) `BDF6`, `RAIL_N`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

Fig. D.10: **(a), (c), (e)** Number of Columns of the Low-rank Factor L_k of the `BDF1` Scheme. **(b), (d), (f)** Number of Columns of the Low-rank Factor L_k of the `BDF6` Scheme.

D.1.3.2 `RAIL_T`

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A + C^{\mathsf{T}}C,\ \ X(0) = 0.$$



`RAIL_T`, $n = 5177$, $t \in [0, 100]$, $h = 2^{0}$

`RAIL_T`, $n = 5177$, $t \in [0, 100]$, $h = 2^{0}$

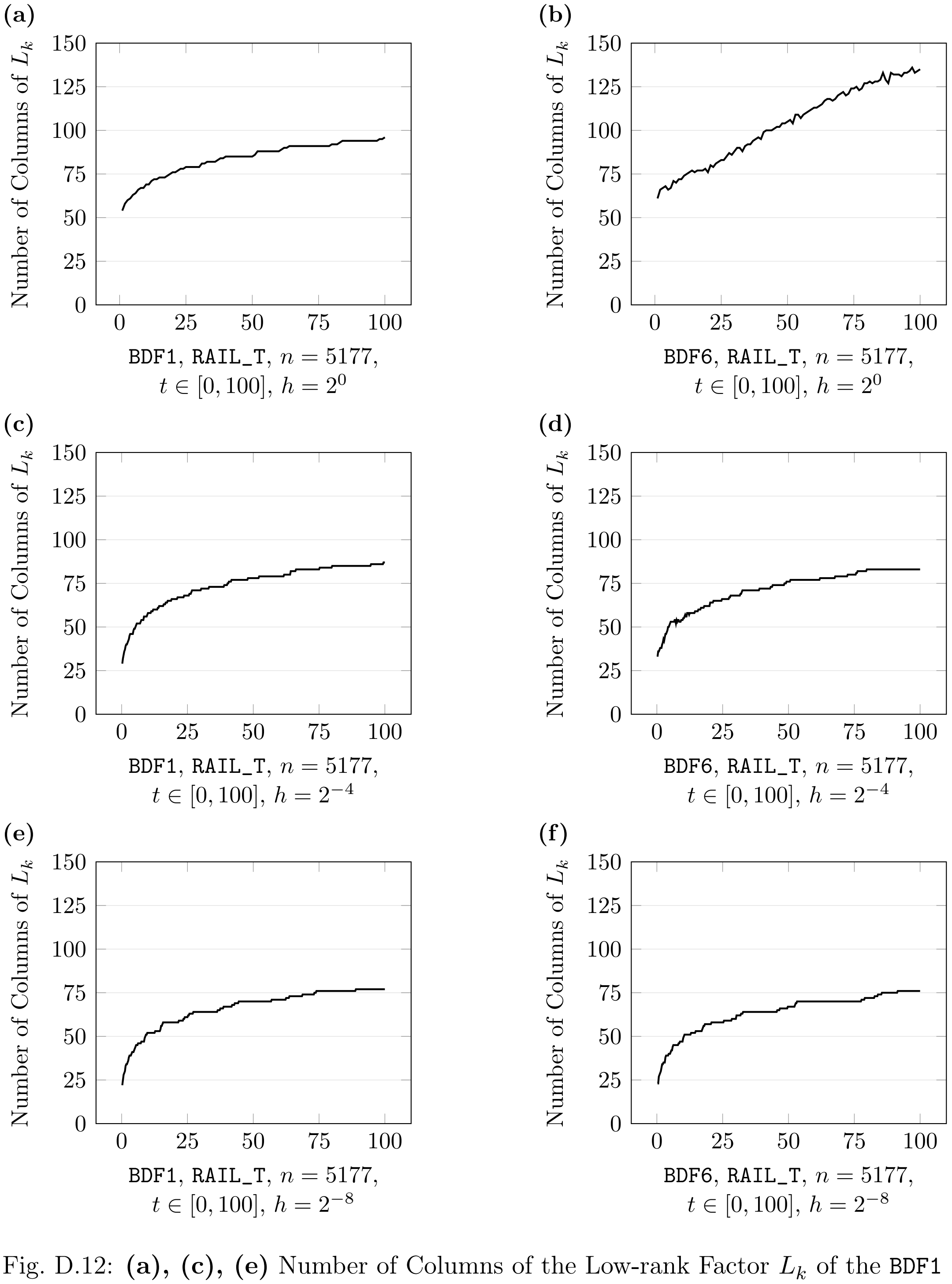
`RAIL_T`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

`RAIL_T`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

`RAIL_T`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

`RAIL_T`, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

Fig. D.11: **(a), (c), (e)** Absolute Error of the `BDF/ADI` Approximation.
(b), (d), (f) Relative Error of the `BDF/ADI` Approximation.



(a) BDF1, RAIL_T, $n = 5177$, $t \in [0, 100]$, $h = 2^0$

(b) BDF6, RAIL_T, $n = 5177$, $t \in [0, 100]$, $h = 2^0$

(c) BDF1, RAIL_T, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

(d) BDF6, RAIL_T, $n = 5177$, $t \in [0, 100]$, $h = 2^{-4}$

(e) BDF1, RAIL_T, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

(f) BDF6, RAIL_T, $n = 5177$, $t \in [0, 100]$, $h = 2^{-8}$

Fig. D.12: **(a), (c), (e)** Number of Columns of the Low-rank Factor L_k of the BDF1 Scheme. **(b), (d), (f)** Number of Columns of the Low-rank Factor L_k of the BDF6 Scheme.

D.2 Differential Riccati Equation

D.2.1 Computational Timings

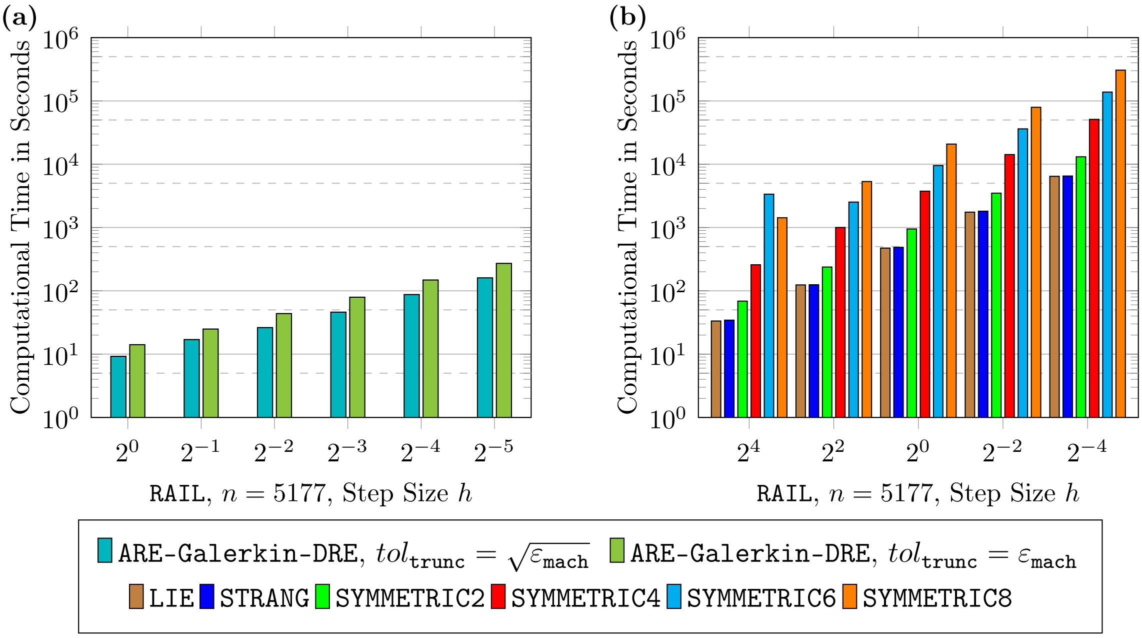


Fig. D.13: `RAIL`: **(a)** Timings for `ARE-Galerkin-DRE`. **(b)** Timings for Splitting Schemes.

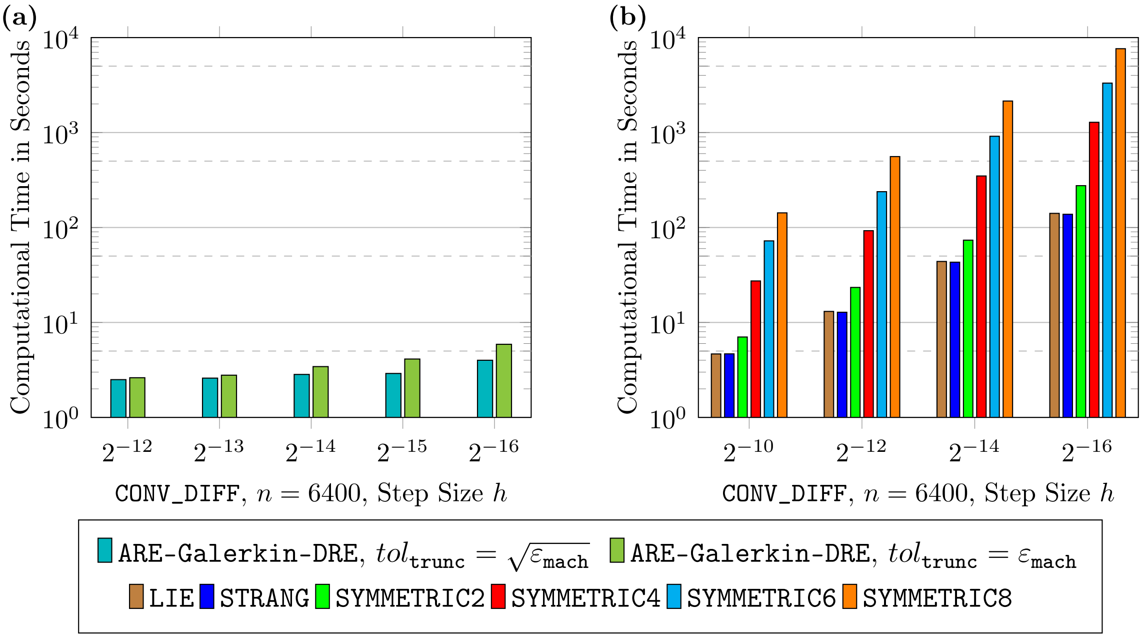


Fig. D.14: `CONV_DIFF`: **(a)** Timings for `ARE-Galerkin-DRE`. **(b)** Timings for Splitting Schemes.

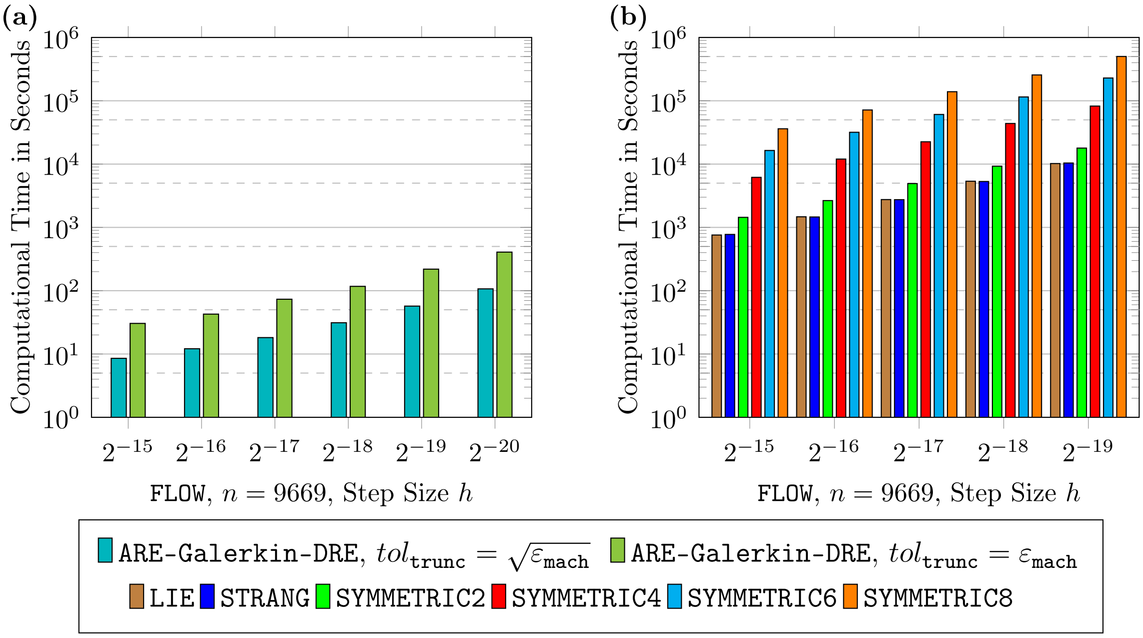


Fig. D.15: FLOW: **(a)** Timings for ARE-Galerkin-DRE. **(b)** Timings for Splitting Schemes.

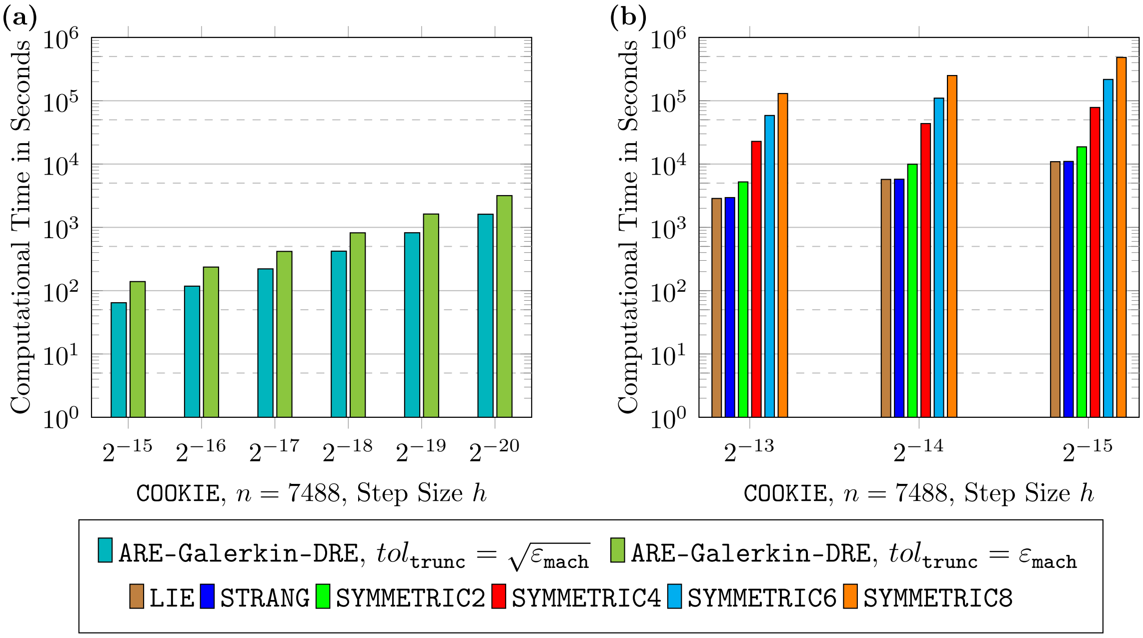
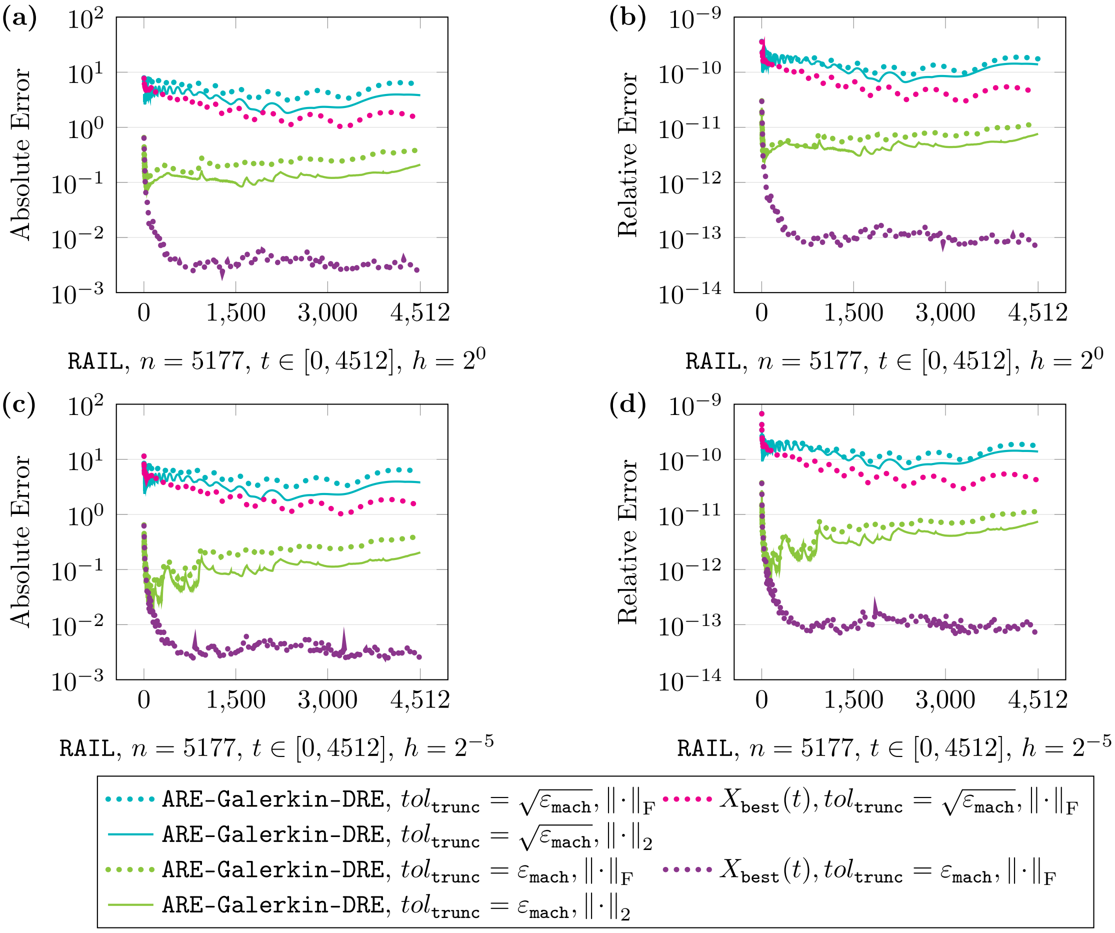


Fig. D.16: COOKIE: **(a)** Timings for ARE-Galerkin-DRE. **(b)** Timings for Splitting Schemes.

D.2.2 `ARE-Galerkin-DRE` (Algorithm 6.4)

D.2.2.1 `RAIL`

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C, \;\; X(0) = 0.$$



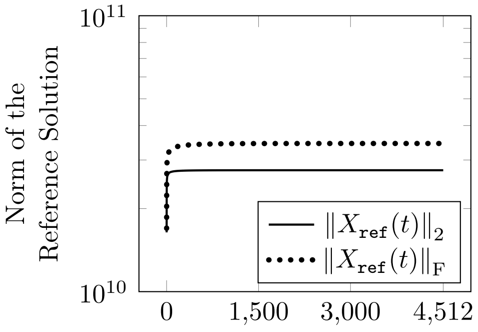
`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{0}$

`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{0}$

`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-5}$

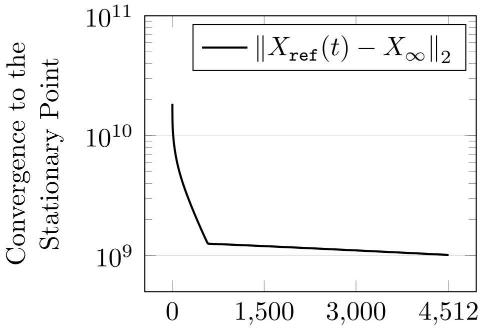
`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-5}$

Fig. D.17: **(a), (c)** Absolute Error of the `ARE-Galerkin-DRE` and Best Approximation. **(b), (d)** Relative Error of the `ARE-Galerkin-DRE` and Best Approximation.



`RAIL`, $n = 5177$, $t \in [0, 4512]$

Fig. D.18: Norm of the Reference Solution.



`RAIL`, $n = 5177$, $t \in [0, 4512]$

Fig. D.19: Convergence to the Stationary Point.

D.2.2.2 `CONV_DIFF`

$$\dot{X}(t) = A^{\mathsf{T}}X(t) + X(t)A - X(t)BB^{\mathsf{T}}X(t) + C^{\mathsf{T}}C,\ X(0) = 0.$$

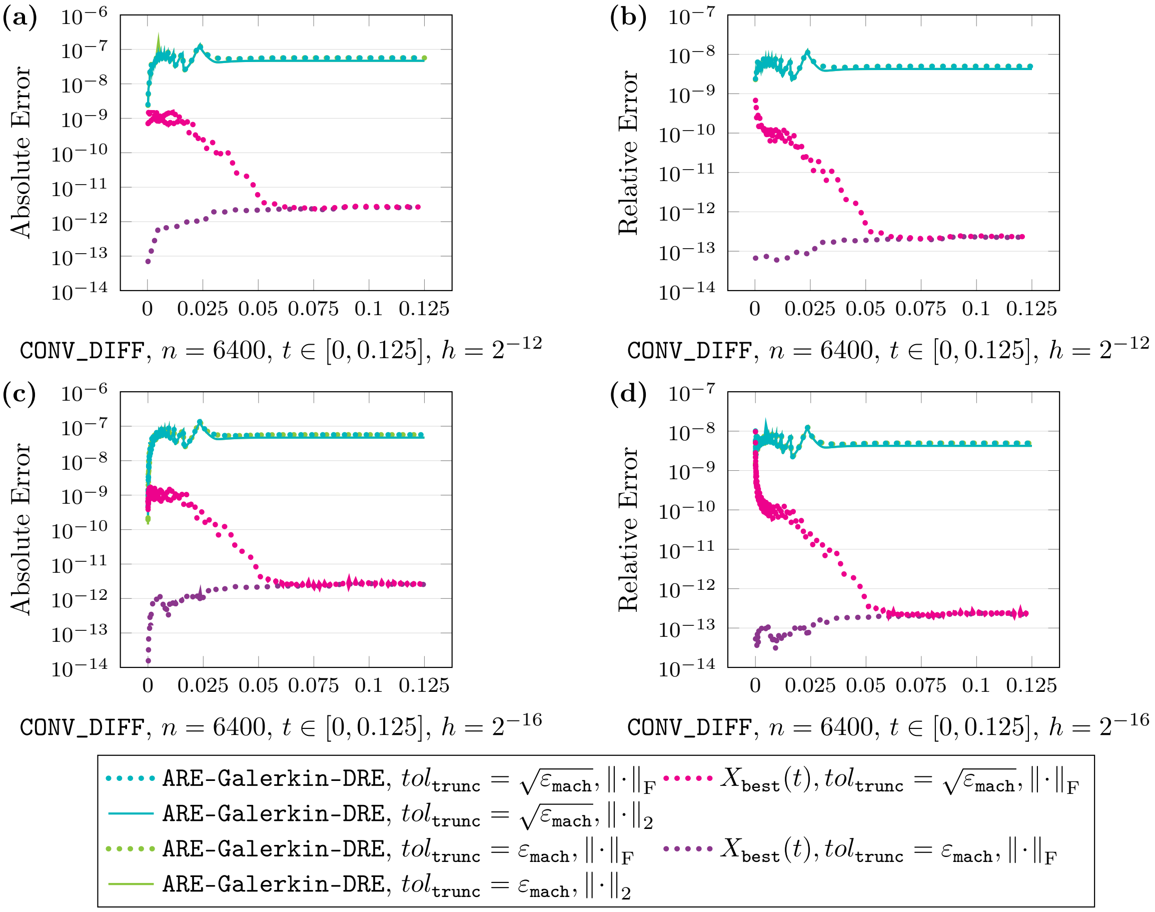


Fig. D.20: **(a), (c)** Absolute Error of the `ARE-Galerkin-DRE` and Best Approximation. **(b), (d)** Relative Error of the `ARE-Galerkin-DRE` and Best Approximation.

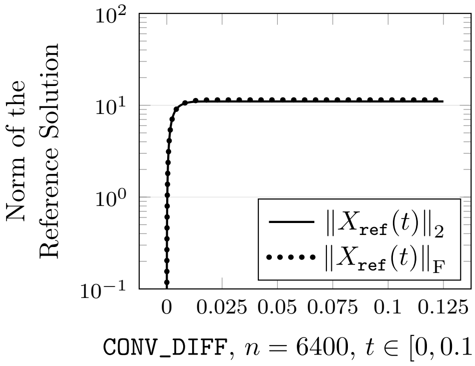


Fig. D.21: Norm of the Reference Solution.

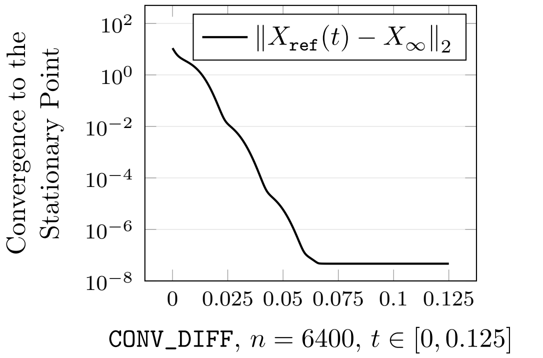
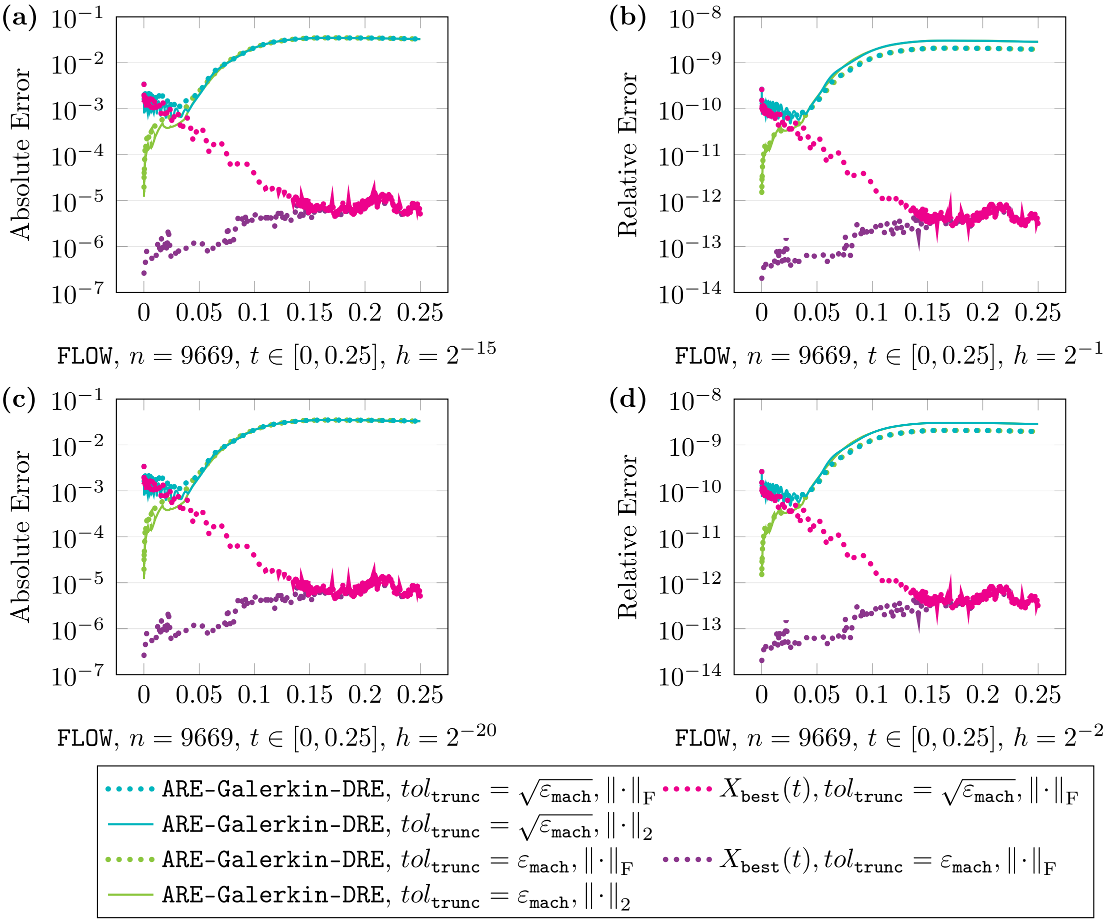


Fig. D.22: Convergence to the Stationary Point.

D.2.2.3 FLOW

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C,\ X(0) = 0.$$



(a) FLOW, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-15}$

(b) FLOW, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-15}$

(c) FLOW, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-20}$

(d) FLOW, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-20}$

Fig. D.23: **(a), (c)** Absolute Error of the ARE-Galerkin-DRE and Best Approximation.
(b), (d) Relative Error of the ARE-Galerkin-DRE and Best Approximation.

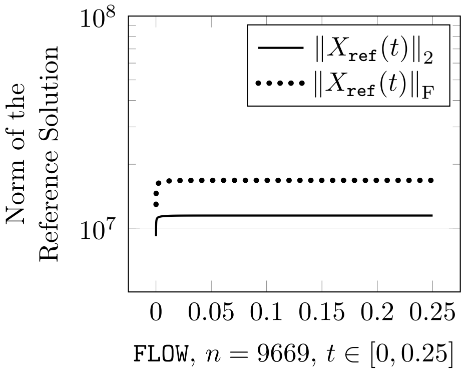


Fig. D.24: Norm of the Reference Solution.

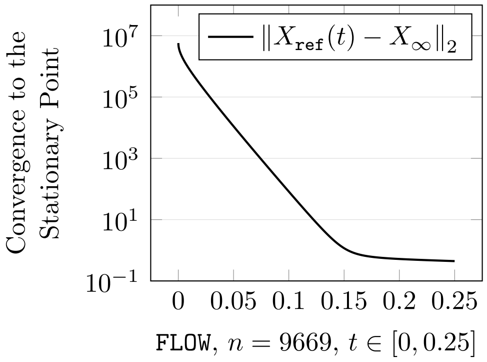
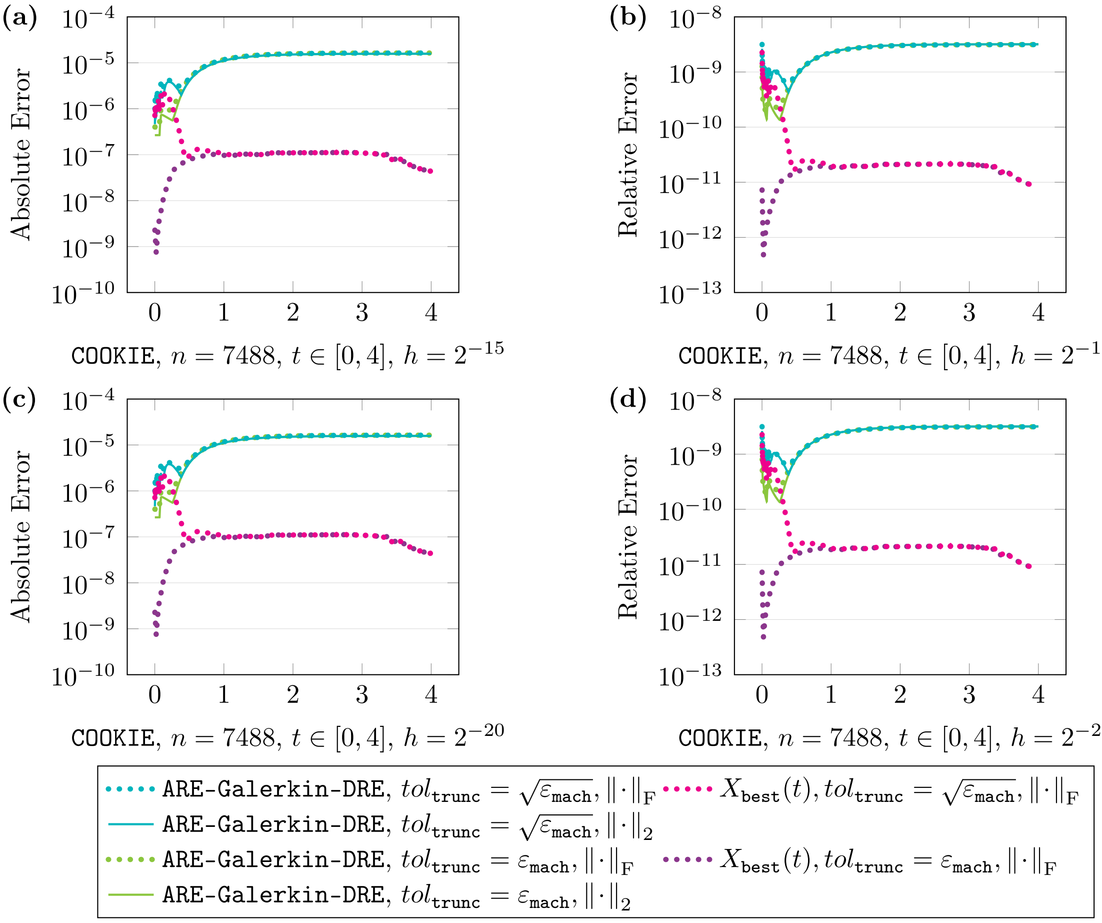


Fig. D.25: Convergence to the Stationary Point.

D.2.2.4 COOKIE

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C,\ X(0) = 0.$$



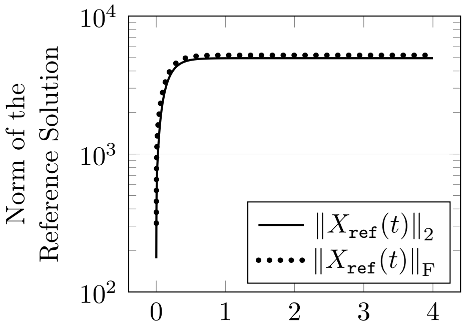
COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-15}$

COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-15}$

COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-20}$

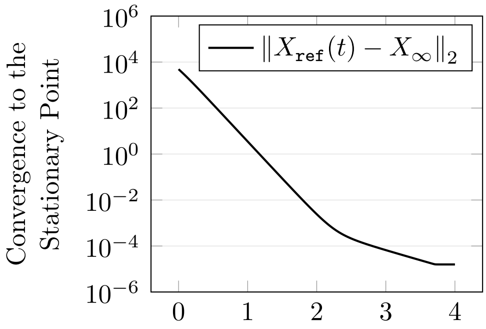
COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-20}$

Fig. D.26: **(a), (c)** Absolute Error of the ARE-Galerkin-DRE and Best Approximation. **(b), (d)** Relative Error of the ARE-Galerkin-DRE and Best Approximation.



COOKIE, $n = 7488$, $t \in [0, 4]$

Fig. D.27: Norm of the Reference Solution.



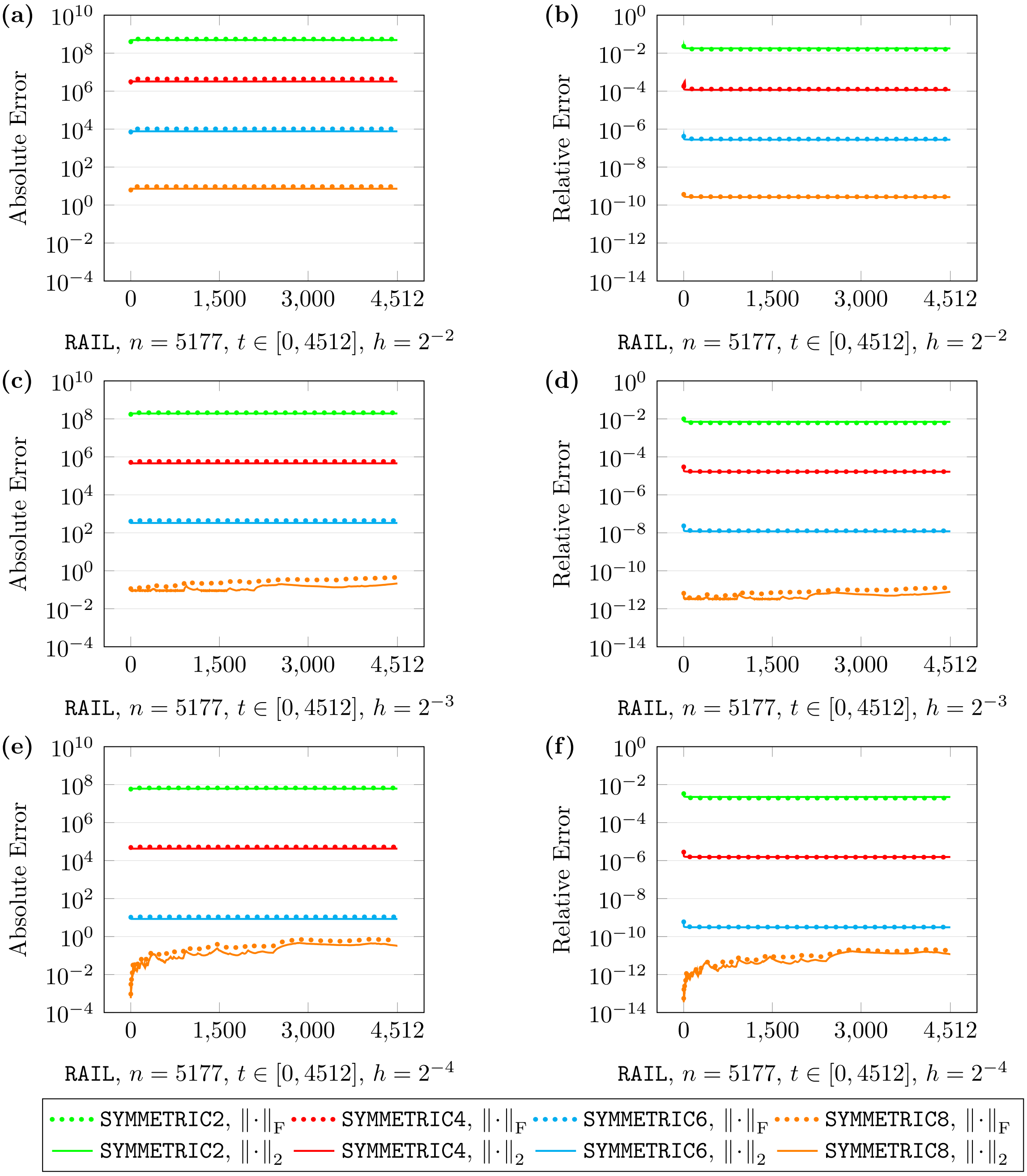
COOKIE, $n = 7488$, $t \in [0, 4]$

Fig. D.28: Convergence to the Stationary Point.

D.2.3 Splitting Schemes

D.2.3.1 `RAIL`

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C,\ \ X(0) = 0.$$



`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-2}$

`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-2}$

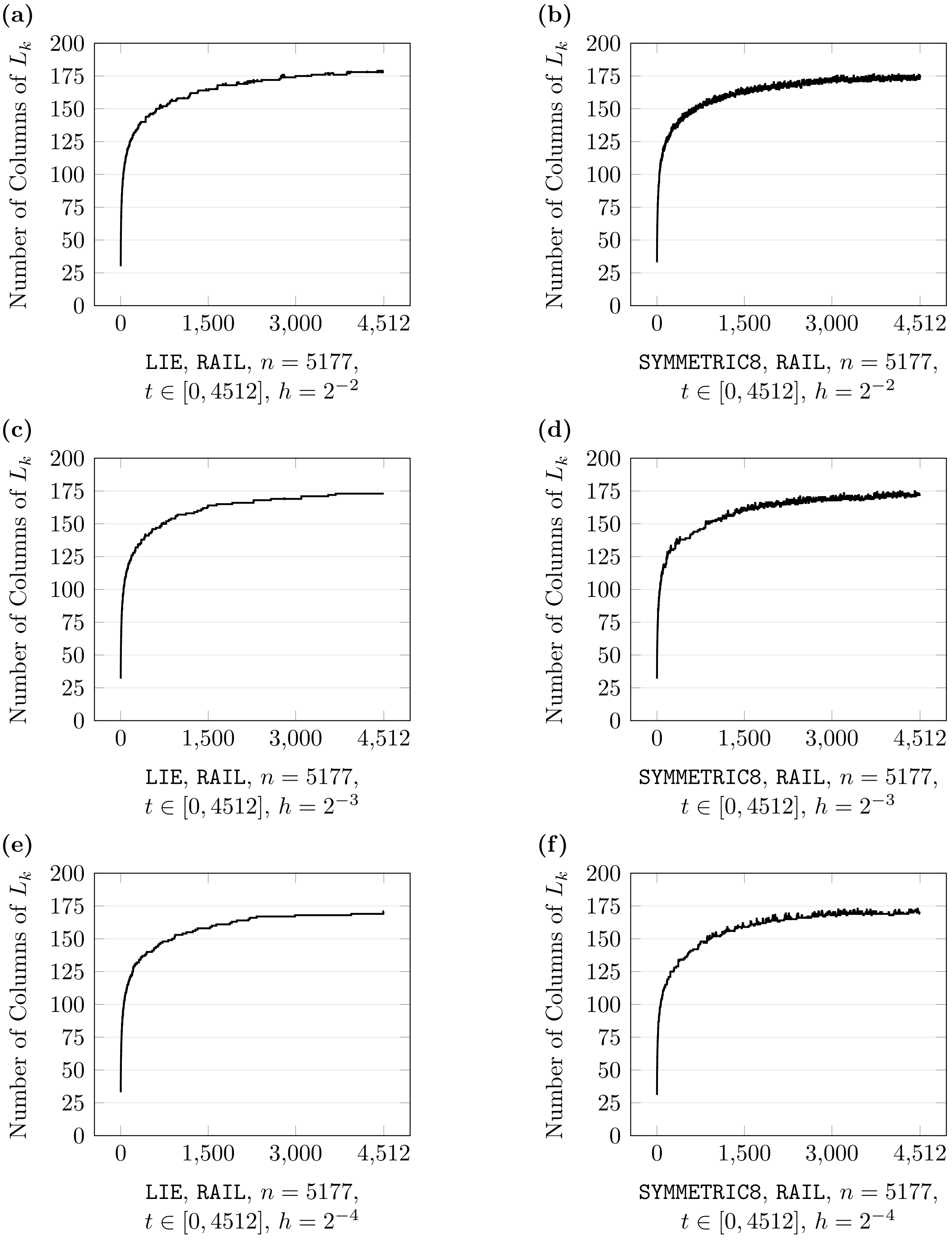
`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-3}$

`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-3}$

`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-4}$

`RAIL`, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-4}$

Fig. D.29: **(a), (c), (e)** Absolute Error of the Splitting Scheme Approximation.
(b), (d), (f) Relative Error of the Splitting Scheme Approximation.



(a) LIE, RAIL, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-2}$

(b) SYMMETRIC8, RAIL, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-2}$

(c) LIE, RAIL, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-3}$

(d) SYMMETRIC8, RAIL, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-3}$

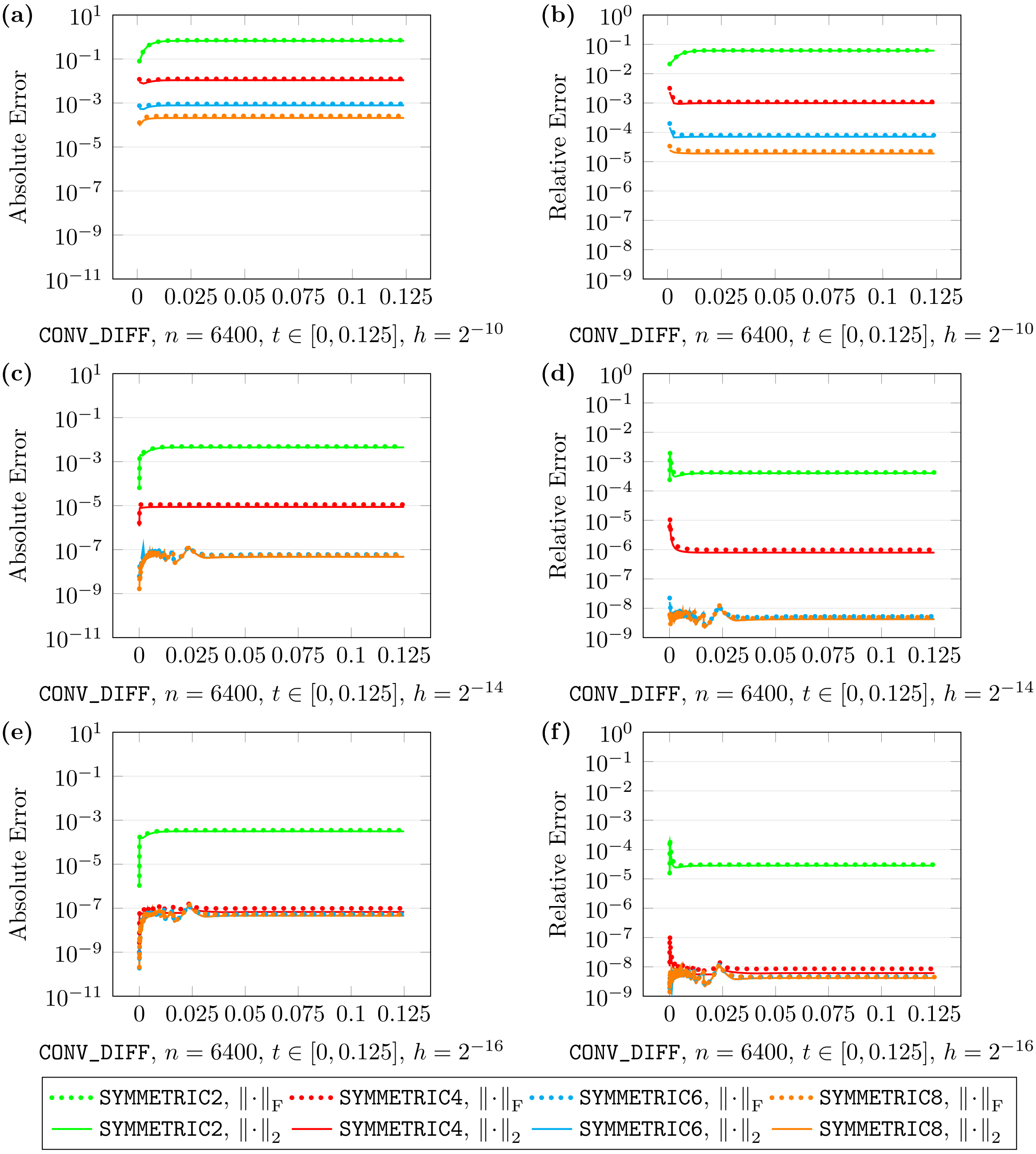
(e) LIE, RAIL, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-4}$

(f) SYMMETRIC8, RAIL, $n = 5177$, $t \in [0, 4512]$, $h = 2^{-4}$

Fig. D.30: **(a), (c), (e)** Number of Columns of the Low-rank Factor L_k of the LIE Scheme. **(b), (d), (f)** Number of Columns of the Low-rank Factor L_k of the SYMMETRIC8 Scheme.

D.2.3.2 `CONV_DIFF`

$$\dot{X}(t) = A^{\mathsf{T}}X(t) + X(t)A - X(t)BB^{\mathsf{T}}X(t) + C^{\mathsf{T}}C, \;\; X(0) = 0.$$



`CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-10}$

`CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-10}$

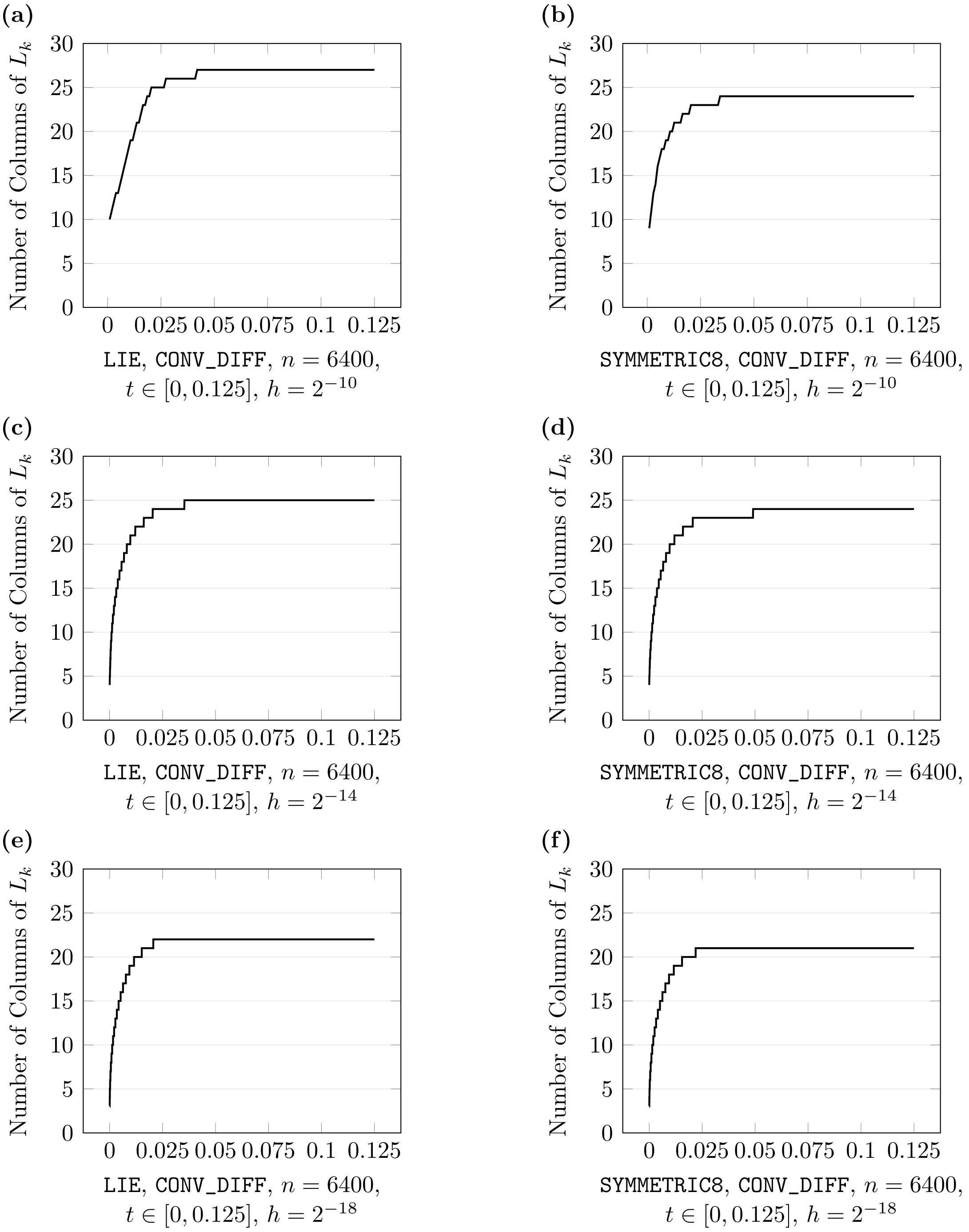
`CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-14}$

`CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-14}$

`CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-16}$

`CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-16}$

Fig. D.31: **(a), (c), (e)** Absolute Error of the Splitting Scheme Approximation.
(b), (d), (f) Relative Error of the Splitting Scheme Approximation.



`LIE`, `CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-10}$

`SYMMETRIC8`, `CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-10}$

`LIE`, `CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-14}$

`SYMMETRIC8`, `CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-14}$

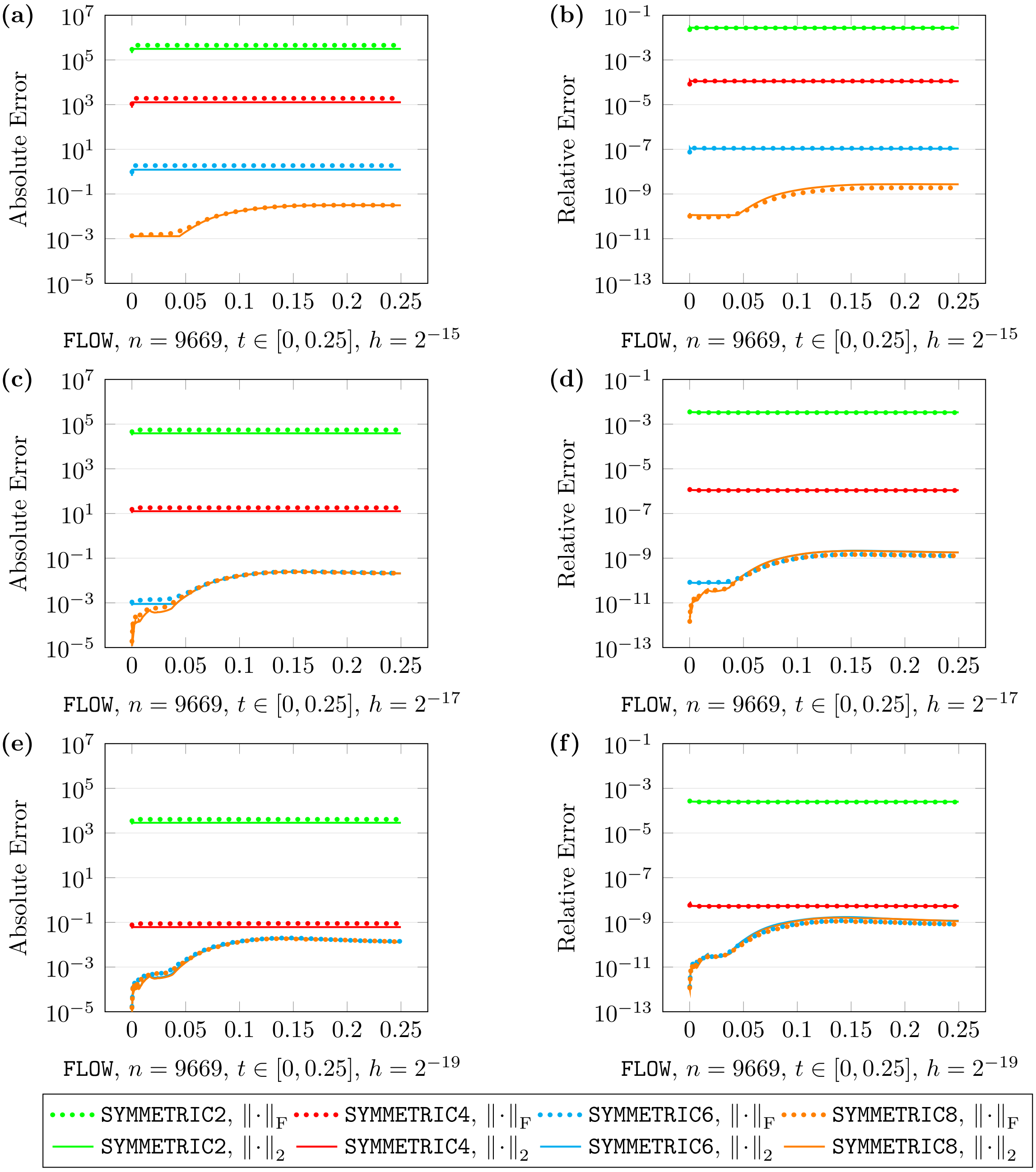
`LIE`, `CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-18}$

`SYMMETRIC8`, `CONV_DIFF`, $n = 6400$, $t \in [0, 0.125]$, $h = 2^{-18}$

Fig. D.32: **(a), (c), (e)** Number of Columns of the Low-rank Factor L_k of the `LIE` Scheme. **(b), (d), (f)** Number of Columns of the Low-rank Factor L_k of the `SYMMETRIC8` Scheme.

D.2.3.3 `FLOW`

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C,\ \ X(0) = 0.$$



(a) `FLOW`, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-15}$

(b) `FLOW`, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-15}$

(c) `FLOW`, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-17}$

(d) `FLOW`, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-17}$

(e) `FLOW`, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-19}$

(f) `FLOW`, $n = 9669$, $t \in [0, 0.25]$, $h = 2^{-19}$

Fig. D.33: **(a), (c), (e)** Absolute Error of the Splitting Scheme Approximation.
(b), (d), (f) Relative Error of the Splitting Scheme Approximation.

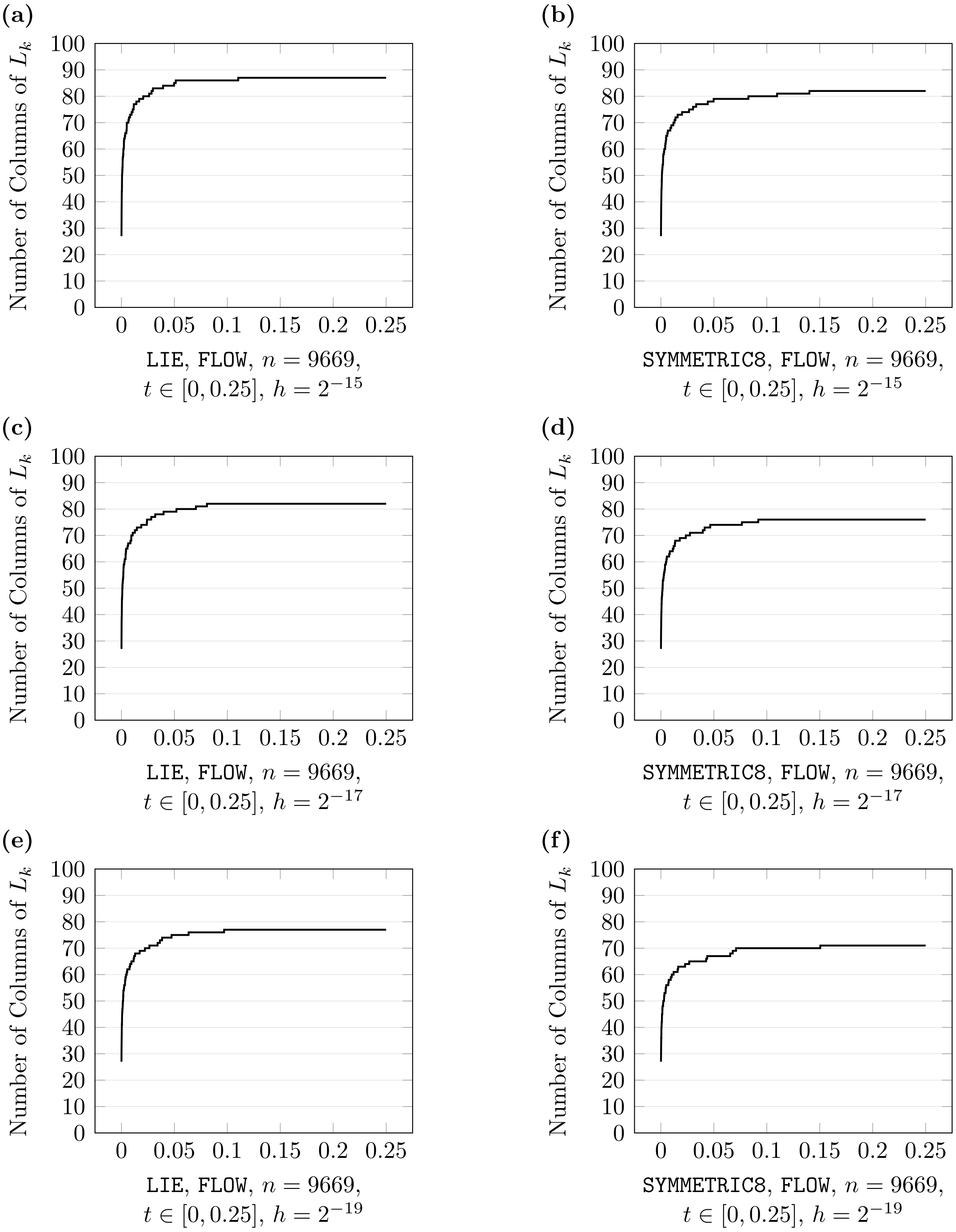
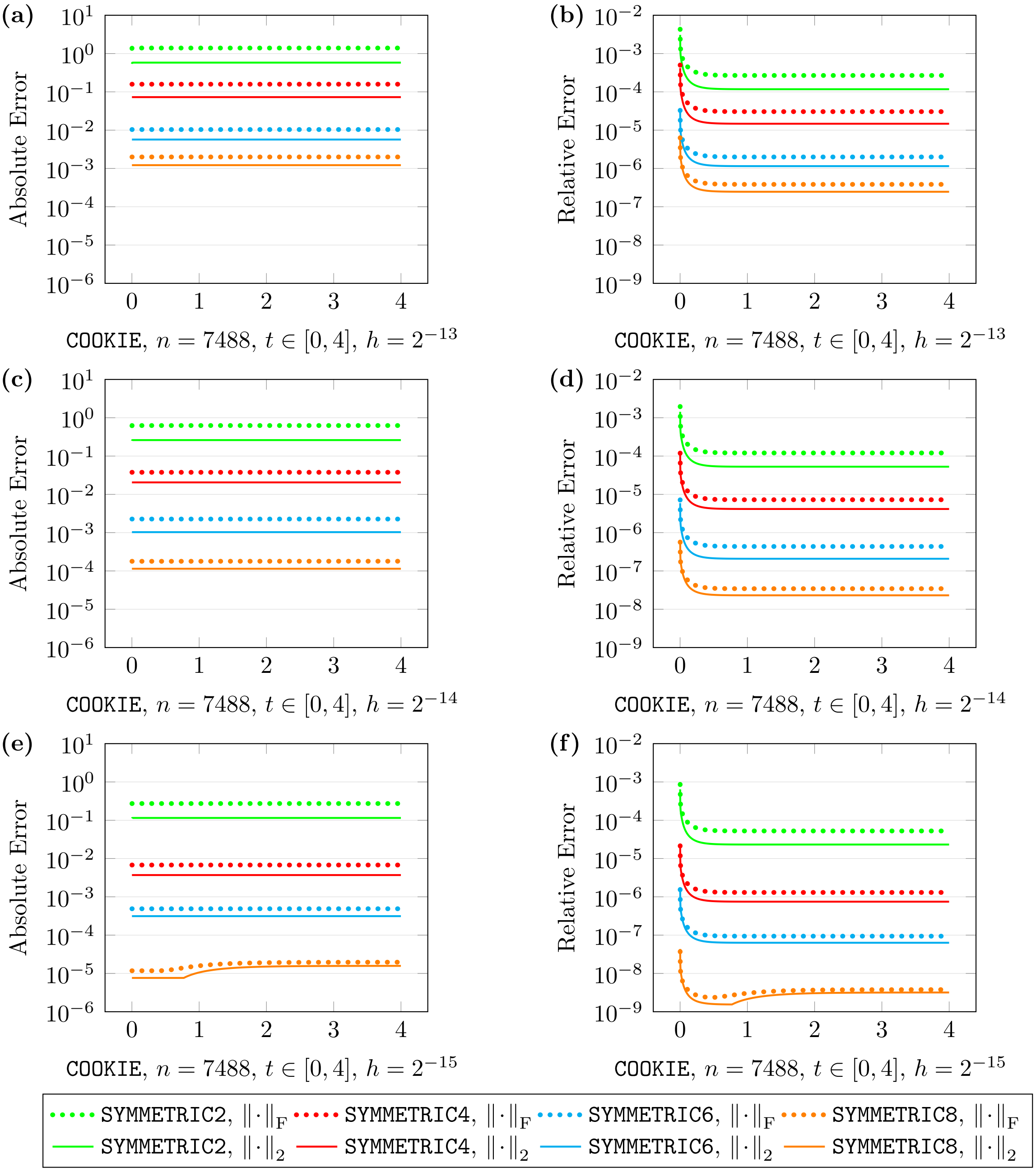


Fig. D.34: **(a), (c), (e)** Number of Columns of the Low-rank Factor L_k of the LIE Scheme. **(b), (d), (f)** Number of Columns of the Low-rank Factor L_k of the SYMMETRIC8 Scheme.

D.2.3.4 COOKIE

$$M^{\mathsf{T}}\dot{X}(t)M = A^{\mathsf{T}}X(t)M + M^{\mathsf{T}}X(t)A - M^{\mathsf{T}}X(t)BB^{\mathsf{T}}X(t)M + C^{\mathsf{T}}C,\ \ X(0) = 0.$$



COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-13}$

COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-13}$

COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-14}$

COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-14}$

COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-15}$

COOKIE, $n = 7488$, $t \in [0, 4]$, $h = 2^{-15}$

Fig. D.35: **(a), (c), (e)** Absolute Error of the Splitting Scheme Approximation.
(b), (d), (f) Relative Error of the Splitting Scheme Approximation.

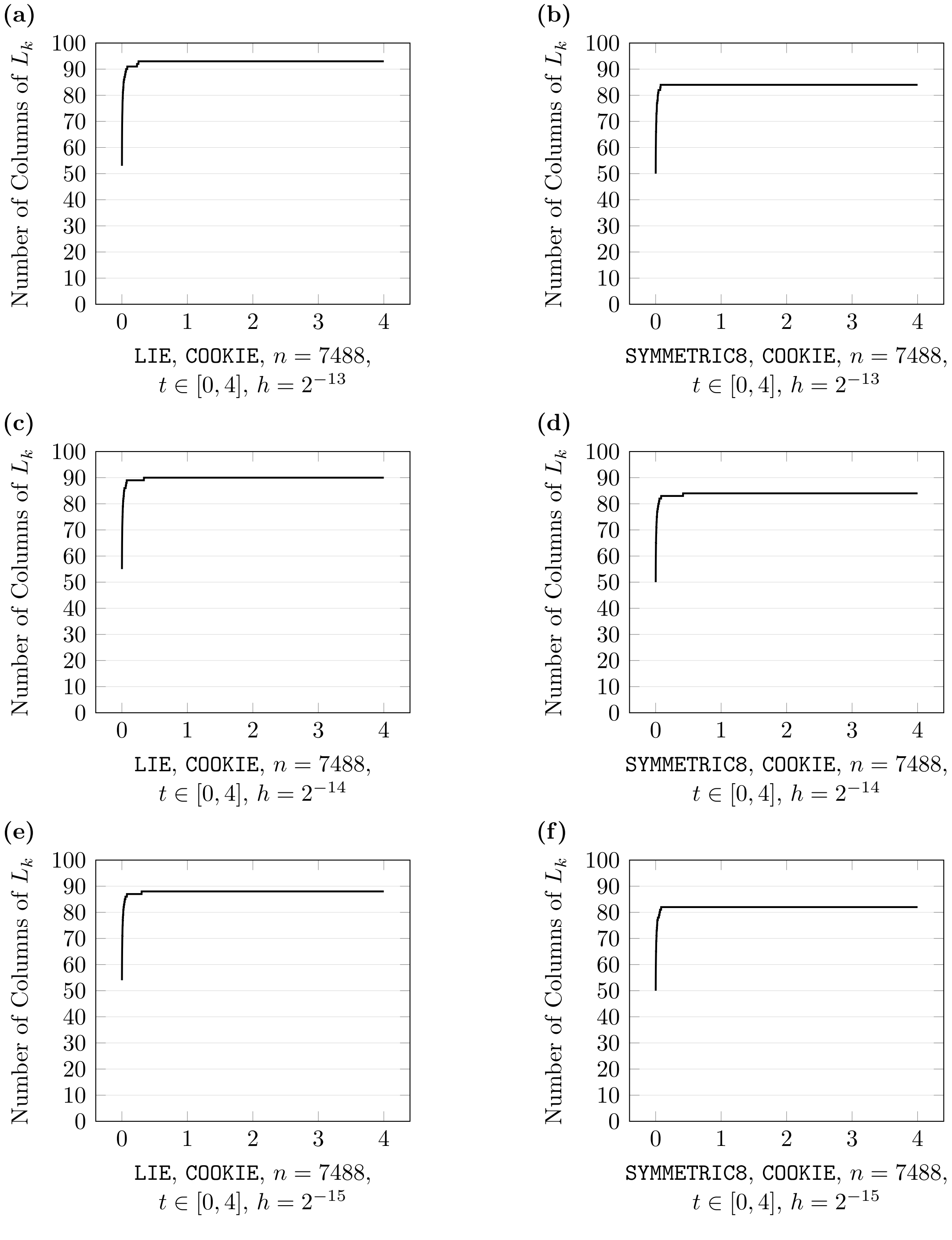


Fig. D.36: **(a), (c), (e)** Number of Columns of the Low-rank Factor L_k of the `LIE` Scheme. **(b), (d), (f)** Number of Columns of the Low-rank Factor L_k of the `SYMMETRIC8` Scheme.

BIBLIOGRAPHY

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank, *Matrix Riccati Equations in Control and Systems Theory*, Birkhäuser, 2003, https://doi.org/10.1007/978-3-0348-8081-7. 1, 19, 25, 40, 43, 53, 54, 55, 57, 58, 64, 65, 66, 67, 68, 70, 77, 78, 108

[2] N. I. Akhiezer, *Elements of the Theory of Elliptic Functions*, vol. 79 of Transl. of Math. Monographs, American Mathematical Society, 1990, https://doi.org/10.1090/mmono/079. 33

[3] H. W. Alt, *Linear Functional Analysis*, Universitext, Springer-Verlag, London, 2016, https://doi.org/10.1007/978-1-4471-7280-2. 7, 9, 11

[4] L. Amodei and J.-M. Buchot, *An invariant subspace method for large-scale algebraic Riccati equation*, Appl. Numer. Math., 60 (2010), pp. 1067–1082, https://doi.org/10.1016/j.apnum.2009.09.006. 61

[5] B. D. O. Anderson and J. B. Moore, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971. 78

[6] V. Angelova, M. Hached, and K. Jbilou, *Approximate solutions to large nonsymmetric differential Riccati problems with applications to transport theory*, Numer. Lin. Alg. Appl., 27 (2020), p. e2272, https://doi.org/10.1002/nla.2272. 71

[7] A. C. Antoulas, *Approximation of Large–Scale Dynamical Systems*, vol. 6 of Adv. Des. Control, SIAM Publications, Philadelphia, 2005, https://doi.org/10.1137/1.9780898718713. 23, 25, 39

[8] A. C. Antoulas, D. C. Sorensen, and Y. Zhou, *On the decay rate of Hankel singular values and related issues*, Syst. Cont. Lett., 46 (2002), pp. 323–342, https://doi.org/10.1016/S0167-6911(02)00147-0. 34

[9] J. Baker, M. Embree, and J. Sabino, *Fast Singular Value Decay for Lyapunov Solutions with Nonnormal Coefficients*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 656–668, https://doi.org/10.1137/140993867. 34

[10] S. Barrachina, P. Benner, and E. S. Quintana-Ortí, *Efficient algorithms for generalized algebraic Bernoulli equations based on the matrix sign function*,

Numer. Algorithms, 46 (2007), pp. 351–368, https://doi.org/10.1007/s11075-007-9143-x. 61

[11] M. Beck and S. J. A. Malham, *Computing the Maslov index for large systems*, Proc. Amer. Math. Soc., 143 (2015), pp. 2159–2173, https://doi.org/10.1090/S0002-9939-2014-12575-5. 1

[12] B. Beckermann and A. Gryson, *Extremal Rational Functions on Symmetric Discrete Sets and Superlinear Convergence of the ADI Method*, Constr. Approx., 32 (2010), pp. 393–428, https://doi.org/10.1007/s00365-010-9087-6. 32

[13] B. Beckermann and A. Townsend, *Bounds on the Singular Values of Matrices with Displacement Structure*, SIAM Rev., 61 (2019), pp. 319–344, https://doi.org/10.1137/19M1244433. 2, 32, 33, 34

[14] M. Behr, P. Benner, and J. Heiland, *On an Invariance Principle for the Solution Space of the Differential Riccati Equation*, Proc. Appl. Math. Mech., 18 (2018), p. e201800031, https://doi.org/10.1002/pamm.201800031. iii, 61

[15] M. Behr, P. Benner, and J. Heiland, *Solution Formulas for Differential Sylvester and Lyapunov Equations*, Calcolo, 56 (2019), p. 51, https://doi.org/10.1007/s10092-019-0348-x. iii, 25, 26, 28, 42, 153

[16] M. Behr, P. Benner, and J. Heiland, *Galerkin trial spaces and Davison–Maki methods for the numerical solution of differential Riccati equations*, Appl. Math. Comp., 410 (2021), p. 126401, https://doi.org/10.1016/j.amc.2021.126401. iii, 154

[17] P. Benner and T. Breiten, *Low rank methods for a class of generalized Lyapunov equations and related issues*, Numer. Math., 124 (2013), pp. 441–470, https://doi.org/10.1007/s00211-013-0521-0. 32

[18] P. Benner and Z. Bujanović, *On the solution of large-scale algebraic Riccati equations by using low-dimensional invariant subspaces*, Linear Algebra Appl., 488 (2016), pp. 430–459, https://doi.org/10.1016/j.laa.2015.09.027. 2, 32, 61

[19] P. Benner, Z. Bujanović, P. Kürschner, and J. Saak, *RADI: A low-rank ADI-type algorithm for large scale algebraic Riccati equations*, Numer. Math., 138 (2018), pp. 301–330, https://doi.org/10.1007/s00211-017-0907-5. 1, 35, 37, 116, 117

[20] P. Benner, J.-R. Li, and T. Penzl, *Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems*, Numer. Lin. Alg. Appl., 15 (2008), pp. 755–777, https://doi.org/10.1002/nla.622. 35

[21] P. Benner, R.-C. Li, and N. Truhar, *On the ADI method for Sylvester equations*, J. Appl. Math. Comput., 233 (2009), pp. 1035–1045, https://doi.org/10.1016/j.cam.2009.08.108. 35

[22] P. Benner and H. Mena, *Numerical solution of the infinite-dimensional LQR-problem and the associated differential Riccati equations*, J. Numer. Math., 26 (2018), pp. 1–20, https://doi.org/10.1515/jnma-2016-1039. 1

[23] R. Bhatia, *Matrix Analysis*, vol. 169 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1997, https://doi.org/10.1007/978-1-4612-0653-8. 29

[24] D. A. Bini, B. Iannazzo, and B. Meini, *Numerical Solution of Algebraic Riccati Equations*, vol. 9 of Fundamentals of Algorithms, SIAM Publications, Philadelphia, 2012, https://doi.org/10.1137/1.9781611972092. 19, 20

[25] Å. Björck, *Numerical Methods in Matrix Computations*, vol. 59 of Texts in Applied Mathematics, Springer International Publishing, 2015, https://doi.org/10.1007/978-3-319-05089-8. 20

[26] R. W. Brockett, *Finite Dimensional Linear Systems*, vol. 74 of Classics in Applied Mathematics, SIAM Publications, Philadelphia, 2015, https://doi.org/10.1137/1.9781611973884. Reprint of the 1970 original. 1, 23, 34, 39, 43

[27] J. C. Butcher, *On the implementation of implicit Runge–Kutta methods*, BIT Numer. Math., 16 (1976), pp. 237–240, https://doi.org/10.1007/bf01932265. 103

[28] R. Byers and S. Nash, *On the Singular "Vectors" of the Lyapunov Operator*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 59–66, https://doi.org/10.1137/0608003. 25

[29] M. Caliari, P. Kandolf, A. Ostermann, and S. Rainer, *The Leja Method Revisited: Backward Error Analysis for the Matrix Exponential*, SIAM J. Sci. Comput., 38 (2016), pp. A1639–A1661, https://doi.org/10.1137/15M1027620. 102

[30] F. M. Callier and J. L. Willems, *Criterion for the Convergence of the Solution of the Riccati Differential Equation*, IEEE Trans. Autom. Control, 26 (1981), pp. 1232–1242, https://doi.org/10.1109/TAC.1981.1102812. 76

[31] F. M. Callier and J. L. Willems, *The Infinite Horizon and the Receding Horizon LQ-Problems with Partial Stabilization Constraints*, in The Riccati Equation, S. Bittanti, A. J. Laub, and J. C. Willems, eds., Springer-Verlag, Berlin Heidelberg, 1991, pp. 243–262, https://doi.org/10.1007/978-3-642-58223-3_9. 76

[32] F. M. Callier, J. L. Willems, and J. Winkin, *Convergence of the time-invariant Riccati differential equation and LQ-problem: mechanisms of attraction*, Internat. J. Control, 59 (1994), pp. 983–1000, https://doi.org/10.1080/00207179408923113. 76, 77

[33] F. M. Callier and J. Winkin, *Convergence of the Time-Invariant Riccati Differential Equation towards Its Strong Solution for Stabilizable Systems*, J. Math. Anal. Appl., 192 (1995), pp. 230–257, https://doi.org/10.1006/jmaa.1995.1169. 76

[34] J. L. Casti, *Linear Dynamical Systems*, vol. 135 of Mathematics in Science and Engineering, Academic Press, second ed., 1987. 74

[35] C. H. Choi, *A Survey of Numerical Methods for Solving Matrix Riccati Differential Equations*, in IEEE Proceedings on Southeastcon, 1990, pp. 696–700 vol.2, https://doi.org/10.1109/SECON.1990.117906. 79

[36] M. C. D'Autilia, I. Sgura, and V. Simoncini, *Matrix-oriented discretization methods for reaction-diffusion PDEs: Comparisons and applications*, Comput. Math. with Appl., 79 (2020), pp. 2067–2085, https://doi.org/10.1016/j.camwa.2019.10.020. 23

[37] E. J. Davison and M. C. Maki, *The Numerical Solution of the Matrix Riccati Differential Equation*, IEEE Trans. Autom. Control, 18 (1973), pp. 71–73, https://doi.org/10.1109/tac.1973.1100210. 79, 80, 81

[38] E. de Souza and S. P. Bhattacharyya, *Controllability, Observability and the Solution of $AX - XB = C$*, Linear Algebra Appl., 39 (1981), pp. 167–188, https://doi.org/10.1016/0024-3795(81)90301-3. 29, 35

[39] C. A. Desoer and M. Vidyasagar, *Feedback Systems*, vol. 55 of Classics in Applied Mathematics, SIAM Publications, Philadelphia, 2009, https://doi.org/10.1137/1.9780898719055. 12

[40] M. Embree and R. Carden, *Pseudospectra and Nonnormal Dynamical Systems*, 2012, https://www.tu-ilmenau.de/fileadmin/media/math/Tagungen/ElgSchool2012/lecture4.pdf (accessed 2021-02-01). 2, 112

[41] D. Fortunato and A. Townsend, *Fast Poisson solvers for spectral methods*, IMA J. Numer. Anal., 40 (2020), pp. 1994–2018, https://doi.org/10.1093/imanum/drz034. 23, 32, 33

[42] Z. Gajić and M. T. J. Qureshi, *Lyapunov Matrix Equation in System Stability and Control*, Math. in Science and Engineering, Academic Press, San Diego, 1995. 23, 25, 31, 39

[43] J. Garloff, *Bounds for the eigenvalues of the solution of the discrete Riccati and Lyapunov equations and the continuous Lyapunov equation*, Internat. J. Control, 43 (1986), pp. 423–431, https://doi.org/10.1080/00207178608933475. 31

[44] L. Grasedyck, *Existence and Computation of Low Kronecker-Rank Approximations for Large Linear Systems of Tensor Product Structure*, Computing, 72 (2004), pp. 247–265, https://doi.org/10.1007/s00607-003-0037-z. 32

[45] L. Grasedyck, *Existence of a low rank or $\mathcal{H}$-matrix approximant to the solution of a Sylvester equation*, Numer. Lin. Alg. Appl., 11 (2004), pp. 371–389, https://doi.org/10.1002/nla.366. 32

[46] L. Grasedyck, W. Hackbusch, and B. N. Khoromskij, *Solution of Large Scale Algebraic Matrix Riccati Equations by Use of Hierarchical Matrices*, Computing, 70 (2003), pp. 121–165, https://doi.org/10.1007/s00607-002-1470-0. 32

[47] W. Greub, *Linear Algebra*, Springer-Verlag, New York, fourth ed., 1975. Graduate Texts in Mathematics, No. 23. 21

[48] L. Grubišić and D. Kressner, *On the eigenvalue decay of solutions to operator Lyapunov equations*, Syst. Cont. Lett., 73 (2014), pp. 42–47, https://doi.org/10.1016/j.sysconle.2014.09.006. 32

[49] Y. Güldoğan, M. Hached, K. Jbilou, and M. Kurulay, *Low-rank approximate solutions to large-scale differential matrix Riccati equations*, Applicationes Mathematicae, 45 (2018), pp. 233–254, https://doi.org/10.4064/am2355-1-2018. 3, 71

[50] S. Güttel, *Rational Krylov Methods for Operator Functions*, Dissertation, Technische Universität Bergakademie Freiberg, Germany, 2010, https://nbn-resolving.org/urn:nbn:de:bsz:105-qucosa-27645. 102

[51] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I*, vol. 8 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin Heidelberg, 1987, https://doi.org/10.1007/978-3-662-12607-3. 51

[52] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin Heidelberg, 1991, https://doi.org/10.1007/978-3-662-09947-6. 103

[53] P. R. Halmos, *Positive Approximants of Operators*, Indiana Univ. Math. J., 21 (1971/72), pp. 951–960, https://doi.org/10.1512/iumj.1972.21.21076. 10

[54] S. Hein, *MPC-LQG-Based Optimal Control of Parabolic PDEs*, Dissertation, Technische Universität Chemnitz, Chemnitz, Germany, Feb. 2009, http://nbn-resolving.de/urn:nbn:de:bsz:ch1-201000134. 61

[55] N. J. Higham, *Computing a Nearest Symmetric Positive Semidefinite Matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118, https://doi.org/10.1016/0024-3795(88)90223-6. 9, 10

[56] N. J. Higham, *Functions of Matrices: Theory and Computation*, Applied Mathematics, SIAM Publications, Philadelphia, 2008, https://doi.org/10.1137/1.9780898717778. 10, 12, 23

[57] M. Hochbruck and C. Lubich, *On Krylov Subspace Approximations to the Matrix Exponential Operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925, https://doi.org/10.1137/S0036142995280572. 102

[58] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1991, https://doi.org/10.1017/cbo9780511840371. 13, 18

[59] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, second ed., 2013, https://doi.org/10.1017/cbo9781139020411. 6, 7, 24, 101

[60] A. Iske, *Approximation Theory and Algorithms for Data Analysis*, vol. 68 of Texts in Applied Mathematics, Springer International Publishing, 2018, https://doi.org/10.1007/978-3-030-05228-7. 12

[61] A. Jameson, *Solution of the Equation $AX + XB = C$ by Inversion of an $M \times M$ or $N \times N$ Matrix*, SIAM J. Appl. Math., 16 (1968), pp. 1020–1023, https://doi.org/10.1137/0116083. 25

[62] W. Kaballo, *Grundkurs Funktionalanalysis*, Springer-Verlag, Berlin Heidelberg, second ed., 2018, https://doi.org/10.1007/978-3-662-54748-9. 8

[63] R. E. Kalman and T. S. Englar, *A user's manual for the automatic synthesis program*, RIAS Report CR-475, 1966. 79, 82

[64] C. Kenney and R. B. Leipnik, *Numerical integration of the differential matrix Riccati equation*, IEEE Trans. Autom. Control, 30 (1985), pp. 962–970, https://doi.org/10.1109/tac.1985.1103822. 79, 82

[65] G. Kirsten and V. Simoncini, *Order reduction methods for solving large-scale differential matrix Riccati equations*, SIAM J. Sci. Comput., 42 (2020), pp. A2182–A2205, https://doi.org/10.1137/19M1264217. 3, 71

[66] H.-W. Knobloch and H. Kwakernaak, *Lineare Kontrolltheorie*, Springer-Verlag, Berlin Heidelberg, 1985, https://doi.org/10.1007/978-3-642-69884-2, https://doi.org/10.1007/978-3-642-69884-2. 1, 53, 66, 70, 74, 75

[67] L. Kohaupt, *Solution of the matrix eigenvalue problem $VA + A^*V = \mu V$ with applications to the study of free linear dynamical systems*, J. Comput. Appl. Math., 213 (2008), pp. 142–165, https://doi.org/10.1016/j.cam.2007.01.001. 25

[68] M. Konstantinov, D. W. Gu, V. Mehrmann, and P. H. Petkov, *Perturbation Theory for Matrix Equations*, vol. 9 of Stud. Comput. Math., Elsevier, Amsterdam, first edition ed., 2003. 25

[69] M. Konstantinov, V. Mehrmann, and P. Petkov, *On properties of Sylvester and Lyapunov operators*, Linear Algebra Appl., 312 (2000), pp. 35–71, https://doi.org/10.1016/S0024-3795(00)00082-3. 25

[70] A. Koskela and H. Mena, *Analysis of Krylov Subspace Approximation to Large Scale Differential Riccati Equations*, e-print arXiv:1705.07507v4, 2018. math.NA. 3, 71

[71] D. Kressner and C. Tobler, *Krylov Subspace Methods for Linear Systems with Tensor Product Structure*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 1688–1714, https://doi.org/10.1137/090756843. 32

[72] C. S. Kubrusly, *Spectral Theory of Bounded Linear Operators*, Springer International Publishing, 2020, https://doi.org/10.1007/978-3-030-33149-8. 26

[73] V. Kučera, *A review of the matrix Riccati equation*, Kybernetica (Prague), 9 (1973), pp. 42–61. 57

[74] P. Kürschner, *Efficient Low-Rank Solution of Large-Scale Matrix Equations*, Dissertation, Otto-von-Guericke-Universität Magdeburg, Germany, 2016, http://hdl.handle.net/11858/00-001M-0000-002A-710D-1 (accessed 2021-02-01). 1, 35, 36

[75] P. Kürschner, *Balanced truncation model order reduction in limited time intervals for large systems*, Adv. Comput. Math., 44 (2018), pp. 1821–1844, https://doi.org/10.1007/s10444-018-9608-6. 104

[76] P. Kürschner, *Approximate residual-minimizing shift parameters for the low-rank ADI iteration*, Electron. Trans. Numer. Anal., 51 (2019), pp. 240–261, https://doi.org/10.1553/etna_vol51s240. 35, 37, 117

[77] W. H. Kwon, Y. S. Moon, and S. C. Ahn, *Bounds in algebraic Riccati and Lyapunov equations: a survey and some new results*, Internat. J. Control, 64 (1997), pp. 377–389, https://doi.org/10.1080/00207179608921634. 31

[78] P. Lancaster, *Explicit Solutions of Linear Matrix Equations*, SIAM Rev., 12 (1970), pp. 544–566, https://doi.org/10.1137/1012104. 29

[79] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York, 1995. 75

[80] N. Lang, *Numerical Methods for Large-Scale Linear Time-Varying Control Systems and related Differential Matrix Equations*, Dissertation, Technische Universität Chemnitz, Germany, 2017, https://www.logos-verlag.de/cgi-bin/buch/isbn/4700 (accessed 2021-02-01). 2, 51, 52

[81] A. J. Laub, *Schur techniques for Riccati differential equations*, in Feedback Control of Linear and Nonlinear Systems, D. Hinrichsen and A. Isidori, eds., Springer-Verlag, New York, 1982, pp. 165–174, https://doi.org/10.1007/bfb0006827. 79

[82] J.-R. Li and J. White, *Low-Rank Solution of Lyapunov Equations*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 260–280, https://doi.org/10.1137/S0895479801384937. 1, 35

[83] Y. Lin and V. Simoncini, *A new subspace iteration method for the algebraic Riccati equation*, Numer. Lin. Alg. Appl., 22 (2015), pp. 26–47, https://doi.org/https://doi.org/10.1002/nla.1936. 115

[84] A. Locatelli, *Optimal Control: An Introduction*, Birkhäuser, Basel, Switzerland, 2001, https://doi.org/10.1007/978-3-0348-8328-3. 1, 53, 55, 58

[85] T. McCauley, *Computing the Maslov index from singularities of a matrix Riccati equation*, J. Dyn. Diff. Equat., 29 (2017), pp. 1487–1502, https://doi.org/10.1007/s10884-016-9568-9. 1

[86] B. P. Molinari, *The Time-Invariant Linear-Quadratic Optimal Control Problem*, Automatica J. IFAC, 13 (1977), pp. 347–357, https://doi.org/10.1016/0005-1098(77)90017-6. 58, 59, 61

[87] I. Moret and P. Novati, *An interpolatory Approximation of the matrix exponential based on Faber polynomials*, J. Comput. Appl. Math., 131 (2001), pp. 361–380, https://doi.org/10.1016/S0377-0427(00)00261-2. 102

[88] T. Mori and I. A. Derese, *A brief summary of the bounds on the solution of the algebraic matrix equations in control theory*, Internat. J. Control, 39 (1984), pp. 247–256, https://doi.org/10.1080/00207178408933163. 31

[89] T. Mori, N. Fukuma, and M. Kuwahara, *Explicit Solution and Eigenvalue Bounds in the Lyapunov Matrix Equation*, IEEE Trans. Autom. Control, 31 (1986), pp. 656–658, https://doi.org/10.1109/TAC.1986.1104369. 34

[90] Oberwolfach Benchmark Collection, *Steel Profile.* hosted at MORwiki – Model Order Reduction Wiki, 2005, http://modelreduction.org/index.php/Steel_Profile (accessed 2021-02-01). 112

[91] M. R. Opmeer, *Decay of singular values of the Gramians of infinite-dimensional systems*, in 2015 European Control Conference (ECC), Linz, 2015, 2015, pp. 1183–1188, https://doi.org/10.1109/ECC.2015.7330700. 32

[92] D. Peaceman and H. Rachford, *The Numerical Solution of Parabolic and Elliptic Differential Equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28–41. 30

[93] T. Penzl, *A Cyclic Low-Rank Smith Method for Large Sparse Lyapunov Equations*, SIAM J. Sci. Comput., 21 (2000), pp. 1401–1418, https://doi.org/10.1137/S1064827598347666. 35, 112

[94] T. Penzl, *Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case*, Syst. Cont. Lett., 40 (2000), pp. 139–144, https://doi.org/10.1016/S0167-6911(00)00010-4. 2, 32, 33

[95] J. W. Prüss and M. Wilke, *Gewöhnliche Differentialgleichungen und dynamische Systeme*, Grundstudium Mathematik., Springer-Verlag, Basel, 2011, https://doi.org/10.1007/978-3-0348-0002-0. 16

[96] V. Radisavljevic, *Improved Potter–Anderson–Moore algorithm for the differential Riccati equation*, Appl. Math. Comput., 218 (2011), pp. 4641–4646, https://doi.org/10.1016/j.amc.2011.09.007. 77

[97] S. Rave and J. Saak, *Thermal Block.* MORwiki – Model Order Reduction Wiki, 2020, http://modelreduction.org/index.php/Thermal_Block (accessed 2021-02-01). 112

[98] E. Ringh, G. Mele, J. Karlsson, and E. Jarlebring, *Sylvester-based preconditioning for the waveguide eigenvalue problem*, Linear Algebra Appl., 542 (2018), pp. 441–463, https://doi.org/10.1016/j.laa.2017.06.027. 23

[99] H. Rome, *A Direct Solution to the Linear Variance Equation of a Time-Invariant Linear System*, IEEE Trans. Autom. Control, 14 (1969), pp. 592–593, https://doi.org/10.1109/TAC.1969.1099271. 43

[100] W. J. Rugh, *Linear System Theory*, Prentice Hall Information and System Sciences Series, Prentice-Hall, Englewood Cliffs, NJ, second ed., 1996. 20, 21, 57, 58

[101] I. Rusnak, *Almost Analytic Representation for the Solution of the Differential Matrix Riccati Equation*, IEEE Trans. Autom. Control, 33 (1988), pp. 191–193, https://doi.org/10.1109/9.388. 78

[102] Y. Saad, *Analysis of Some Krylov Subspace Approximations to the Matrix Exponential Operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228, https://doi.org/10.1137/0729014. 102

[103] J. Sabino, *Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method*, Dissertation, Rice University, Houston, Texas, 2007, https://hdl.handle.net/1911/20641 (accessed 2021-02-01). 32, 34, 35

[104] T. C. Sideris, *Ordinary Differential Equations and Dynamical Systems*, Atlantis Press, 2013, https://doi.org/10.2991/978-94-6239-021-8. 20

[105] V. Simoncini, *Analysis of the rational Krylov subspace projection method for large-scale algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 1655–1674, https://doi.org/10.1137/16M1059382. 1, 61

[106] V. Simoncini, D. B. Szyld, and M. Monsalve, *On two numerical methods for the solution of large-scale algebraic Riccati equations*, IMA J. Numer. Anal., 34 (2014), pp. 904–920, https://doi.org/10.1093/imanum/drt015. 61

[107] G. Söderlind, *The logarithmic norm. History and modern theory*, BIT Numer. Math., 46 (2006), pp. 631–652, https://doi.org/10.1007/s10543-006-0069-9. 12

[108] D. C. Sorensen and Y. Zhou, *Bounds on eigenvalue decay rates and sensitivity of solutions to Lyapunov equations*, Tech. Report TR02-07, Dept. of Comp. Appl. Math., Rice University, Houston, TX, June 2002, https://hdl.handle.net/1911/101987 (accessed 2021-02-01). 2, 32, 33

[109] M. Sorine and P. Winternitz, *Superposition Laws for Solutions of Differential Matrix Riccati Equations Arising in Control Theory*, IEEE Trans. Autom. Control, 30 (1985), pp. 266–272, https://doi.org/10.1109/TAC.1985.1103934. 78

[110] G. Starke, *Near-Circularity for the Rational Zolotarev Problem in the Complex Plane*, J. Approx. Theory, 70 (1992), pp. 115–130, https://doi.org/10.1016/0021-9045(92)90059-W. 33

[111] G. W. Stewart, *Matrix Algorithms. Vol. II*, SIAM Publications, Philadelphia, 2001, https://doi.org/10.1137/1.9780898718058. 25

[112] T. Stillfjord, *Low-Rank Second-Order Splitting of Large-Scale Differential Riccati Equations*, IEEE Trans. Autom. Control, 60 (2015), pp. 2791–2796, https://doi.org/10.1109/TAC.2015.2398889. 100, 101, 102, 105

[113] T. Stillfjord, *Adaptive high-order splitting schemes for large-scale differential Riccati equations*, Numer. Algorithms, 78 (2018), pp. 1129–1151, https://doi.org/10.1007/s11075-017-0416-8. 2, 100

[114] T. Stillfjord, *Singular Value Decay of Operator-Valued Differential Lyapunov and Riccati Equations*, SIAM J. Control Optim., 56 (2018), pp. 3598–3618, https://doi.org/10.1137/18M1178815. 32

[115] The MORwiki Community, *Convection.* MORwiki – Model Order Reduction Wiki, 20XX, http://modelreduction.org/index.php/Convection (accessed 2021-02-01). 112

[116] A. Townsend and H. Wilber, *On the singular values of matrices with high displacement rank*, Linear Algebra Appl., 548 (2018), pp. 19–41, https://doi.org/10.1016/j.laa.2018.02.025. 32, 33

[117] L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM Publications, Philadelphia, 1997, https://doi.org/10.1137/1.9780898719574. 9

[118] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra*, Princeton University Press, Princeton, NJ, 2005, https://doi.org/10.1515/9780691213101. 12

[119] N. Truhar, Z. Tomljanović, and R.-C. Li, *Analysis of the solution of the Sylvester equation using low-rank ADI with exact shifts*, Syst. Cont. Lett., 59 (2010), pp. 248–257, https://doi.org/10.1016/j.sysconle.2010.02.002. 30, 32

[120] N. Truhar and K. Veselić, *Bounds on the trace of a solution to the Lyapunov equation with a general stable matrix*, Syst. Cont. Lett., 56 (2007), pp. 493–503, https://doi.org/10.1016/j.sysconle.2007.02.003. 32

[121] A. Varga, *On solving periodic Riccati equations*, Numer. Lin. Alg. Appl., 15 (2008), pp. 809–835, https://doi.org/10.1002/nla.604. 108

[122] D. R. Vaughan, *A Negative Exponential Solution for the Matrix Riccati Equation*, IEEE Trans. Autom. Control, 14 (1969), pp. 72–75, https://doi.org/10.1109/tac.1969.1099117. 79

[123] E. L. Wachspress, *Iterative Solution of the Lyapunov Matrix Equation*, Appl. Math. Letters, 1 (1988), pp. 87–90, https://doi.org/10.1016/0893-9659(88)90183-8. 1, 25, 30

[124] E. L. Wachspress, *The ADI Model Problem*, Springer-Verlag, New York, 2013, https://doi.org/10.1007/978-1-4614-5122-8. 30, 35

[125] W. Walter, *Ordinary Differential Equations*, vol. 182 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1998, https://doi.org/10.1007/978-1-4612-0601-9. 11, 12, 16, 17, 75

[126] N. Wong and V. Balakrishnan, *Quadratic alternating direction implicit iteration for the fast solution of algebraic Riccati equations*, in Proc. Int. Symposium on Intelligent Signal Processing and Communication Systems, 2005, pp. 373–376, https://doi.org/10.1109/ISPACS.2005.1595424. 115

[127] N. Wong and V. Balakrishnan, *Fast positive-real balanced truncation via quadratic alternating direction implicit iteration*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 26 (2007), pp. 1725–1731, https://doi.org/10.1109/TCAD.2007.895617. 115

STATEMENT OF SCIENTIFIC COOPERATIONS

This work is based on articles and reports (published and unpublished) that have been obtained in cooperation with various coauthors. To guarantee a fair assessment of this thesis, this statement clarifies the contributions that each individual coauthor has made. The following people contributed to the content of this work:

- Peter Benner (PB), Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
- Jan Heiland (JH), Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
- Federico Poloni (FP), University of Pisa, Pisa, Italy
- Valeria Simoncini (VS), University of Bologna, Bologna, Italy

Chapter 3

Lemma 3.4 and Theorem 3.5 in Section 3.2 are based on

[15] M. Behr, P. Benner, and J. Heiland, *Solution Formulas for Differential Sylvester and Lyapunov Equations*, Calcolo, 56 (2019), p. 51, https://doi.org/10.1007/s10092-019-0348-x

and were established by myself.

Chapter 4

The theoretical results in Section 4.2 and the corresponding computational results in Appendix D.1 were obtained by myself. Section 4.2 and Appendix D.1 are based on

[15] M. Behr, P. Benner, and J. Heiland, *Solution Formulas for Differential Sylvester and Lyapunov Equations*, Calcolo, 56 (2019), p. 51, https://doi.org/10.1007/s10092-019-0348-x.

PB suggested me to do a numerical comparison with the `BDF/ADI` method for the computational results presented in [15]. Several discussions with PB and JH and their proofreads improved the presentation of [15].

Chapter 6

All theoretical results in Section 6.3.2 were obtained by myself.

FP answered my question about the numerical approximation of the matrix exponential of an Hamiltonian matrix on StackExchange [1]. FP pointed out the Riccati–Lyapunov transformation to me, which is a part of the proof of Theorem 6.15.

All theoretical results in Section 6.5 and the corresponding computational results in Appendix D.2 were obtained by myself. Section 6.5 and Appendix D.2 is a partly based and revised version of

[16] M. Behr, P. Benner, and J. Heiland, *Galerkin trial spaces and Davison–Maki methods for the numerical solution of differential Riccati equations*, Appl. Math. Comp., 410 (2021), p. 126401, https://doi.org/10.1016/j.amc.2021.126401.

Discussions with PB and JH and their proofreads improved the presentation of [16].

A discussion with VS during the "METT VIII – 8th Workshop on Matrix Equations and Tensor Techniques" [2] led to an improvement of the considerations that have been made on the choice of the step size h for the modified Davison–Maki method (Section 6.4.2 below Algorithm 6.3 and [16, Sec. 3.2]) and resulted in a simplified version of the proof of Lemma 6.16.

[1] https://scicomp.stackexchange.com/questions/29320/matrix-exponential-of-a-hamiltonian-matrix

[2] https://www.mpi-magdeburg.mpg.de/csc/events/mett2019

DECLARATION OF HONOR

I hereby declare that I produced this thesis without prohibited assistance and that all sources of information that were used in producing this thesis, including my own publications, have been clearly marked and referenced.

In particular, I have not willfully:

- Fabricated data or ignored or removed undesired results.
- Misused statistical methods with the aim of drawing other conclusions than those warranted by the available data.
- Plagiarized data or publications or presented them in a distorted way.

I know that violations of copyright may lead to injunction and damage claims from the author or prosecution by the law enforcement authorities.

This work has not previously been submitted as a doctoral thesis in the same or a similar form in Germany or in any other country. It has not previously been published as a whole.

Magdeburg, 27.10.2021

Maximilian Behr