Galerkin Trial Spaces and Davison-Maki Methods for the Numerical Solution of Differential Riccati Equations

Maximilian Behr[a,*], Peter Benner[a,b], Jan Heiland[a,b]

[a] *Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany*
[b] *Faculty of Mathematics, Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany*

Abstract

The differential Riccati equation appears in different fields of applied mathematics like control and system theory. Recently, Galerkin methods based on Krylov subspaces were developed for the autonomous differential Riccati equation. These methods overcome the prohibitively large storage requirements and computational costs of the numerical solution. Known solution formulas are reviewed and extended. Because of memory-efficient approximations, invariant subspaces for a possibly low-dimensional solution representation are identified. A Galerkin projection onto a trial space related to a low-rank approximation of the solution of the algebraic Riccati equation is proposed. The *modified Davison-Maki method* is used for time discretization. Known stability issues of the *Davison-Maki method* are discussed. Numerical experiments for large-scale autonomous differential Riccati equations and a comparison with high-order splitting schemes are presented.

Keywords: differential Riccati equation, Davison-Maki method, low-rank methods, optimal control, linear-quadratic regulator

2020 MSC: 15A24, 65F60, 65L05, 93B52

1. Introduction

In this paper, we consider the autonomous differential Riccati equation

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

The most prominent application of the differential Riccati equations is the linear-quadratic regulator problem both in finite (cp., e.g., [1–3, 26]) and infinite dimensions [4]. More recently, the differential Riccati equation has been used to analyze steady state solutions to reaction-diffusion equations [5, 6].

*Corresponding Author
Email addresses: behr@mpi-magdeburg.mpg.de (Maximilian Behr), benner@mpi-magdeburg.mpg.de (Peter Benner), heiland@mpi-magdeburg.mpg.de (Jan Heiland)

We focus on the large-scale case that occurs, e.g., when infinite dimensional problems are spatially discretized. In such settings, the numerical approximation of X comes with high memory requirements and high computational costs. Just the storage of the solution at the relevant time instances would scale with $N_t n^2$, if n is the dimension of the problem and N_t is the number of time steps.

Most approaches discretize in time and then focus on an efficient approximation of the resulting algebraic equations. This typically comes with a restriction on the choice of the time discretization method in order to preserve the definiteness of the discrete solution; cp. [7–10], although there have been efforts to overcome this restriction; cp. [11, 14]. Nonetheless, in these methods, the use of higher order schemes implies additional effort in every solve of the algebraic Riccati equation, so that only *backward differencing schemes* are considered suitable choices for the time discretization. More flexibility is provided by the *splitting schemes* (see, e.g., [13, 15, 16]) that separate the linear and nonlinear parts. Still, at least one large-scale algebraic equation has to be solved and stored for every time step in all the approaches mentioned in this paragraph. Conceptually, it seems more beneficial for the autonomous differential Riccati equation to first reduce the problem dimensions to, say, $k \ll n$, and then approach the reduced equation as this leads to storage requirements of the order of $N_t k^2$ for the reduced problem and nk for the basis vectors. In this respect, Krylov subspace methods have been proposed [17–23] that generate a trial space for the numerical solution using an Arnoldi method. The resulting Galerkin projected system is of lower order and can be solved with low memory demands and with various methods that exist for differential Riccati equations of small or moderate size.

In this work, we develop a Galerkin approach, where the trial space is based on the numerical solution of the algebraic Riccati equation. This extends the concepts of our previous work on a numerical scheme for the differential Lyapunov equation [24].

The paper is organized as follows. In Section 2, we introduce the algebraic and differential Riccati equations and review the relevant fundamental properties of their solutions. In Section 3, we review *Radon's Lemma*. Moreover, in Section 3.2, we apply *Radon's Lemma* to obtain solution formulas for the differential Riccati equation based on the solution of the algebraic Riccati equation that we will use to explain and illustrate the major source of numerical instabilities of the *Davison-Maki method* for the numerical solution of the differential Riccati equation; see Section 3.3. We derive the *modified Davison-Maki method* in a way that overcomes these instabilities. In Section 4, we develop a Galerkin approach for the solution of the differential Riccati equation in the matrix exponential representation that results from *Radon's Lemma*. We combine the monotonicity of the solution of the differential Riccati equation and relevant properties of the solution of the algebraic Riccati equation to define a suitable and numerically computable trial space for the approximation of the solution of the differential Riccati equation. We propose to solve the resulting Galerkin system with the *modified Davison-Maki method*. Numerical results are presented in Section 5, Appendix A, and Appendix B.

We set the notation and review some basic results from linear algebra. The identity matrix and zero matrix

of size $n \times n$ are written by I_n and 0_n, respectively. The image or column space of a matrix $A \in \mathbb{R}^{n\times m}$ is denoted by range (A), and its kernel or null space by $\ker(A)$. The 1–norm, 2–norm, Frobenius norm, and Frobenius inner product are denoted by $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_F$, and $\langle\cdot,\cdot\rangle_F$, respectively. The spectrum of a square matrix A is denoted by $\Lambda(A)$. Generally, the spectrum is a subset of $\mathbb{C}$. A matrix is called *stable* if its spectrum is contained in the left open complex half plane $\mathbb{C}^-$, i.e., $\Lambda(A) \subseteq \mathbb{C}^-$. If A is real and symmetric, all eigenvalues are real, and $\lambda_k^{\downarrow}(A)$ represents the k–largest eigenvalue. Therefore, $\lambda_1^{\downarrow}(A) \geq \lambda_2^{\downarrow}(A) \geq \cdots \geq \lambda_n^{\downarrow}(A)$ are the eigenvalues of A ordered in a weakly decreasing fashion. The Loewner partial ordering on the set of real symmetric matrices is defined by $A \preccurlyeq B$, which means $B - A$ is positive semidefinite, [25, Ch. 7.7]. The unique symmetric positive semidefinite square root of a symmetric positive semidefinite matrix $X \in \mathbb{R}^{n\times n}$ is denoted by $X^{1/2}$; cf. [25, Thm. 7.2.6]. The orthogonal complement of a linear subspace $U \subseteq \mathbb{R}^n$ is denoted by $U^{\perp} \subseteq \mathbb{R}^n$. For $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times b}$, the image of the Krylov matrix generated by A and B is denoted by $\mathcal{K}(A,B) := \text{range}\left(\left[B, AB, \ldots, A^{n-1}B\right]\right) \subseteq \mathbb{R}^n$. The linear space $\mathcal{K}(A,B)$ is A–invariant.

2. Algebraic and Differential Riccati Equations

This section introduces the algebraic and differential Riccati equation (ARE/DRE) and the algebraic Lyapunov equation (ALE).

Consider $A, X_0 \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times b}$ and $C \in \mathbb{R}^{c\times n}$. Throughout this paper, we assume that X_0 is a symmetric positive semidefinite matrix and consider the DRE

$$\dot{X}(t) = \mathcal{R}(X(t)) := A^\top X(t) + X(t)A - X(t)BB^\top X(t) + C^\top C, \tag{1a}$$

$$X(0) = X_0. \tag{1b}$$

Stationary points of (1a) are solutions of the corresponding ARE

$$0_n = \mathcal{R}(X) = A^\top X + XA - XBB^\top X + C^\top C. \tag{2}$$

The linear version of the ARE is the ALE

$$0_n = A^\top X + XA + C^\top C. \tag{3}$$

We review some fundamental results about existence, uniqueness, and properties of the solution of the DRE (1), ARE (2), and the ALE (3).

Theorem 2.1 ([26, Thm. 1.1.3, Thm. 1.1.7])**.**
If $\Lambda(A) \cap \Lambda(-A) = \emptyset$, then the ALE (3) *has a unique solution $X_L \in \mathbb{R}^{n\times n}$. The solution X_L is symmetric.*

If A is stable, then X_L is symmetric positive semidefinite and given by

$$X_L = \int_0^\infty e^{tA^\top} C^\top C e^{tA} \mathrm{d}t.$$

Theorem 2.2 ([26, Lem. 2.4.1, Cor. 2.4.3], [2, Ch. 10])**.**
Let (A, B) be stabilizable and (A, C) be detectable, then the ARE (2) *has a unique stabilizing solution $X_\infty \in \mathbb{R}^{n\times n}$. This means $\mathcal{R}(X_\infty) = 0_n$, and $A - BB^\top X_\infty$ is stable. Moreover, X_∞ is symmetric positive semidefinite, and there is no other symmetric positive semidefinite solution of the* ARE (2)*.*

Theorem 2.3 ([27, Thm. 3.2])**.**
Let (A, B) be stabilizable, (A, C) be detectable, and $X_\infty \in \mathbb{R}^{n\times n}$ be the unique stabilizing solution of the ARE (2)*. Then the following relation holds:*

$$\operatorname{range}(X_\infty) = \mathcal{K}\left(A^\top, C^\top\right).$$

The inclusion $\mathcal{K}\left(A^\top, C^\top\right) \subseteq \operatorname{range}(X_\infty)$ in Theorem 2.3 holds for each symmetric solution of the ARE (2); cf. [26, Lemma 2.4.9]. In [28, Sec. 3.3], a Kalman decomposition is used to show that $\operatorname{rank}(X_\infty) = \dim\left(\mathcal{K}\left(A^\top, C^\top\right)\right)$. A connection between the space $\mathcal{K}\left(A^\top, C^\top\right)$ and a certain Krylov subspace generated by the associated Hamiltonian matrix, which can be used for the numerical approximation of the solution of the ARE (2), was presented in [29, Thm. 10].
Typically, solutions of quadratic differential equations like the DRE (1) exhibit a finite-time escape phenomenon. Through comparison arguments and the fact that $-BB^\top$ is negative semidefinite, one can show that the solution exists for all $t \geq 0$. With additional assumptions, the solution converges monotonically to the unique solution of the ARE (2) and is, thus, bounded.

Theorem 2.4 ([26, Thm. 4.1.6, Thm. 4.1.8], [2, Ch. 10])**.**
The DRE (1) *has a unique solution $X\colon (t^-, \infty) \to \mathbb{R}^{n\times n}$. The solution X has the following properties:*

- *$X(t)$ is symmetric for all $t \in (t^-, \infty)$.*
- *$X(t)$ is symmetric positive semidefinite for all $t \geq 0$.*
- *If $0_n \preccurlyeq \dot{X}(0) = \mathcal{R}(X_0)$, then $t \mapsto X(t)$ is monotonically increasing on $[0, \infty)$, i.e. $X(t_1) \preccurlyeq X(t_2)$ for all t_1, t_2 such that $0 \leq t_1 \leq t_2$.*

Theorem 2.5 ([27, Thm. 3.1])**.**
Let the columns of $Q \in \mathbb{R}^{n\times p}$ be an orthonormal basis of $\mathcal{K}\left(A^\top, C^\top\right)$ and define the linear space $\mathcal{Q} :=$

$\{QYQ^\top \mid Y \in \mathbb{R}^{p\times p}\} \subseteq \mathbb{R}^{n\times n}$ *or* $\mathcal{Q} := \{0_n\} \subseteq \mathbb{R}^{n\times n}$*, if* C *is zero. Then the following holds:*

$$X(t) \in \mathcal{Q} \text{ for all } t \geq 0,$$

where X *is the unique solution of the* DRE (1) *with* $X_0 = 0_n$*.*

With this relation, one can readily confirm that the solution of the DRE (1) evolves in an invariant subspace of $\mathbb{R}^{n\times n}$.

For numerical approximations of the solutions of large-scale ALEs, AREs, and DREs, one typically seeks low-rank approximations to avoid overly demanding memory requirements. Therefore, the relevant literature features numerous contributions that study the decay rate of $\lambda_k^{\downarrow}(X)$ or $\lambda_k^{\downarrow}(X)/\lambda_1^{\downarrow}(X)$ for increasing k; see, e.g., [30–37] on the eigenvalue decay of the solution of the ALE and [15, 29, 35] for results on the ARE and DRE.

For the DRE (1), one can derive estimates based on the monotonicity. Assume that $0_n \preccurlyeq \mathcal{R}(X_0)$, then by Theorem 2.4, the function $t \mapsto X(t)$ is monotonically weakly increasing on $[0, \infty)$, where X is the unique solution of the DRE (1). A direct consequence of the Courant-Fischer-Weyl min-max principle [25, Cor. 7.7.4] implies that $t \mapsto \lambda_k^{\downarrow}(X(t))$ is also monotonically weakly increasing on $[0, \infty)$. Therefore, the number of eigenvalues of $X(t)$ greater than or equal to a given threshold $\varepsilon > 0$ is weakly increasing over time.

Example 2.1 (Eigenvalue Decay)**.**
We illustrate this by an example in Figure 1*. We have chosen* $C = \begin{bmatrix}1, \ldots, 1\end{bmatrix} = B^\top$*,* $X_0 = 0_n$ *and* A *to be tridiagonal with entries* $5, -1, -5$ *on the subdiagonal, diagonal, and superdiagonal, respectively. The matrices are of size* $n = 100$*, and the* DRE *was solved numerically to high precision on the time interval* $[0, 15]$*. For this, we have used the variable-precision arithmetic* `vpa` *of* MATLAB® 2018a *with* 512 *significant digits and* Algorithm 2 *with step size* $h = 2^{-5}$*. The eigenvalues of* $X(t)$ *are arranged in a weakly decreasing order and plotted for* $t \in \{0.5, 1, \ldots, 15\}$*. The functions* $t \mapsto \lambda_k^{\downarrow}(X(t))$ *are highlighted in red for* $k \in \{10, 20, 30, 40, 50\}$*. All eigenvalues below* 10^{-60} *were truncated from* Figure 1*. The shadowed red plane is drawn at the level* $2 \cdot 10^{-16}$*, which is approximately machine precision in double arithmetic.*

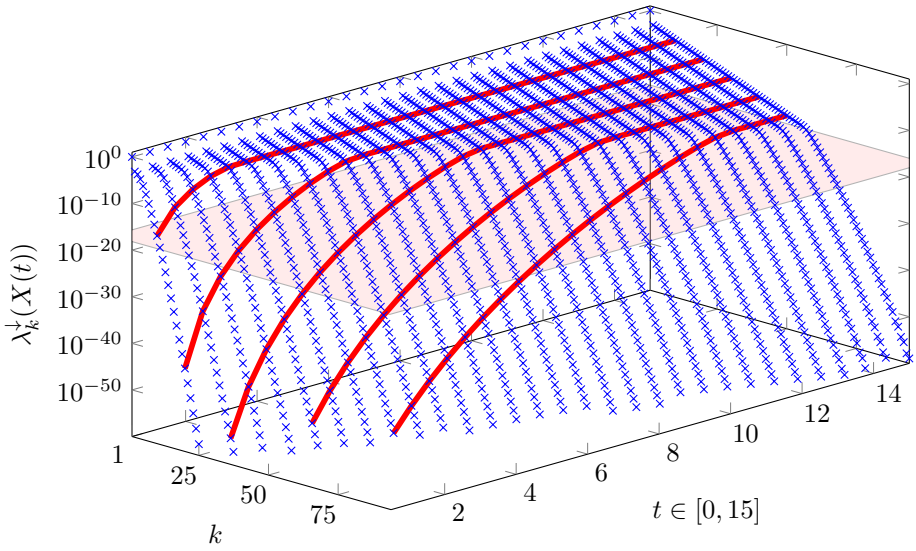


Fig. 1. Eigenvalues $\lambda_k^{\downarrow}(X(t))$ of the numerical solution of DRE (1).

3. Solution Formulas and Davison-Maki methods

3.1. Radon's Lemma

In this section, we consider the nonsymmetric differential Riccati equation abbreviated by NDRE as a generalization of the DRE. We will make heavy use of *Radon's Lemma* that shows that the NDRE is locally equivalent to a linear differential equation of twice the size. Vice versa, the solution of the NDRE defines the solution of an associated linear system. *Radon's Lemma* (Theorem 3.1) has several consequences. Section 3.2 shows how solution formulas can be obtained by applying suitable linear transformations, which decouple the linear differential equation.

Then, because of numerical approximations, we review the *Davison-Maki method* and the *modified Davison-Maki method* in Section 3.3. We use the solution formula from Section 3.2 to explain why the *Davison-Maki method* applied to the DRE usually suffers from numerical instabilities.

We use the solution representation (Theorem 3.3) to motivate the Galerkin approach described in Section 4.1. We make use of the *modified Davison-Maki method* in Algorithm 3 in Section 4.

Theorem 3.1 (Radon's Lemma, [26, Thm. 3.1.1])**.**
Let $M_{11} \in \mathbb{R}^{n\times n}$, $M_{12} \in \mathbb{R}^{n\times m}$, M_{21}, $M_0 \in \mathbb{R}^{m\times n}$, $M_{22} \in \mathbb{R}^{m\times m}$, *and* $\mathbb{I} \subseteq \mathbb{R}$ *be an open interval such that* $0 \in \mathbb{I}$. *We consider the* NDRE

$$\dot{W}(t) = M_{22}W(t) - W(t)M_{11} - W(t)M_{12}W(t) + M_{21}, \tag{4a}$$

$$W(0) = M_0. \tag{4b}$$

The following holds:

1. *Let $W\colon \mathbb{I} \to \mathbb{R}^{m\times n}$ be the solution of* (4) *and $U\colon \mathbb{I} \to \mathbb{R}^{n\times n}$ be the solution of the linear initial value problem*

$$\dot{U}(t) = (M_{11} + M_{12}W(t))\,U(t),\ U(0) = I_n. \tag{5}$$

Moreover, let $V(t) := W(t)U(t)$. 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

$$\begin{bmatrix}\dot{U}(t)\\ \dot{V}(t)\end{bmatrix} = M\begin{bmatrix}U(t)\\ V(t)\end{bmatrix} := \begin{bmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{bmatrix}\begin{bmatrix}U(t)\\ V(t)\end{bmatrix},\ \begin{bmatrix}U(0)\\ V(0)\end{bmatrix} = \begin{bmatrix}I_n\\ M_0\end{bmatrix}. \tag{6}$$

2. *If $\begin{bmatrix}U\\ V\end{bmatrix}\colon \mathbb{I} \to \mathbb{R}^{(n+m)\times n}$ is a solution of* (6)*, and the matrix $U(t)$ is nonsingular for all $t \in \mathbb{I}$, then $W\colon \mathbb{I} \to \mathbb{R}^{m\times n},\ W(t) = V(t)U(t)^{-1}$ is a solution of* (4)*.*

Radon's Lemma (Theorem 3.1) also holds for time-dependent continuous matrix-valued functions as coefficients. Note that, usually, the solution of the NDRE (4) has finite time escape, while the solution of the system (6) exists for all $t \in \mathbb{R}$. As the function U is a solution of the linear initial value problem (5) and $U(0) = I_n$ 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 $t \in \mathbb{I}$, c.f. [38, §15]. Therefore, as long as the solution of the NDRE (4) exists, it can be recovered from the solution of the system (6).

3.2. Solution Formulas

Radon's Lemma (Theorem 3.1) enables certain solution representations for the DRE (1): Theorem 2.4 ensures that the DRE (1) has a unique solution defined on the interval (t^-, ∞). By *Radon's Lemma* (Theorem 3.1), we have that $U(t)$ is nonsingular on the same interval.

Let $H := \begin{bmatrix}A & -BB^\top\\ -C^\top C & -A^\top\end{bmatrix} \in \mathbb{R}^{2n\times 2n}$ be the Hamiltonian matrix corresponding to the DRE (1). The matrices $U(t)$ and $V(t)$ are determined by the linear initial value problem

$$\begin{bmatrix}\dot{U}(t)\\ \dot{V}(t)\end{bmatrix} = -H\begin{bmatrix}U(t)\\ V(t)\end{bmatrix},\ \begin{bmatrix}U(0)\\ V(0)\end{bmatrix} = \begin{bmatrix}I_n\\ X_0\end{bmatrix}. \tag{7}$$

We obtain

$$\begin{bmatrix}U(t)\\ V(t)\end{bmatrix} = e^{-tH}\begin{bmatrix}I_n\\ X_0\end{bmatrix}.$$

The strategy is to decompose the Hamiltonian matrix H, such that (7) decouples.

Theorem 3.2 (Solution Representation I for DRE (1), [39])**.**
Let $X_1 \in \mathbb{R}^{n\times n}$ be any solution of the ARE (2)*. Then the solution of the* DRE (1) *is given by*

$$X(t) = X_1 - e^{t(A-BB^\top X_1^\top)^\top}\tilde{X}\left(I_n - \int_0^t e^{s(A-BB^\top X_1)}BB^\top e^{s(A-BB^\top X_1^\top)^\top}\mathrm{d}s\tilde{X}\right)^{-1} e^{t(A-BB^\top X_1)},$$
$$\tilde{X} := X_1 - X_0.$$

125 The formula was presented in [39] without proof. Since the existence of the involved inverse is not trivially established, we provide a proof.

Proof. We use $T := \begin{bmatrix} I_n & 0_n \\ X_1 & I_n \end{bmatrix}$ and apply a similarity transformation to H,

$$T^{-1}HT = \begin{bmatrix} I_n & 0_n \\ -X_1 & I_n \end{bmatrix}\begin{bmatrix} A & -BB^\top \\ -C^\top C & -A^\top \end{bmatrix}\begin{bmatrix} I_n & 0_n \\ X_1 & I_n \end{bmatrix} = \begin{bmatrix} A-BB^\top X_1 & -BB^\top \\ 0_n & -(A-BB^\top X_1^\top)^\top \end{bmatrix} =: \tilde{H}.$$

This gives

$$\begin{bmatrix} U(t) \\ V(t) \end{bmatrix} = e^{-tH}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} = e^{-tT\tilde{H}T^{-1}}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} = Te^{-t\tilde{H}}T^{-1}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} = Te^{-t\tilde{H}}\begin{bmatrix} I_n \\ X_0 - X_1 \end{bmatrix} =: T\begin{bmatrix} \tilde{U}(t) \\ \tilde{V}(t) \end{bmatrix}.$$

Clearly, $\tilde{U}$ and $\tilde{V}$ are determined by the solution of the initial value problem

$$\begin{bmatrix} \dot{\tilde{U}}(t) \\ \dot{\tilde{V}}(t) \end{bmatrix} = -\tilde{H}\begin{bmatrix} \tilde{U}(t) \\ \tilde{V}(t) \end{bmatrix} = \begin{bmatrix} -(A-BB^\top X_1) & BB^\top \\ 0_n & (A-BB^\top X_1^\top)^\top \end{bmatrix}\begin{bmatrix} \tilde{U}(t) \\ \tilde{V}(t) \end{bmatrix}, \quad \begin{bmatrix} \tilde{U}(0) \\ \tilde{V}(0) \end{bmatrix} = \begin{bmatrix} I_n \\ X_0 - X_1 \end{bmatrix}.$$

By using the variation of constants formula [38, §18], we obtain that $\tilde{U}$ and $\tilde{V}$ are given by

$$\begin{aligned}
\tilde{V}(t) &= -e^{t(A-BB^\top X_1^\top)^\top}(X_1 - X_0), \\
\tilde{U}(t) &= e^{-t(A-BB^\top X_1)} + \int_0^t e^{-(t-s)(A-BB^\top X_1)}BB^\top\tilde{V}(s)\mathrm{d}s \\
&= e^{-t(A-BB^\top X_1)}\left(I_n - \int_0^t e^{s(A-BB^\top X_1)}BB^\top e^{s(A-BB^\top X_1^\top)^\top}\mathrm{d}s(X_1 - X_0)\right).
\end{aligned}$$

Since $\tilde{U}(t) = U(t)$ is nonsingular for all $t \in (t^-, \infty)$ and the matrix exponential is nonsingular, the matrix in

brackets is also nonsingular. Finally, we obtain

$$\begin{aligned} V(t) &= X_1\tilde{U}(t) + \tilde{V}(t), \\ X(t) &= V(t)U(t)^{-1} = X_1 + \tilde{V}(t)\tilde{U}(t)^{-1}. \end{aligned}$$

□

Theorem 3.3 (Solution Representation II for DRE (1), [40, Thm. 1], [41])**.**
Let (A, B) be stabilizable, (A, C) be detectable, and $X_\infty \in \mathbb{R}^{n\times n}$ be the unique symmetric positive definite stabilizing solution of the ARE (2)*. Moreover, let $\hat{A} := A - BB^\top X_\infty$ and $X_L \in \mathbb{R}^{n\times n}$ be the unique symmetric positive semidefinite solution of the* ALE

$$\hat{A}X_L + X_L\hat{A}^\top + BB^\top = 0_n.$$

The solution of the DRE (1) *is represented by*

$$X(t) = X_\infty - e^{t\hat{A}^\top}(X_\infty - X_0)\Big(I_n - \Big(X_L - e^{t\hat{A}}X_L e^{t\hat{A}^\top}\Big)(X_\infty - X_0)\Big)^{-1}e^{t\hat{A}}.$$

Proof. Similar to the proof of Theorem 3.2, we use similarity transformations to decompose the Hamiltonian matrix H. This is also known as a Riccati-Lyapunov transformation [26, Ch. 3.1.1]. We obtain

$$\begin{aligned} T &:= \begin{bmatrix} I_n & 0_n \\ X_\infty & I_n \end{bmatrix},\ T^{-1}HT = \begin{bmatrix} \hat{A} & -BB^\top \\ 0_n & -\hat{A}^\top \end{bmatrix} =: \tilde{H}, \\ \tilde{T} &:= \begin{bmatrix} I_n & -X_L \\ 0_n & I_n \end{bmatrix},\ \tilde{T}^{-1}\tilde{H}\tilde{T} = \begin{bmatrix} \hat{A} & 0_n \\ 0_n & -\hat{A}^\top \end{bmatrix} =: \hat{H}. \end{aligned}$$

We thus get

$$\begin{aligned} e^{-tH} &= e^{-t(T\tilde{T})\hat{H}(T\tilde{T})^{-1}} = (T\tilde{T})e^{-t\hat{H}}(T\tilde{T})^{-1} \\ &= \begin{bmatrix} I_n & -X_L \\ X_\infty & I_n - X_\infty X_L \end{bmatrix} \begin{bmatrix} e^{-t\hat{A}} & 0_n \\ 0_n & e^{t\hat{A}^\top} \end{bmatrix} \begin{bmatrix} I_n - X_L X_\infty & X_L \\ -X_\infty & I_n \end{bmatrix} \\ &= \begin{bmatrix} e^{-t\hat{A}}(I_n - X_L X_\infty) + X_L e^{t\hat{A}}X_\infty & e^{-t\hat{A}}X_L - X_L e^{t\hat{A}^\top} \\ X_\infty e^{-t\hat{A}}(I_n - X_L X_\infty) - (I_n - X_\infty X_L)e^{t\hat{A}}X_\infty & X_\infty e^{-t\hat{A}}X_L + (I_n - X_\infty X_L)e^{t\hat{A}} \end{bmatrix} \end{aligned} \tag{8}$$

and

$$\begin{bmatrix} U(t) \\ V(t) \end{bmatrix} = e^{-tH} \begin{bmatrix} I_n \\ X_0 \end{bmatrix} = \begin{bmatrix} e^{-t\hat{A}}(I_n - X_L(X_\infty - X_0)) + X_L e^{t\hat{A}^\top}(X_\infty - X_0) \\ X_\infty e^{-t\hat{A}}(I_n - X_L(X_\infty - X_0)) - (X_\infty X_L + I_n)e^{t\hat{A}^\top}(X_\infty - X_0) \end{bmatrix}. \tag{9}$$

Now observe that

$$\begin{aligned} U(t) &= e^{-t\hat{A}} \left(I_n - \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top} \right) (X_\infty - X_0) \right), \\ V(t) &= X_\infty e^{-t\hat{A}} \left(I_n - \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top} \right) (X_\infty - X_0) \right) - e^{t\hat{A}^\top}(X_\infty - X_0) \\ &= X_\infty U(t) - e^{t\hat{A}^\top}(X_\infty - X_0), \end{aligned}$$

therefore,

$$X(t) = V(t)U(t)^{-1} = X_\infty - e^{t\hat{A}^\top}(X_\infty - X_0)\Big(I_n - \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top} \right) (X_\infty - X_0) \Big)^{-1} e^{t\hat{A}}.$$

□

In [42, Ch. 15.4], one can find another solution formula, which holds under more restrictive assumptions. A solution formula based on the Jordan canonical form is given in [26, Thm. 3.2.1].

3.3. Davison-Maki Methods

The *Davison-Maki method* for the NDRE (4) was proposed in [43]. 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 3.1), we have that

$$\begin{bmatrix} U(h) \\ V(h) \end{bmatrix} = e^{hM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix}, \quad W(h) = V(h)U(h)^{-1}.$$

The next step is then to make use of the semigroup property of the matrix exponential:

$$\begin{bmatrix} U(2h) \\ V(2h) \end{bmatrix} = e^{2hM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix} = \left(e^{hM}\right)^2 \begin{bmatrix} I_n \\ M_0 \end{bmatrix}, \quad W(2h) = V(2h)U(2h)^{-1}.$$

For the further steps, we obtain

$$\begin{bmatrix} U(kh) \\ V(kh) \end{bmatrix} = \left(e^{hM}\right)^k \begin{bmatrix} I_n \\ M_0 \end{bmatrix}, \quad W(kh) = V(kh)U(kh)^{-1}.$$

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 Algorithm 1.

Algorithm 1 Davison-Maki method for the NDRE (4) [43, 44]

Assumption: The solution W of the NDRE (4) exists on $[0, t_f)$.

Input: Real matrices M_0 and M_{ij} as in Theorem 3.1, step size $h > 0$ and final time $t_f > 0$.

Output: Matrices W_k, such that $W(kh) = W_k$ for $k \in \mathbb{N}_0$ and $kh < t_f$.

$W_0 = M_0$;

$k = 1$;

% Compute matrix exponential:

$\Theta_h = \exp\left(h \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}\right)$;

`Variant with matrix exponential update:`

$\Theta = \Theta_h$;

while $kh < t_f$ **do**

Partition $\begin{matrix} \\ n \\ m \end{matrix} \begin{matrix} \begin{matrix} n & m \end{matrix} \\ \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12} & \Theta_{22} \end{bmatrix} \end{matrix} = \Theta$;

$U_{\mathtt{dm}} = \Theta_{11} + \Theta_{12} M_0$;

$V_{\mathtt{dm}} = \Theta_{21} + \Theta_{22} M_0$;

$W_k = V_{\mathtt{dm}} {U_{\mathtt{dm}}}^{-1}$;

$\Theta = \Theta\Theta_h$;

$k = k + 1$;

end while

`Variant with updating` U `and` V`:`

$U_{\mathtt{dm}} = I_n$;

$V_{\mathtt{dm}} = M_0$;

Partition $\begin{matrix} \\ n \\ m \end{matrix} \begin{matrix} \begin{matrix} n & m \end{matrix} \\ \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12} & \Theta_{22} \end{bmatrix} \end{matrix} = \Theta$;

while $kh < t_f$ **do**

$U_{\mathtt{dm}} = \Theta_{11} U_{\mathtt{dm}} + \Theta_{12} V_{\mathtt{dm}}$;

$V_{\mathtt{dm}} = \Theta_{21} U_{\mathtt{dm}} + \Theta_{22} V_{\mathtt{dm}}$;

$W_k = V_{\mathtt{dm}} U_{\mathtt{dm}}^{-1}$;

$k = k + 1$;

end while

When the *Davison-Maki method* (Algorithm 1) is applied to the DRE (1), usually numerical instabilities occur because each block of e^{-tH} as well as $U(t)$ and $V(t)$ contains the matrix $e^{-t\hat{A}}$; cf. Equations (8) and (9). Since $\hat{A} = A - BB^\top X_\infty$ 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* (Algorithm 1) was also pointed out in [44–47]. Another reason is that the spectrum of a real Hamiltonian matrix comes in quadruples, that is $\Lambda(H) = \{\lambda_1, \dots, \lambda_n, -\lambda_1, \dots, -\lambda_n\}$ with $\operatorname{Re}(\lambda_i) \leq 0$. Therefore, the spectrum of the Hamiltonian matrix usually contains eigenvalues with positive real part, and its matrix exponential grows for large times [48, Prop. 2.3.1].

A suitable modification of the *Davison-Maki method* (Algorithm 1) was proposed in [44], but the modified method originates back to [49, p. 9].

By *Radon's Lemma* (Theorem 3.1), we have the identity

$$\begin{bmatrix} \tilde{U} \\ \tilde{V} \end{bmatrix} := e^{-hH} \begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix}, \quad W(kh) = \tilde{V}\tilde{U}^{-1}.$$

The *modified Davison-Maki method* is given in Algorithm 2.

Algorithm 2 Modified Davison-Maki method for the NDRE (4) [44, 49]

Assumption: The NDRE (4) has a solution $W\colon [0, t_f) \to \mathbb{R}^{m\times n}$.

Input: Real matrices M_0 and M_{ij} as in Theorem 3.1, step size $h > 0$, final time $t_f > 0$ and a moderate large number $tol_{\texttt{exp}} > 0$.

Output: Matrices W_k, such that $W(kh) = W_k$ for $k \in \mathbb{N}_0$ and $kh < t_f$.

$W_0 = M_0$;

$k = 1$;

% 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 $\|\Theta\|_1 > tol_{\texttt{exp}}$ **then**

    return Error(„1-Norm of the matrix exponential is too large, decrease the step size h“);

end if

Partition $\begin{matrix} \\ n \\ m \end{matrix} \begin{matrix} \begin{matrix} n & m \end{matrix} \\ \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12} & \Theta_{22} \end{bmatrix} \end{matrix} = \Theta$;

while $kh < t_f$ **do**

    $U_{\texttt{mod\_dm}} = \Theta_{11} + \Theta_{12} W_{k-1}$;

    $V_{\texttt{mod\_dm}} = \Theta_{21} + \Theta_{22} W_{k-1}$;

    $W_k = V_{\texttt{mod\_dm}} U_{\texttt{mod\_dm}}^{-1}$;

    $k = k + 1$;

end while

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 norm, which will lead 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. However, if one can afford a single computation of the matrix exponential and the occasional application with high accuracy for a larger step size, the solution can be corrected at the corresponding grid points.

In the k-th iteration of Algorithm 2, we have

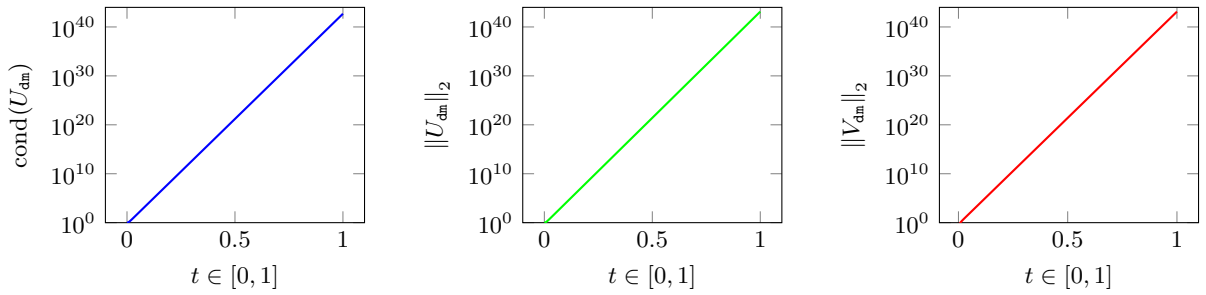
$$\begin{bmatrix} U_{\texttt{mod\_dm}} \\ V_{\texttt{mod\_dm}} \end{bmatrix} = e^{hM} \begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix} = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} \begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix},$$

and the norm of the iterates can be bounded by

$$\begin{aligned} \|U_{\texttt{mod\_dm}}\| &\leq \|\Theta_{11}\| + \|\Theta_{12}\| \, \|W((k-1)h)\| \,, \\ \|V_{\texttt{mod\_dm}}\| &\leq \|\Theta_{21}\| + \|\Theta_{22}\| \, \|W((k-1)h)\| \,. \end{aligned}$$

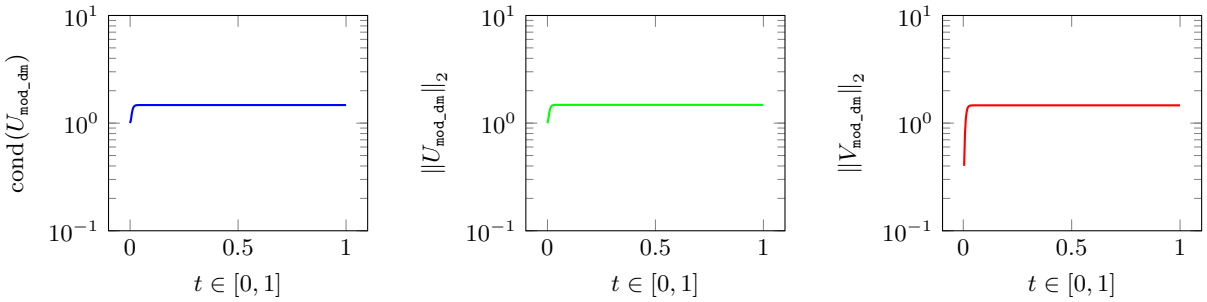
For small step sizes of $h > 0$ it holds $e^{hM} \approx I_{n+m} + hM$ and $\Theta_{11} \approx I_n + hM_{11}$, $\Theta_{12} \approx hM_{12}$, $\Theta_{21} \approx hM_{21}$ and $\Theta_{22} \approx I_m + hM_{22}$. Therefore, for small enough step size and moderate norm of the solution $\|W(t)\|$, the norm of the iterates cannot grow heavily in contrast to Algorithm 1. Assuming that the matrix exponential in line 3 of Algorithm 2 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 3.1 (Exponential Growth Davison-Maki method)**.**
We applied the Davison-Maki method (Algorithm 1) *with step size* $h = 2^{-8}$ *to a* DRE *with the same matrices* A, B, C, *and* X_0 *as in* Example 2.1. *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* $[0, 1]$. *The plot 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 3 *shows the same quantities for the iterates* $U_{\texttt{mod\_dm}}$ *and* $V_{\texttt{mod\_dm}}$ *of the modified Davison-Maki method* (Algorithm 2).



(a) Condition number of $U_{\texttt{dm}}$ (b) 2-norm of $U_{\texttt{dm}}$ (c) 2-norm of $V_{\texttt{dm}}$

Fig. 2. Davison-Maki method (Algorithm 1) and the growth of $U_{\texttt{dm}}$ and $V_{\texttt{dm}}$.



(a) Condition number of $U_{\mathtt{mod\_dm}}$ (b) 2-norm of $U_{\mathtt{mod\_dm}}$ (c) 2-norm of $V_{\mathtt{mod\_dm}}$

Fig. 3. Modified Davison-Maki method (Algorithm 2) and the growth of $U_{\mathtt{mod\_dm}}$ and $V_{\mathtt{mod\_dm}}$.

If a symmetric solution is expected, then line 11 in Algorithm 2 should be altered to $W_k = \frac{1}{2}\left(W_k + W_k^\top\right)$ because due to numerical errors, the symmetry will be lost after some iterations.

Any computationally efficient norm can also be used for the matrix exponential in Algorithm 2 line 4. The *modified Davison-Maki method* is also more efficient than the *Davison-Maki method* in both variants because fewer matrix-matrix products are needed per time step; cf. Algorithm 2 lines 8-13 with Algorithm 1 lines 5-12 and lines 16-21.

4. Galerkin Approach for Large-Scale Differential Riccati Equations

In this section, we develop a feasible numerical approach for large-scale DREs. We consider the DRE (1) and assume that $X_0 = 0_n$. We develop the Galerkin approach based on two theoretical considerations. First, we use the solution formula of Theorem 3.3. We show that the range of the solution X_∞ of the ARE is invariant under the action of the closed-loop matrix $A - BB^\top X_\infty$. It follows then that the action of the matrix exponential of the closed-loop matrix on X_∞ has the same property. This makes the approach consistent in the sense that the evolution does not leave the space and provides reasoning that the consistency error made by a numerical approximation to these subspaces can be made arbitrarily small. Moreover, this invariance property allows for a straight-forward low-dimensional approximation of the matrix exponential. After that, we show that, for our proposed choice of a Galerkin basis, the approximation quality of the space can be quantified by the eigenvalue decay of the solution of the ARE.

The result is a low-dimensional solution space with an accessible formula for the relevant matrix exponential so that we can use the *modified Davison-Maki* (Algorithm 2) for an efficient solution of the projected Galerkin system.

4.1. Invariant Subspaces for the Galerkin Approach

First, we prove that the range space of the solution X_∞ of the ARE is invariant under the action of the transposed closed-loop matrix $\left(A - BB^\top X_\infty\right)^\top$.

190 **Lemma 4.1.**

Let (A, B) be stabilizable, (A, C) be detectable, and $X_\infty \in \mathbb{R}^{n\times n}$ be the unique stabilizing solution of the ARE (2). *Then* range(X_∞) *is $(A - BB^\top X_\infty)^\top$–invariant.*

Proof. We can assume that $X_\infty \neq 0_n$. Let the columns of $Q_\infty \in \mathbb{R}^{n\times p}$ be an orthonormal basis for range(X_∞). Then $Q_\infty Q_\infty^\top$ is the orthogonal projection onto range(X_∞). We obtain

$$Q_\infty Q_\infty^\top X_\infty = X_\infty.$$

By Theorem 2.3, the columns of Q_∞ are also an orthonormal basis for $\mathcal{K}\left(A^\top, C^\top\right)$. The space $\mathcal{K}\left(A^\top, C^\top\right)$ is $A^\top$–invariant. We obtain

$$A^\top Q_\infty = Q_\infty Q_\infty^\top A^\top Q_\infty.$$

Finally, we have

$$\begin{aligned}(A - BB^\top X_\infty)^\top Q_\infty &= Q_\infty Q_\infty^\top A^\top Q_\infty - Q_\infty Q_\infty^\top X_\infty BB^\top Q_\infty \\ &= Q_\infty \left(Q_\infty^\top A^\top Q_\infty - Q_\infty^\top X_\infty BB^\top Q_\infty\right).\end{aligned}$$

This shows range(X_∞) is $(A - BB^\top X_\infty)^\top$–invariant. □

According to Theorem 3.3, the solution of the DRE (1) is given by

$$X(t) = X_\infty - e^{t\hat{A}^\top} X_\infty \Big(I_n - \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top}\right) X_\infty\Big)^{-1} e^{t\hat{A}},$$

where $\hat{A} = A - BB^\top X_\infty$. The identity $(I_n - P(t))^{-1} = I_n + (I_n - P(t))^{-1} P(t)$ leads to

$$\begin{aligned}X(t) = X_\infty &- e^{t\hat{A}^\top} X_\infty e^{t\hat{A}} \\ &- e^{t\hat{A}^\top} X_\infty \Big(I_n - \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top}\right) X_\infty\Big)^{-1} \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top}\right) X_\infty e^{t\hat{A}}.\end{aligned} \tag{10}$$

4.1.1. Derivation by using the exact solution X_∞ of the ARE

195 By Lemma 4.1, it holds that range (X_∞) is invariant under $\hat{A}^\top$. Assume now that X_∞ is given in factorized form, this means that $X_\infty = Z_\infty Z_\infty^\top$ and $Z_\infty \in \mathbb{R}^{n\times p}$ and $1 \leq p = \text{rank}\,(X_\infty) \leq n$. If rank$(X_\infty) = 0$, then also $X_\infty = 0_n$ as well as the solution $X(t)$. Now it holds that range(X_∞) = range(Z_∞) and consequently range(Z_∞) is invariant under $\hat{A}^\top$.

Utilizing the compact singular value decomposition of Z_∞, we obtain matrices $Q_\infty \in \mathbb{R}^{n\times p}$, $S_\infty \in \mathbb{R}^{p\times p}$, and $V_\infty \in \mathbb{R}^{p\times p}$, such that $Z_\infty = Q_\infty S_\infty V_\infty^\top$, range (Q_∞) = range (Z_∞) and $Z_\infty Z_\infty^\top = Q_\infty S_\infty^2 Q_\infty^\top$. Because of

the invariance, we get

$$e^{t\hat{A}^\top} Q_\infty = Q_\infty e^{tQ_\infty^\top \hat{A}^\top Q_\infty}.$$

Now observe that

$$e^{t\hat{A}^\top} X_\infty = e^{t\hat{A}^\top} Z_\infty Z_\infty^\top = e^{t\hat{A}^\top} Q_\infty S_\infty^2 Q_\infty^\top = Q_\infty e^{tQ_\infty^\top \hat{A}^\top Q_\infty} S_\infty^2 Q_\infty^\top. \tag{11}$$

Therefore, the solution $X(t)$ can be written in the form

$$X(t) = X_\infty - Q_\infty \tilde{X}(t) Q_\infty^\top. \tag{12}$$

We use the DRE (1) and Equation (12) and obtain a differential equation for $\tilde{X}(t)$:

$$\begin{aligned}
\dot{\tilde{X}}(t) &= Q_\infty^\top \hat{A}^\top Q_\infty \tilde{X}(t) + \tilde{X}(t) Q_\infty^\top \hat{A} Q_\infty + \tilde{X} Q_\infty^\top B B^\top Q_\infty \tilde{X}(t), \\
\tilde{X}(0) &= Q_\infty^\top X_\infty Q_\infty.
\end{aligned}$$

4.1.2. Derivation using a low-rank approximation X_N of the exact solution X_∞ of the ARE

Let now $Z_N Z_N^\top = X_N \approx X_\infty$ be a low-rank approximation obtained by a numerical method. We replace X_∞ by X_N in Equation (10) and obtain

$$\begin{aligned}
X(t) \approx X_N &- e^{t\left(A - BB^\top X_N\right)^\top} X_N e^{t\left(A - BB^\top X_N\right)} \\
&- e^{t\left(A - BB^\top X_N\right)^\top} X_N \left(I_n - \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top}\right) X_\infty\right)^{-1} \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^\top}\right) X_N e^{t\left(A - BB^\top X_N\right)}.
\end{aligned}$$

Let $Z_N = Q_N S_N V_N^\top$ be the compact singular value decomposition of the low-rank factor. According to Equation (11), we propose to approximate the action of the matrix exponential by

$$\begin{aligned}
e^{t\left(A - BB^\top X_N\right)^\top} X_N &= e^{t\left(A - BB^\top X_N\right)^\top} Z_N Z_N^\top = e^{t\left(A - BB^\top X_N\right)^\top} Q_N S_N^2 Q_N^\top \\
&\approx Q_N e^{tQ_N^\top \left(A - BB^\top X_N\right)^\top Q_N} S_N^2 Q_N^\top.
\end{aligned}$$

Therefore, we obtain the Galerkin approximation $X(t) \approx X_N - Q_N \tilde{X}_N(t) Q_N^\top$ for the numerical approximation. Again we use the DRE (1) and get a differential equation for $\tilde{X}_N(t)$

$$\begin{aligned}
\dot{\tilde{X}}_N(t) &= Q_N^\top \left(A - BB^\top X_N\right)^\top Q_N \tilde{X}_N(t) + \tilde{X}_N(t) Q_N^\top \left(A - BB^\top X_N\right) Q_N \\
&\quad + \tilde{X}_N(t) Q_N^\top B B^\top Q_N \tilde{X}_N(t) + Q_N^\top \mathcal{R}(X_N) Q_N. \\
\tilde{X}_N(0) &= Q_N^\top X_N Q_N.
\end{aligned}$$

We assume that the numerical low-rank approximation is accurate enough such that $\mathcal{R}(X_N) \leq \tau \ll 1$. Then it holds:

$$\left\|Q_N^\top \mathcal{R}(X_N) Q_N\right\|_2 \leq \|\mathcal{R}(X_N)\|_2 \leq \tau \ll 1.$$

This means that the projected residual $Q_N^\top \mathcal{R}(X_N) Q_N$ is even smaller than the residual of the ARE $\mathcal{R}(X_N)$, and, therefore, we neglect the projected residual.

4.2. Reduced Trial Space for the Galerkin Approach using Eigenvalue Decay

Let $X_\infty = Z_\infty Z_\infty^\top$ be the exact stabilizing solution of the ARE (2). Moreover, let $Z_\infty = Q_\infty S_\infty V_\infty^\top$ be its compact singular value decomposition, such that $Q_\infty \in \mathbb{R}^{n\times p},\ S_\infty \in \mathbb{R}^{p\times p}, V_\infty \in \mathbb{R}^{p\times p}$, and $Z_\infty = Q_\infty S_\infty V_\infty^\top$. The compact singular value decomposition of Z_∞ gives a spectral decomposition of X_∞ that is

$$X_\infty = Z_\infty Z_\infty^\top = Q_\infty S_\infty^2 Q_\infty^\top, \text{ and } S_\infty^2 = \operatorname{diag}\left(\lambda_1^\downarrow(X_\infty), \ldots, \lambda_p^\downarrow(X_\infty)\right).$$

This means that the diagonal matrix S_∞^2 contains all nonzero eigenvalues of X_∞ in a weakly decreasing fashion. We have that $\operatorname{range}(X_\infty) = \operatorname{range}(Z_\infty) = \operatorname{range}(Q_\infty)$. Because of Theorem 2.3, it holds that $\operatorname{range}(Q_\infty) = \mathcal{K}\left(A^\top, C^\top\right)$. According to Theorem 2.5, we can represent the solution of the DRE in the form

$$X(t) = Q_\infty Q_\infty^\top X(t) Q_\infty Q_\infty^\top.$$

This representation has the advantage that the absolute value of the entries of $Q_\infty^\top X(t) Q_\infty$ can be bounded by the eigenvalues of X_∞.

Theorem 4.1.

Let (A, B) be stabilizable and (A, C) be detectable. Moreover, let $X_\infty \in \mathbb{R}^{n\times n}$ be the unique symmetric positive semidefinite solution of the ARE (2) *and $q_1, \ldots, q_n \in \mathbb{R}^n$ be a system of orthonormal eigenvectors of X_∞ corresponding to the eigenvalues $\lambda_1^\downarrow(X_\infty), \ldots, \lambda_n^\downarrow(X_\infty) \in \mathbb{R}$. Then for all $i, j = 1, \ldots, n$ and $t \geq 0$, the following holds:*

$$\left|q_i^\top X(t) q_j\right| \leq \sqrt{\lambda_i^\downarrow(X_\infty)\, \lambda_j^\downarrow(X_\infty)}, \tag{14}$$

where X is the unique solution of the DRE (1) *with $X_0 = 0_n$.*

Proof. We use the fact that for any symmetric positive semidefinite matrix $A \in \mathbb{R}^{n\times n}$ the inequality $\left|A_{i,j}\right| \leq \sqrt{A_{i,i} A_{j,j}}$ holds for all $i, j = 1, \ldots, n$; see [25, Obs. 7.1.2, Problem 7.1.P1]. According to Theorem 2.4, the inequality $0_n \preccurlyeq X(t) \preccurlyeq X_\infty$ holds for all $t \geq 0$. This implies $0_n \preccurlyeq X_\infty - X(t)$. Let the columns of $Q \in \mathbb{R}^{n\times n}$ be $q_1, \ldots, q_n$, then the matrix $Q^\top X_\infty Q - Q^\top X(t) Q$ is symmetric positive semidefinite.

Therefore, the diagonal entries $\lambda_i^{\downarrow}(X_\infty) - q_i^\top X(t) q_i$ are nonnegative. We apply the inequality and obtain

$$\left|q_i^\top X(t) q_j\right| \leq \sqrt{\left(\lambda_i^{\downarrow}(X_\infty) - q_i^\top X(t) q_i\right)\left(\lambda_j^{\downarrow}(X_\infty) - q_j^\top X(t) q_j\right)}$$

for $i \neq j$. The inequality $0_n \preccurlyeq X(t)$ implies that $0 \leq q_i^\top X(t) q_i$ and the claim follows. □

Let the columns of Q_∞ be $q_1, \ldots, q_p$. Due to the decay of the eigenvalues $\lambda_k^{\downarrow}(X_\infty)$ of the solution of the ARE (2) and the inequality (14) from Theorem 4.1, the values $\left|q_i^\top X(t) q_j\right|$ also decay for $i + j$ increasing. We have that

$$X(t) = Q_\infty Q_\infty^\top X(t) Q_\infty Q_\infty^\top = \sum_{i,j=1}^{p} \left(q_i^\top X(t) q_j\right) q_i q_j^\top.$$

For quick enough eigenvalue decay, we expect that $\left|q_i^\top X(t) q_j\right| \leq \sqrt{\lambda_i^{\downarrow}(X_\infty)\,\lambda_j^{\downarrow}(X_\infty)} \approx 0$ for $i + j$ large enough. We truncate the sum and obtain

$$X(t) \approx \sum_{i,j=1}^{k} \left(q_i^\top X(t) q_j\right) q_i q_j^\top = Q_{\infty,k} Q_{\infty,k}^\top X(t) Q_{\infty,k} Q_{\infty,k}^\top,$$

where $Q_{\infty,k} = \begin{bmatrix} q_1, \ldots, q_k \end{bmatrix} \in \mathbb{R}^{n \times k}$. We also consider the appropriate real linear space

$$\mathcal{Q}_{\infty,k} := \left\{ Q_{\infty,k} Y Q_{\infty,k}^\top \mid Y \in \mathbb{R}^{k \times k} \right\} \subseteq \mathbb{R}^{n \times n}$$

together with the orthogonal projection

$$\mathcal{P}_k : \mathbb{R}^{n \times n} \to \mathcal{Q}_{\infty,k},\ \mathcal{P}_{\infty,k}(X) = Q_{\infty,k} Q_{\infty,k}^\top X Q_{\infty,k} Q_{\infty,k}^\top.$$

The columns of $Q_{\infty,k}$ are orthonormal. Consequently, it holds that $\mathcal{P}_{\infty,k}^2(X) = \mathcal{P}_{\infty,k}(X)$. Moreover, the projection $\mathcal{P}_{\infty,k}$ is orthogonal because of

$$\begin{aligned}
\langle X - \mathcal{P}_{\infty,k}(X), Q_{\infty,k} Y Q_{\infty,k} \rangle_F &= \langle X - Q_{\infty,k} Q_{\infty,k}^\top X Q_{\infty,k} Q_{\infty,k}^\top, Q_{\infty,k} Y Q_{\infty,k}^\top \rangle_F \\
&= \langle X, Q_{\infty,k} Y Q_{\infty,k}^\top \rangle_F - \langle Q_{\infty,k} Q_{\infty,k}^\top X Q_{\infty,k} Q_{\infty,k}^\top, Q_{\infty,k} Y Q_{\infty,k}^\top \rangle_F \\
&= \langle X, Q_{\infty,k} Y Q_{\infty,k}^\top \rangle_F - \langle X, Q_{\infty,k} Y Q_{\infty,k}^\top \rangle_F = 0
\end{aligned}$$

for all $Y \in \mathbb{R}^{k \times k}$. Therefore, the best approximation of $X(t)$ in $\mathcal{Q}_{\infty,k}$ is given by

$$\sum_{i,j=1}^{k} \left(q_i^\top X(t) q_j\right) q_i q_j^\top = \mathcal{P}_{\infty,k}(X(t)) = \operatorname*{argmin}_{X \in \mathcal{Q}_{\infty,k}} \|X - X(t)\|_F.$$

For the projection error, we obtain

$$\left\|X(t) - \mathcal{P}_{\infty,k}(X(t))\right\|_F = \left\| \sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{p} \left(q_i^\top X(t) q_j\right) q_i q_j^\top \right\|_F = \sqrt{\sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{p} \left|q_i^\top X(t) q_j\right|^2}$$
$$\leq \sqrt{\sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{p} \lambda_i^\downarrow(X_\infty)\, \lambda_j^\downarrow(X_\infty)}.$$

Since the eigenvalues $\lambda_{p+1}^\downarrow(X_\infty), \ldots, \lambda_n^\downarrow(X_\infty)$ are 0, we obtain

$$\left\|X(t) - \mathcal{P}_{\infty,k}(X(t))\right\|_F \leq \sqrt{\sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{n} \lambda_i^\downarrow(X_\infty)\, \lambda_j^\downarrow(X_\infty)}. \tag{15}$$

We measure the projection error in the 2-norm.

Theorem 4.2.

Let (A, B) be stabilizable and (A, C) be detectable. Moreover, let $X_\infty \in \mathbb{R}^{n\times n}$ be the unique symmetric positive semidefinite solution of the ARE (2)*. Then for all $k = 1, \ldots, n-1$ and all $t \geq 0$, the projection error is bounded by*

$$\left\|X(t) - \mathcal{P}_{\infty,k}(X(t))\right\|_2 \leq 2\sqrt{\lambda_{k+1}^\downarrow(X_\infty)\, \lambda_1^\downarrow(X_\infty)},$$

where X is the unique solution of the DRE (1) *with $X_0 = 0_n$.*

Proof. Again, Theorem 2.4 yields $0_n \preccurlyeq X(t) \preccurlyeq X_\infty$ for all $t \geq 0$. Moreover,

$$X_\infty = \mathcal{P}_{\infty,k}(X_\infty) + X_\infty - \mathcal{P}_{\infty,k}(X_\infty) \preccurlyeq \mathcal{P}_{\infty,k}(X_\infty) + \left\|X_\infty - \mathcal{P}_{\infty,k}(X_\infty)\right\|_2 I_n.$$

This leads to the inequality

$$0_n \preccurlyeq X(t) \preccurlyeq Q_{\infty,k} Q_{\infty,k}^\top X_\infty Q_{\infty,k} Q_{\infty,k}^\top + \lambda_{k+1}^\downarrow(X_\infty)\, I_n. \tag{16}$$

The orthogonal projection matrix $I_n - Q_{\infty,k}Q_{\infty,k}^\top$ is symmetric, and $\left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right) Q_{\infty,k}$ vanishes. We multiply (16) from the left and right with $I_n - Q_{\infty,k}Q_{\infty,k}^\top$, and obtain

$$0_n \preccurlyeq \left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right) X(t) \left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right) \preccurlyeq \lambda_{k+1}^\downarrow(X_\infty) \left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right). \tag{17}$$

The 2-norm of any positive semidefinite matrix is equal to its largest eigenvalue. Using (17), we have the

bound

$$\left\|\left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right) X(t)^{1/2}\right\|_2^2 = \left\|\left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right) X(t) \left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right)\right\|_2 \leq \lambda_{k+1}^{\downarrow}\left(X_\infty\right).$$

Finally,

$$\begin{aligned}
\left\|X(t) - \mathcal{P}_{\infty,k}(X(t))\right\|_2 &= \left\|X(t) - Q_{\infty,k}Q_{\infty,k}^\top X(t) Q_{\infty,k}Q_{\infty,k}^\top\right\|_2 \\
&= \left\|\left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right) X(t) + Q_{\infty,k}Q_{\infty,k}^\top X(t) \left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right)\right\|_2 \\
&\leq 2\left\|\left(I_n - Q_{\infty,k}Q_{\infty,k}^\top\right) X(t)^{1/2}\right\|_2 \left\|X(t)^{1/2}\right\|_2 \\
&\leq 2\sqrt{\lambda_{k+1}^{\downarrow}\left(X_\infty\right)\left\|X(t)\right\|_2} \leq 2\sqrt{\lambda_{k+1}^{\downarrow}\left(X_\infty\right)\lambda_1^{\downarrow}\left(X_\infty\right)}.
\end{aligned}$$

□

Therefore, we propose to set up a trial space for the Galerkin approach using a system of 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 singular value decomposition of the numerical low-rank approximation of X_∞ can be used to obtain an approximation of the eigenvectors corresponding to the largest eigenvalues. By virtue of Theorems 4.1 and 4.2, we remove the small singular values from the singular value decomposition. This also reduces the dimension of the trial space. Let $Z_N = Q_N S_N V_N^\top$ be the truncated reduced singular value decomposition of the low-rank approximation. With that, the trial space for the Galerkin approach is given by $\left\{Q_N \tilde{X} Q_N^\top \mid \tilde{X} \in \mathbb{R}^{p\times p}\right\}$, and, as $X(t)$ converges to X_∞ and $X_\infty \approx Z_N Z_N^\top$, we propose the Galerkin approach $X(t) \approx Z_N Z_N^\top - Q_N \tilde{X}(t) Q_N^\top$.

Example 4.1 (Decay of Absolute Values of Entries)**.**

We illustrate the decay of $\left|q_i^\top X(t) q_j\right|$ *in* Figures 4a-4e. *We have chosen the same matrices as in* Example 2.1. *To improve the visualization, all values below machine precision were set to machine precision. The eigenvalue decay of the solution* X_∞ *of the corresponding* ARE *is shown in* Figure 4f.

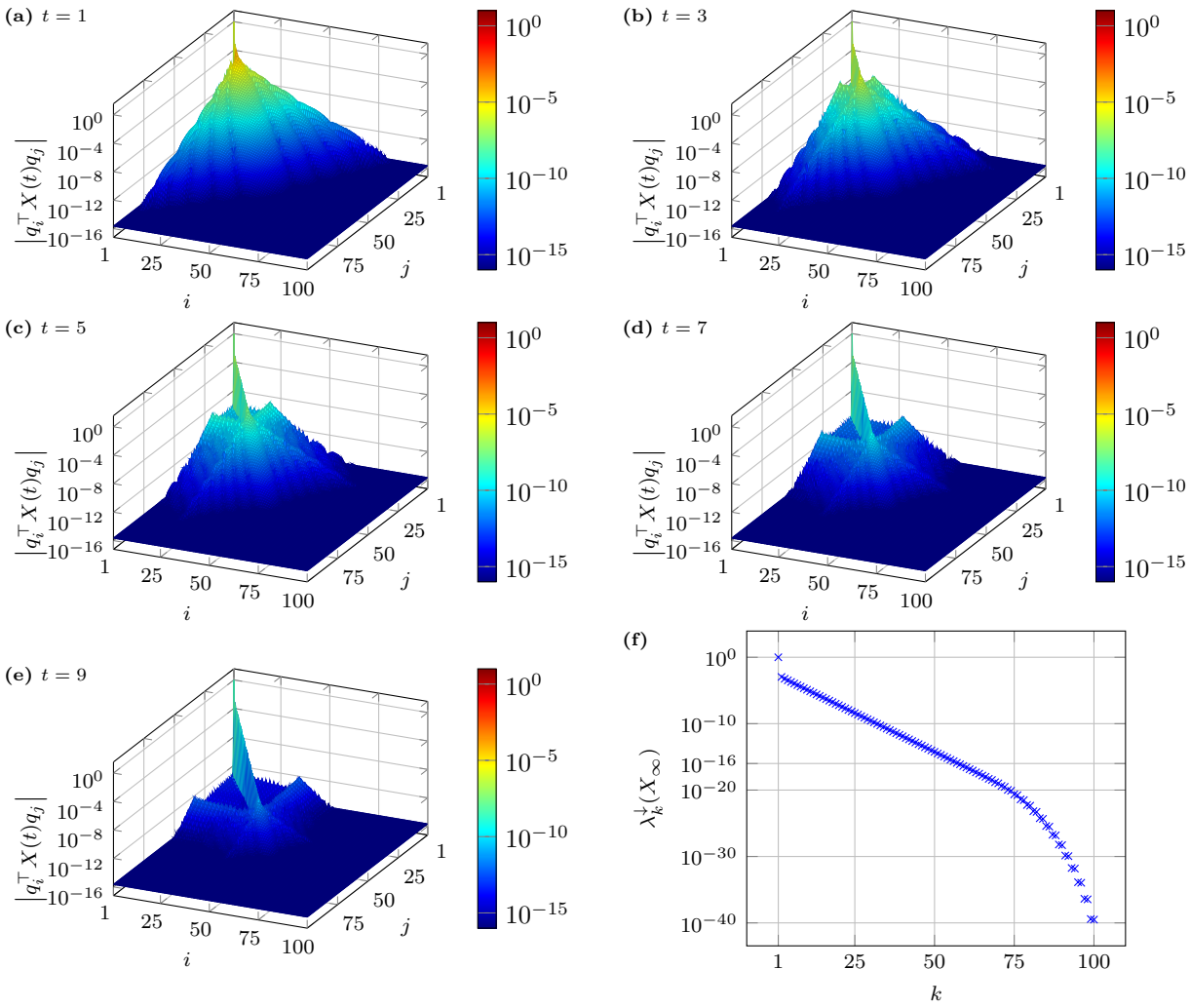


Fig. 4. **(a)–(e)** Decay of $\left|q_i^\top X(t) q_j\right|$ for $t \in \{1, 3, 5, 7, 9\}$. **(f)** The eigenvalue decay of X_∞.

Remark 4.1.

With minor adjustments, all arguments also hold for the generalized DRE

$$M^\top \dot{X}(t) M = A^\top X(t) M + M^\top X(t) A - M^\top X(t) B B^\top X(t) M + C^\top C, \tag{18a}$$

$$X(0) = 0_n, \tag{18b}$$

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 approach can be implemented based on Algorithm 3.

Algorithm 3 Galerkin approach for the generalized DRE (18) (`ARE-Galerkin`)

Assumption: $\left(AM^{-1}, B\right)$ is stabilizable and (AM^{-1}, CM^{-1}) is detectable.

Input: $M,\ A \in \mathbb{R}^{n\times n},\ B \in \mathbb{R}^{n\times b},\ C \in \mathbb{R}^{c\times n}$, truncation tolerance $tol_{\texttt{trunc}} > 0$.

Output: $X(t) \approx Z_\infty Z_\infty^\top - Q_\infty \tilde{X}(t) Q_\infty^\top$ that approximates the solution to the generalized DRE (18).

% Solve the ARE:

$A^\top X_\infty M + M^\top X_\infty A - M^\top X_\infty B B^\top X_\infty M + C^\top C = 0_n$ for $X_\infty \approx Z_\infty Z_\infty^\top$;

% Compute compact singular value decomposition:

$[Q_\infty, S_\infty, \sim] = \texttt{svd}(Z_\infty, 0)$;

% Truncate all singular values smaller than tolerance and get truncated low-rank factor:

$idx = \texttt{diag}(S_\infty) \geq tol_{\texttt{trunc}} \cdot S_\infty(1,1)$;

$S_\infty = S_\infty(idx, idx)$;

$Q_\infty = Q_\infty(:, idx)$;

$Z_\infty = Q_\infty S_\infty$;

% Compute matrices:

$\tilde{A} = Q_\infty^\top \left(AM^{-1} - BB^\top Z_\infty Z_\infty^\top\right) Q_\infty$;

$\tilde{B} = Q_\infty^\top B$;

% Solve the differential equation using Algorithm 2:

$\dot{\tilde{X}}(t) = \tilde{A}^\top \tilde{X}(t) + \tilde{X}(t)\tilde{A} + \tilde{X}(t)\tilde{B}\tilde{B}^\top \tilde{X}(t),\ \tilde{X}(0) = S_\infty^2$;

4.3. Nonzero Initial Condition

In this section, we extend our discussion to the case of positive semidefinite nonzero initial conditions $X(0) = Z_0 Z_0^\top$. Here, the inequality

$$0_n \preccurlyeq X(t) \preccurlyeq X_\infty \tag{19}$$

needs to be established by other means than Theorem 2.4, which requires $\mathcal{R}(X(0)) \succcurlyeq 0_n$.

We distinguish the cases of $\mathcal{R}(X(0))$ positive (semi)-definite, negative (semi)-definite, and indefinite.

If $\mathcal{R}(X(0)) \succcurlyeq 0_n$, then, as for $X_0 = 0$, Theorem 2.4 readily applies.

Curiously, the case of $\mathcal{R}(X(0)) \preccurlyeq 0_n$ fits the framework without relying on the solution of the ARE. In fact, in this case, the solution $X(t)$ is monotonically weakly decreasing; cf. [26, Thm. 4.1.8]. Still, the solution X is positive semidefinite for all $t \geq 0$, so that (19) can be replaced by $0_n \preccurlyeq X(t) \preccurlyeq Z_0 Z_0^\top$ for all $t \geq 0$. In particular, it follows that $\mathrm{im}(X(t)) \subseteq \mathrm{im}(Z_0)$ for all $t \geq 0$, so that a trial space is readily defined by a basis of $\mathrm{im}(Z_0)$. Thus, a basis can be computed by a QR factorization or a compact singular value decomposition of Z_0.

The symmetric but indefinite case that we write as $\mathcal{R}(X(0)) = Z_+ Z_+^\top - Z_- Z_-^\top$, requires additional reasoning. One may compute a suitable upper bound $\tilde{X}_\infty$ that replaces X_∞ in (19) as follows. Consider the modified DRE

$$\begin{aligned}\dot{\tilde{X}}(t) &= \tilde{\mathcal{R}}(\tilde{X}(t)) := A^\top \tilde{X}(t) + \tilde{X}(t)A - \tilde{X}(t)BB^\top \tilde{X}(t) + C^\top C + Z_- Z_-^\top,\\ \tilde{X}(0) &= Z_0 Z_0^\top.\end{aligned}$$

By construction, it holds $\tilde{\mathcal{R}}(\tilde{X}(0)) = Z_+ Z_+^\top \succcurlyeq 0_n$ so that $\tilde{X}(t)$ is monotonically weakly increasing for all $t \geq 0$. 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 $0_n = \tilde{\mathcal{R}}(\tilde{X}_\infty)$. This means that $0_n \preccurlyeq \tilde{X}(t) \preccurlyeq \tilde{X}_\infty$. A standard comparison argument gives that $X(t) \leq \tilde{X}(t)$ for all $t \geq 0$; cf. [26, Thm. 4.1.4]. With that, we have $0_n \preccurlyeq X(t) \preccurlyeq \tilde{X}(t) \preccurlyeq \tilde{X}_\infty$ for all $t \geq 0$, and the bounds on the projection error (Theorem 4.1, Equation 15, and Theorem 4.2) can be established analogously.

5. Numerical Experiments

To quantify the performance of Algorithm 3, we consider a number of DREs that are used to define optimal controls. Concretely, we consider the generalized DRE (18) and their realizations. First, we consider the `RAIL` benchmark example, that is a finite element discretization of a heat equation; see [50] for the model description. The second example, `CONV_DIFF`, results from a finite-differences discretized heat equation with convection on the unit square with homogeneous Dirichlet boundary conditions; see [51]. The third and the fourth example, `FLOW` and `COOKIE`, are taken from [52] and [53].

We compare the proposed method with the splitting methods developed in [13, 54]. Splitting methods are based on a splitting of the DRE into an affine and nonlinear subproblem. The advantages of that approach lie in the fact that the nonlinear subproblem can be solved by an explicit solution formula. The numerical solution of the linear subproblem is based on approximating the action of the matrix exponential by means of Gauss-Legendre Runge-Kutta methods. We employed the *Lie* and *Strang* splittings of order 1 and 2, respectively, as well as 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 $X_{\texttt{ref}}(t)$ using `SYMMETRIC8` with a constant time step size h. The basic information about the setup of the benchmark problems are given in Table 1.

The reference solutions were computed on a machine with 2× Xeon® Skylake Gold 6130 @ 2.10GHz CPU with 16 cores, 192 GB RAM, and MATLAB 2019b. All other computations are carried out on a machine with 2× Xeon® Skylake Silver 4110 @ 2.10GHz CPU with 8 cores, 192 GB RAM, and MATLAB 2019b.

Instance	n	Matrices	Interval	Reference Solution
RAIL	5177	M symmetric positive definite, A symmetric, $M^{-1}A$ stable, $B \in \mathbb{R}^{n\times 6},\ C \in \mathbb{R}^{7\times n}$	$[0, 4512]$	SYMMETRIC8, $h = 2^{-5}$
CONV_DIFF	6400	$M = I_n$, A nonsymmetric and stable, $B \in \mathbb{R}^{n\times 1},\ C \in \mathbb{R}^{1\times n}$	$[0, 0.125]$	SYMMETRIC8, $h = 2^{-20}$
FLOW	9669	M diagonal positive definite, A symmetric and stable, $M^{-1}A$ stable, $B \in \mathbb{R}^{n\times 1},\ C \in \mathbb{R}^{5\times n}$	$[0, 0.25]$	SYMMETRIC8, $h = 2^{-21}$
COOKIE	7488	M nonsymmetric, A nonsymmetric and stable, $M^{-1}A$ stable, $B \in \mathbb{R}^{n\times 1},\ C \in \mathbb{R}^{4\times n}$	$[0, 4]$	SYMMETRIC8, $h = 2^{-16}$

Table 1: Information about benchmark problems.

We report the absolute and relative errors

$$\|X(t) - X_{\texttt{ref}}(t)\| \quad \text{and} \quad \frac{\|X(t) - X_{\texttt{ref}}(t)\|}{\|X_{\texttt{ref}}(t)\|},$$

where $X(t)$ is the numerical approximation, and $X_{\texttt{ref}}(t)$ is the reference solution, in 2-norm and Frobenius norm. We also report the norm of the reference solution $\|X_{\texttt{ref}}(t)\|$ and the convergence to the stationary point $\|X_{\texttt{ref}}(t) - X_\infty\|_2$.

Numerical results for the Galerkin approximation from Algorithm 3 and for the splitting scheme based solvers can be found in Appendix A and Appendix B. The computational costs for both methods are given in Section 5.2. Also, we evaluate the best approximation in the trial space of the reference solution, which is given by

$$X_{\texttt{best}}(t) := Q_\infty Q_\infty^\top X_{\texttt{ref}}(t) Q_\infty Q_\infty^\top = \operatorname*{argmin}_{X \in \left\{Q_\infty \tilde{X} Q_\infty^\top \mid \tilde{X} \in \mathbb{R}^{k\times k}\right\}} \|X - X_{\texttt{ref}}(t)\|_F,$$

where Q_∞ is the matrix from Algorithm 3 line 6.

The code of the implementation is available, as mentioned in Figure 5.

5.1. Galerkin Approach and Splitting Schemes

The initial step of Algorithm 3 requires the solution to the associated ARE. For this task, we use the RADI algorithm that iteratively computes the numerical solution to the following absolute and relative residuals

$$\left\|A^\top Z_\infty Z_\infty^\top M + M^\top Z_\infty Z_\infty^\top A - M^\top Z_\infty Z_\infty^\top B B^\top Z_\infty Z_\infty^\top M + C^\top C\right\|_2$$

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

`doi:10.5281/zenodo.4460618`

under the GPLv2+ license and is authored by Maximilian Behr.

Fig. 5. Link to code and data.

and

$$\frac{\left\|A^\top Z_\infty Z_\infty^\top M + M^\top Z_\infty Z_\infty^\top A - M^\top Z_\infty Z_\infty^\top B B^\top Z_\infty Z_\infty^\top M + C^\top C\right\|_2}{\left\|C^\top C\right\|_2}.$$

The achieved values for the different test setups and the number of columns of the corresponding Z_∞ after truncation (see Algorithm 3 line 3), that define the dimension of the reduced model, are listed in Table 2.

Instance	n	$tol_{\texttt{trunc}}$	Size of Galerkin system	Absolute residual	Relative residual
RAIL		10^{-8}	193	$2.91 \cdot 10^{-14}$	$2.43 \cdot 10^{-15}$
		$\varepsilon_{\texttt{mach}}$	279	$3.25 \cdot 10^{-14}$	$2.71 \cdot 10^{-15}$
CONV_DIFF		10^{-8}	36	$1.37 \cdot 10^{-10}$	$3.06 \cdot 10^{-14}$
		$\varepsilon_{\texttt{mach}}$	54	$1.39 \cdot 10^{-10}$	$3.11 \cdot 10^{-14}$
FLOW		10^{-8}	115	$2.85 \cdot 10^{-8}$	$2.85 \cdot 10^{-8}$
		$\varepsilon_{\texttt{mach}}$	252	$1.06 \cdot 10^{-11}$	$1.06 \cdot 10^{-11}$
COOKIE		10^{-8}	122	$1.29 \cdot 10^{-14}$	$3.07 \cdot 10^{-12}$
		$\varepsilon_{\texttt{mach}}$	169	$1.33 \cdot 10^{-14}$	$3.16 \cdot 10^{-12}$

Table 2: Residuals for the ARE $0_n = A^\top XM + M^\top XA - M^\top XBB^\top XM + C^\top C$.

Instance	n	Time to solve ARE (s)
RAIL		0.86
CONV_DIFF		1.59
FLOW		2.10
COOKIE		2.68

Table 3: Timings for the ARE $0_n = A^\top XM + M^\top XA - M^\top XBB^\top XM + C^\top C$.

The 1-norm bound for the matrix exponential $tol_{\texttt{exp}}$ of Algorithm 2 was set to 10^{10}. The resulting step sizes are given in Table 4. Here, we used two values for the truncation threshold $tol_{\texttt{trunc}}$, namely the machine precision $\varepsilon_{\texttt{mach}}$ that in this setup is approximately $2 \cdot 10^{-16}$ and the rougher value $tol_{\texttt{trunc}} = 10^{-8}$.

Instance	n	Step sizes h
RAIL	5177	$\{2^0, 2^{-1}, \ldots, 2^{-5}\}$
CONV_DIFF	6400	$\{2^{-12}, 2^{-13}, \ldots, 2^{-16}\}$
FLOW	9669	$\{2^{-15}, 2^{-16}, \ldots, 2^{-20}\}$
COOKIE	7488	$\{2^{-15}, 2^{-16}, \ldots, 2^{-20}\}$

Table 4: Step sizes h for the modified Davison-Maki method (Algorithm 2).

We plot the numerical errors in Figures A.7, A.10, A.13, and A.16. The Figures A.8, A.9, A.11, A.12, A.14, A.15, A.17, and A.18 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 3.3, the accuracy of the *modified Davison-Maki method* is independent of the step size; cf. Figures A.7b and A.7d. Still, we compute the solution on different time grids, since for control applications, the values of the solution might be needed at many time instances.

The computational times for ARE-Galerkin 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 6a, $tol_{\texttt{trunc}} = \varepsilon_{\texttt{mach}}$), 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 what makes the absolute error comparatively large; see Figures A.8 and A.14 in Appendix A.

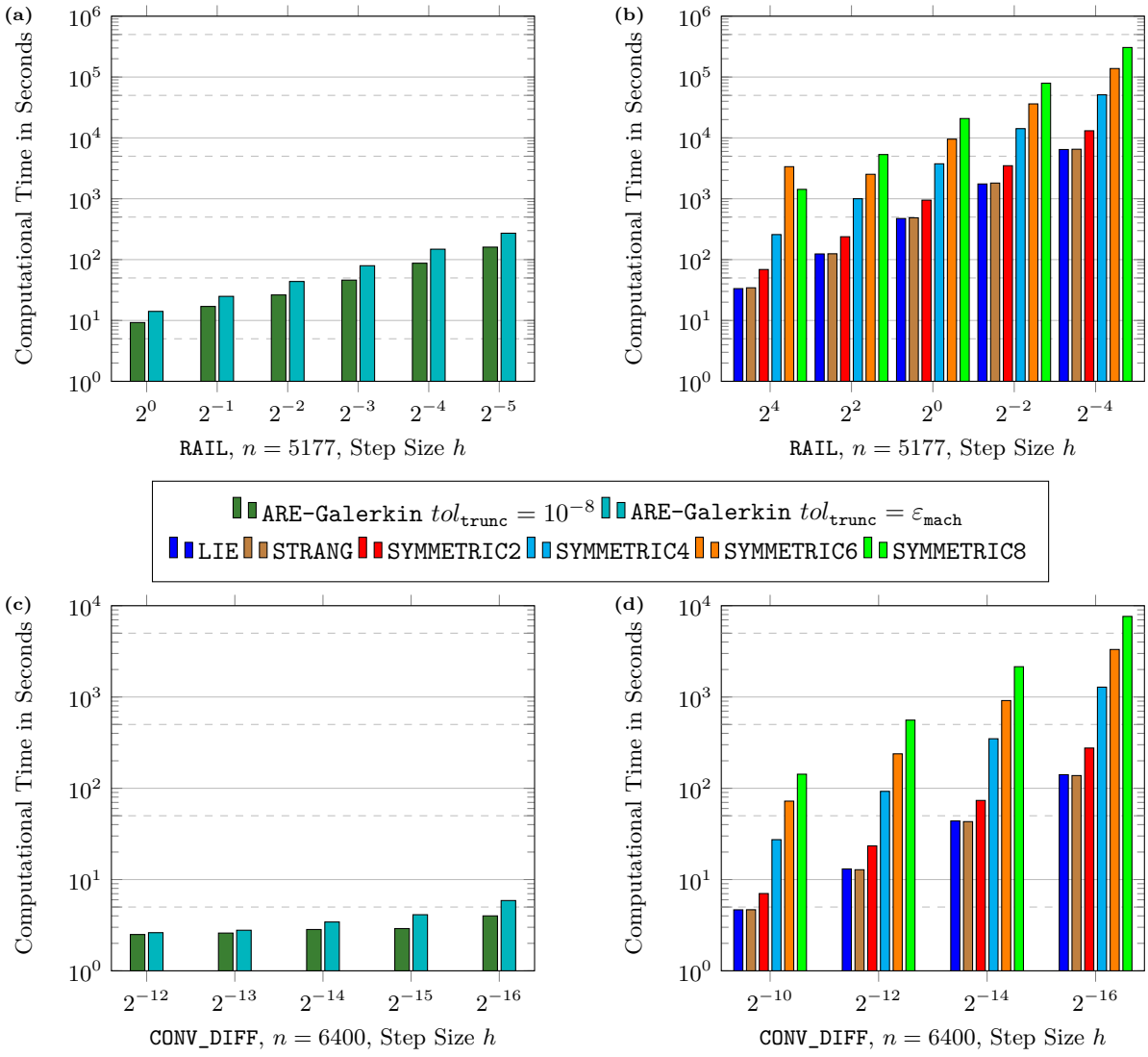
The LIE, STRANG, and SYMMETRIC2 splitting schemes gave an absolute and relative error nearly at the same level, therefore the Figures B.19–B.22 only show the error of the SYMMETRIC2 splitting scheme.

In all examples, in terms of accuracy, the ARE-Galerkin ($tol_{\texttt{trunc}} = \varepsilon_{\texttt{mach}}$) approximation is nearly at the same level as the high order splitting schemes; cf. Figures A.7b, B.19f and Figures A.10b, B.20f. However, we note that the ARE-Galerkin method does not give the best possible approximation in the trial space; compare the error levels for X_{best}.

In any case, the ARE-Galerkin method outperforms the splitting methods in terms of computational time versus accuracy in all test examples. The performance can be further improved by adapting the truncation threshold $tol_{\texttt{trunc}}$; cf. line 3 of Algorithm 3. Apart from the savings in the timings (Figures 6a, 6c, 6e, and 6g) the reduced memory requirements can be significant. For the RAIL example, the rougher tolerance, namely 10^{-8} instead of machine precision, means a reduction in storage by a factor of $279^2/193^2 \approx 2$; cf. Table 2. 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 with respect to machine precision; cf. Figure A.7d. For the other examples, the approximation accuracy was only slightly affected by the larger truncation

threshold The most favorable example is the FLOW example, where the relaxed truncation threshold led to savings of a factor $^{407s}/_{107s} \approx 4$ ($h = 2^{-20}$) in the timings (Figure 6e) and a factor of $^{252^2}/_{115^2} \approx 5$ in memory requirements (Table 2) while, except for a short initial phase, maintaining the same approximation accuracy (Figure A.13d).

5.2. Computational Time



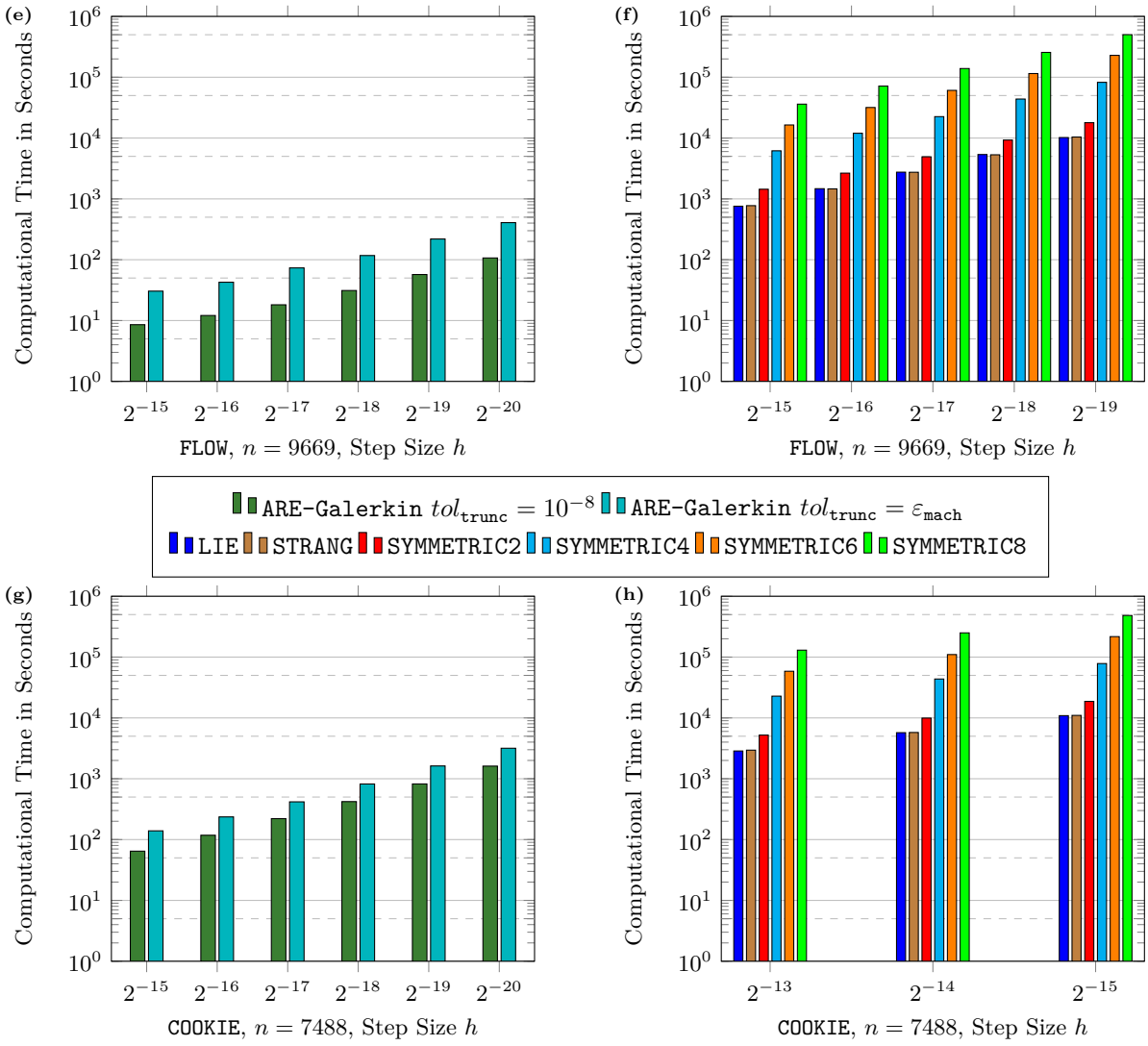


Fig. 6. **(a), (c), (e), (g)** Timings for `ARE-Galerkin`. **(b), (d), (f), (h)** Timings for splitting schemes.

5.3. Large-scale Examples

We consider the benchmark problems `RAIL`, `CONV_DIFF`, and `FLOW` for finer space discretization resulting in a larger state-space dimension n. Moreover, we consider the `CHIP` model [52]. The structural properties of the matrices A, E, B, and C of the `CHIP` model are the same as for the `FLOW` model; cf. Table 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. Table 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 Algorithm 3 are given in Table 6. We report the time for the numerical approximation of the ARE ($t_{\texttt{ARE}}$, Algorithm 3), the computation of the singular value decomposition ($t_{\texttt{svd}}$), the assembly of the system matrices of the

Galerkin system ($t_{\texttt{gal}}$), the approximation of the matrix exponential and the norm computation ($t_{\texttt{expm}}$); cf. Algorithm 3 line 1, lines 2–6, lines 7–8, and Algorithm 2 lines 3–4. The computational time for the time-stepping Algorithm 2 lines 8–13 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 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 only increase slightly. Also, the computed residuals turn out to be not affected by the sizes of the initial systems.

Instance	n	$tol_{\texttt{trunc}}$	Size of Galerkin system	Absolute residual	Relative residual
RAIL_20K	20 209	10^{-8}	224	$6.75 \cdot 10^{-14}$	$5.63 \cdot 10^{-15}$
		$\varepsilon_{\texttt{mach}}$	323	$6.37 \cdot 10^{-14}$	$5.31 \cdot 10^{-15}$
RAIL_79K	79 841	10^{-8}	254	$6.13 \cdot 10^{-14}$	$5.11 \cdot 10^{-15}$
		$\varepsilon_{\texttt{mach}}$	353	$5.90 \cdot 10^{-14}$	$4.92 \cdot 10^{-15}$
CONV_DIFF_160K	160 000	10^{-8}	48	$2.16 \cdot 10^{-9}$	$1.93 \cdot 10^{-14}$
		$\varepsilon_{\texttt{mach}}$	79	$2.16 \cdot 10^{-9}$	$1.93 \cdot 10^{-14}$
CONV_DIFF_1M	1 000 000	10^{-8}	52	$1.94 \cdot 10^{-8}$	$2.77 \cdot 10^{-14}$
		$\varepsilon_{\texttt{mach}}$	82	$1.93 \cdot 10^{-8}$	$2.76 \cdot 10^{-14}$
CONV_DIFF_9M	9 000 000	10^{-8}	57	$3.72 \cdot 10^{-7}$	$5.91 \cdot 10^{-14}$
		$\varepsilon_{\texttt{mach}}$	101	$3.02 \cdot 10^{-7}$	$4.80 \cdot 10^{-14}$
COOKIE_425K	425 272	10^{-8}	156	$5.05 \cdot 10^{-14}$	$6.37 \cdot 10^{-10}$
		$\varepsilon_{\texttt{mach}}$	238	$5.04 \cdot 10^{-14}$	$6.37 \cdot 10^{-10}$
COOKIE_1185K	1 185 586	10^{-8}	163	$1.53 \cdot 10^{-13}$	$5.41 \cdot 10^{-9}$
		$\varepsilon_{\texttt{mach}}$	298	$1.53 \cdot 10^{-13}$	$5.42 \cdot 10^{-9}$
COOKIE_2656K	2 656 643	10^{-8}	168	$4.51 \cdot 10^{-13}$	$3.58 \cdot 10^{-8}$
		$\varepsilon_{\texttt{mach}}$	306	$1.04 \cdot 10^{-13}$	$8.28 \cdot 10^{-9}$
CHIP	20 082	10^{-8}	103	$6.81 \cdot 10^{-10}$	$6.81 \cdot 10^{-10}$
		$\varepsilon_{\texttt{mach}}$	198	$1.79 \cdot 10^{-12}$	$1.79 \cdot 10^{-12}$

Table 5: Residuals for the ARE $0_n = A^\top XM + M^\top XA - M^\top XBB^\top XM + C^\top C$ and the size of the Galerkin system.

Instance	n	$tol_{\texttt{trunc}}$	$t_{\texttt{ARE}}$	$t_{\texttt{svd}}$	$t_{\texttt{gal}}$	$t_{\texttt{expm}}$	$t_{\texttt{total}}$
RAIL_20K	20 209	10^{-8}	3.14	0.40	0.16	$5.31 \cdot 10^{-2}$	**3.75**
		$\varepsilon_{\texttt{mach}}$	2.77	0.41	0.22	$6.80 \cdot 10^{-2}$	**3.47**
RAIL_79K	79 841	10^{-8}	11.20	1.82	0.67	$3.12 \cdot 10^{-2}$	**13.72**
		$\varepsilon_{\texttt{mach}}$	11.72	1.85	0.92	$7.20 \cdot 10^{-2}$	**14.56**
CONV_DIFF_160K	160 000	10^{-8}	22.08	0.79	0.08	$3.69 \cdot 10^{-3}$	**22.96**
		$\varepsilon_{\texttt{mach}}$	22.27	0.79	0.14	$5.97 \cdot 10^{-3}$	**23.21**
CONV_DIFF_1M	1 000 000	10^{-8}	160.00	5.19	0.49	$6.49 \cdot 10^{-3}$	**165.68**
		$\varepsilon_{\texttt{mach}}$	159.45	5.26	0.85	$1.87 \cdot 10^{-2}$	**165.58**
CONV_DIFF_9M	9 000 000	10^{-8}	2732.65	43.23	5.10	$3.00 \cdot 10^{-2}$	**2781.01**
		$\varepsilon_{\texttt{mach}}$	2745.94	52.55	9.18	$1.67 \cdot 10^{-2}$	**2807.68**
COOKIE_425K	425 272	10^{-8}	192.21	6.90	10.58	$4.24 \cdot 10^{-2}$	**209.73**
		$\varepsilon_{\texttt{mach}}$	191.63	6.80	13.93	$5.30 \cdot 10^{-2}$	**212.40**
COOKIE_1185K	1 185 586	10^{-8}	868.96	25.60	41.42	$3.54 \cdot 10^{-2}$	**936.02**
		$\varepsilon_{\texttt{mach}}$	866.70	26.83	56.95	$4.92 \cdot 10^{-2}$	**950.53**
COOKIE_2656K	2 656 643	10^{-8}	2374.40	54.99	111.64	$4.17 \cdot 10^{-2}$	**2541.08**
		$\varepsilon_{\texttt{mach}}$	2350.73	60.22	153.44	$5.07 \cdot 10^{-2}$	**2564.44**
CHIP	20 082	10^{-8}	5.58	0.29	0.08	$9.35 \cdot 10^{-3}$	**5.96**
		$\varepsilon_{\texttt{mach}}$	5.35	0.21	0.08	$2.21 \cdot 10^{-2}$	**5.66**

Table 6: Timings for the large-scale examples.

6. Conclusions

We have reviewed, and extended fundamental properties of the solution to the DRE and ARE and heavily relied on the solution representation provided by *Radon's Lemma* to analyze variants of *Davison-Maki methods* and to derive an efficient Galerkin projection scheme. Numerical tests confirmed that the resulting projected scheme outperforms splitting methods in terms of computation time, memory requirements, and approximation quality. In particular, storage requirements have been the bottleneck in the numerical considerations of large-scale DREs.

Our proposed Galerkin method bases on a low-rank approximation of the associated ARE for which efficient solvers exist. Moreover, the information on the residual and on the eigenvalue decay that come with the low-rank iteration for the ARE, can be directly transferred into estimates for the approximation quality

of our approach. Future work will deal with error analysis of the Galerkin approximation. Moreover, the stability and rounding error analysis for the *modified Davison-Maki method* is an open question.

Acknowledgments.

We thank Prof. Dr. Valeria Simoncini for a helpful discussion on a previous version of this manuscript.

References

[1] A. Locatelli, Optimal Control: An Introduction, Birkhäuser, Basel, Switzerland, 2001.

[2] H. W. Knobloch, H. Kwakernaak, Lineare Kontrolltheorie, Springer-Verlag, Berlin, 1985. doi:10.1007/978-3-642-69884-2.

[3] R. A. Brockett, Finite Dimensional Linear Systems, Wiley, New York, 1970. doi:10.1137/1.9781611973884.

[4] P. Benner, H. Mena, Numerical solution of the infinite-dimensional LQR-problem and the associated differential Riccati equations, J. Numer. Math. 26 (1) (2018) 1–20. doi:10.1515/jnma-2016-1039.

[5] T. McCauley, Computing the Maslov index from singularities of a matrix Riccati equation, J. Dyn. Diff. Equat. 29 (4) (2017) 1487–1502. doi:10.1007/s10884-016-9568-9.

[6] M. Beck, S. J. A. Malham, Computing the Maslov index for large systems, Proc. Amer. Math. Soc. 143 (5) (2015) 2159–2173. doi:10.1090/S0002-9939-2014-12575-5.

[7] P. Benner, H. Mena, BDF methods for large-scale differential Riccati equations, in: B. De Moor, B. Motmans, J. Willems, P. Van Dooren, V. Blondel (Eds.), Proc. 16th Intl. Symp. Mathematical Theory of Network and Systems, MTNS 2004, 2004, pp. 1–12.

[8] P. Benner, H. Mena, Rosenbrock methods for solving Riccati differential equations, IEEE Trans. Autom. Control 58 (11) (2013) 2950–2957. doi:10.1109/TAC.2013.2258495.

[9] J. Heiland, Decoupling and optimization of differential-algebraic equations with application in flow control, Dissertation, TU Berlin (2014). doi:10.14279/depositonce-4069.

[10] M. Köhler, N. Lang, J. Saak, Solving differential matrix equations using Parareal, Proc. Appl. Math. Mech. 16 (1) (2016) 847–848. doi:10.1002/pamm.201610412.

[11] N. Lang, H. Mena, J. Saak, On the benefits of the $LDL^\top$ factorization for large-scale differential matrix equation solvers, Linear Algebra Appl. 480 (2015) 44–71. doi:10.1016/j.laa.2015.04.006.

[12] N. Lang, Numerical methods for large-scale linear time-varying control systems and related differential matrix equations, Dissertation, Technische Universität Chemnitz, Germany (2017).

[13] T. Stillfjord, Low-rank second-order splitting of large-scale differential Riccati equations, IEEE Trans. Autom. Control 60 (10) (2015) 2791–2796. doi:10.1109/TAC.2015.2398889.

[14] T. Stillfjord, Singular value decay of operator-valued differential Lyapunov and Riccati equations, SIAM J. Control Optim. 56 (2018) 3598–3618. doi:10.1137/18M1178815.

[15] H. Mena, L.-M. Pfurtscheller, T. Stillfjord, GPU acceleration of splitting schemes applied to differential matrix equations, Numer. Algorithms 83 (1) (2019) 395–419. doi:10.1007/s11075-019-00687-w.

[16] Y. Güldoğan, M. Hached, K. Jbilou, M. Kurulay, Low-rank approximate solutions to large-scale differential matrix Riccati equations, Applicationes Mathematicae 45 (2) (2018) 233–254. doi:10.4064/am2355-1-2018.

[17] M. Hached, K. Jbilou, Numerical solutions to large-scale differential Lyapunov matrix equations, Numer. Algorithms 79 (3) (2018) 741–757. doi:10.1007/s11075-017-0458-y.

[18] M. Hached, K. Jbilou, Computational Krylov-based methods for large-scale differential Sylvester matrix problems, Numer. Lin. Alg. Appl. 25 (5) (2018) e2187, 14. doi:10.1002/nla.2187.

[19] M. Hached, K. Jbilou, Numerical methods for differential linear matrix equations via Krylov subspace methods, J. Comput. Appl. Math. 370 (2020) 112674. doi:10.1016/j.cam.2019.112674.

[20] A. Koskela, H. Mena, Analysis of Krylov Subspace Approximation to Large Scale Differential Riccati Equations, e-print arXiv:1705.07507v4, math.NA (2018).

[21] G. Kirsten, V. Simoncini, Order reduction methods for solving large-scale differential matrix Riccati equations, SIAM J. Sci. Comput. 42 (4) (2020) A2182–A2205. doi:10.1137/19M1264217.

[22] V. Angelova, M. Hached, K. Jbilou, Approximate solutions to large nonsymmetric differential Riccati problems with applications to transport theory, Numer. Lin. Alg. Appl. 27 (1) (2020) e2272. doi:10.1002/nla.2272.

[23] M. Behr, P. Benner, J. Heiland, Solution formulas for differential Sylvester and Lyapunov equations, Calcolo 56 (4) (2019) 51. doi:10.1007/s10092-019-0348-x.

[24] R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1985. doi:10.1017/CBO9780511810817.

[25] H. Abou-Kandil, G. Freiling, V. Ionescu, G. Jank, Matrix Riccati Equations in Control and Systems Theory, Birkhäuser, Basel, Switzerland, 2003. doi:10.1007/978-3-0348-8081-7.

[26] M. Behr, P. Benner, J. Heiland, On an Invariance Principle for the Solution Space of the Differential Riccati Equation, Proc. Appl. Math. Mech. 18 (1) (2018) e201800031. doi:10.1002/pamm.201800031.

[27] L. Amodei, J.-M. Buchot, An invariant subspace method for large-scale algebraic Riccati equation, Appl. Numer. Math. 60 (11) (2010) 1067–1082. doi:10.1016/j.apnum.2009.09.006.

[28] P. Benner, Z. Bujanović, On the solution of large-scale algebraic Riccati equations by using low-dimensional invariant subspaces, Linear Algebra Appl. 488 (2016) 430–459. doi:10.1016/j.laa.2015.09.027.

[29] B. Beckermann, A. Townsend, On the singular values of matrices with displacement structure, SIAM J. Matrix Anal. Appl. 38 (4) (2017) 1227–1248. doi:10.1137/16M1096426.

[30] J. Baker, M. Embree, J. Sabino, Fast singular value decay for Lyapunov solutions with nonnormal coefficients, SIAM J. Matrix Anal. Appl. 36 (2) (2015) 656–668. doi:10.1137/140993867.

[31] A. C. Antoulas, D. C. Sorensen, Y. Zhou, On the decay rate of Hankel singular values and related issues, Syst. Cont. Lett. 46 (5) (2002) 323–342. doi:10.1016/S0167-6911(02)00147-0.

[32] L. Grubišić, D. Kressner, On the eigenvalue decay of solutions to operator Lyapunov equations, Syst. Cont. Lett. 73 (2014) 42–47. doi:10.1016/j.sysconle.2014.09.006.

[33] T. Penzl, Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case, Syst. Cont. Lett. 40 (2000) 139–144. doi:10.1016/S0167-6911(00)00010-4.

[34] M. Opmeer, Decay of singular values of the Gramians of infinite-dimensional systems, in: Proceedings 2015 European Control Conference (ECC), IEEE, Linz, Austria, 2015, pp. 1183–1188. doi:10.1109/ECC.2015.7330700.

[35] L. Grasedyck, Existence of a low rank or $\mathcal{H}$-matrix approximant to the solution of a Sylvester equation, Numer. Lin. Alg. Appl. 11 (4) (2004) 371–389. doi:10.1002/nla.366.

[36] D. C. Sorensen, Y. Zhou, Bounds on eigenvalue decay rates and sensitivity of solutions to Lyapunov equations, Tech. Rep. TR02-07, Dept. of Comp. Appl. Math., Rice University, Houston, TX (2002).

[37] W. Walter, Ordinary differential equations, Vol. 182 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1998. doi:10.1007/978-1-4612-0601-9.

[38] I. Rusnak, Almost analytic representation for the solution of the differential matrix Riccati equation, IEEE Trans. Autom. Control 33 (2) (1988) 191–193. doi:10.1109/9.388.

[39] F. M. Callier, J. Winkin, J. L. Willems, Convergence of the time-invariant Riccati differential equation and LQ-problem: mechanisms of attraction, Internat. J. Control 59 (4) (1994) 983–1000. doi:10.1080/00207179408923113.

[40] V. Radisavljevic, Improved Potter-Anderson-Moore algorithm for the differential Riccati equation, Appl. Math. Comput. 218 (8) (2011) 4641–4646. doi:10.1016/j.amc.2011.09.007.

[41] B. D. O. Anderson, J. B. Moore, Linear Optimal Control, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[42] E. J. Davison, M. C. Maki, The numerical solution of the matrix Riccati differential equation, IEEE Trans. Autom. Control 18 (1973) 71–73. doi:10.1109/tac.1973.1100210.

[43] C. Kenney, R. B. Leipnik, Numerical integration of the differential matrix Riccati equation, IEEE Trans. Autom. Control 30 (1985) 962–970. doi:10.1109/tac.1985.1103822.

[44] 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. doi:10.1109/SECON.1990.117906.

[45] D. R. Vaughan, A negative exponential solution for the matrix Riccati equation, IEEE Trans. Autom. Control 14 (1969) 72–75. doi:10.1109/tac.1969.1099117.

[46] A. J. Laub, Schur techniques for Riccati differential equations, in: D. Hinrichsen, A. Isidori (Eds.), Feedback Control of Linear and Nonlinear Systems, Springer-Verlag, New York, 1982, pp. 165–174. doi:10.1007/bfb0006827.

[47] K. R. Meyer, D. C. Offin, Introduction to Hamiltonian Dynamical Systems and the N-Body Problem, 3rd Edition, Vol. 90 of Applied Mathematical Sciences, Springer, Cham, 2017. doi:10.1007/978-3-319-53691-0.

[48] R. E. Kalman, T. S. Englar, A user's manual for the automatic synthesis program, RIAS Report CR-475, NASA (1966).

[49] P. Benner, J. Saak, A semi-discretized heat transfer model for optimal cooling of steel profiles, in: P. Benner, V. Mehrmann, D. Sorensen (Eds.), Dimension Reduction of Large-Scale Systems, Vol. 45 of Lect. Notes Comput. Sci. Eng., Springer-Verlag, Berlin/Heidelberg, Germany, 2005, pp. 353–356. doi:10.1007/3-540-27909-1_19.

[50] T. Penzl, LYAPACK Users Guide, Tech. Rep. SFB393/00-33, Sonderforschungsbereich 393 *Numerische Simulation auf massiv parallelen Rechnern*, TU Chemnitz, 09107 Chemnitz, Germany (2000).
URL http://www.tu-chemnitz.de/sfb393/sfb00pr.html

[51] The MORwiki Community, Convection, MORwiki – Model Order Reduction Wiki (20XX).
URL http://modelreduction.org/index.php/Convection

[52] S. Rave, J. Saak, Thermal block, MORwiki – Model Order Reduction Wiki (2020).
URL http://modelreduction.org/index.php/Thermal_Block

[53] T. Stillfjord, Adaptive high-order splitting schemes for large-scale differential Riccati equations, Numer. Algorithms 78 (2018) 1129–1151. doi:10.1007/s11075-017-0416-8.

Appendix A. Numerical Results for Galerkin Approach

RAIL, $n = 5177$, $M^\top \dot{X}(t) M = A^\top X(t) M + M^\top X(t) A - M^\top X(t) B B^\top X(t) M + C^\top C,\ X(0) = 0_n$.

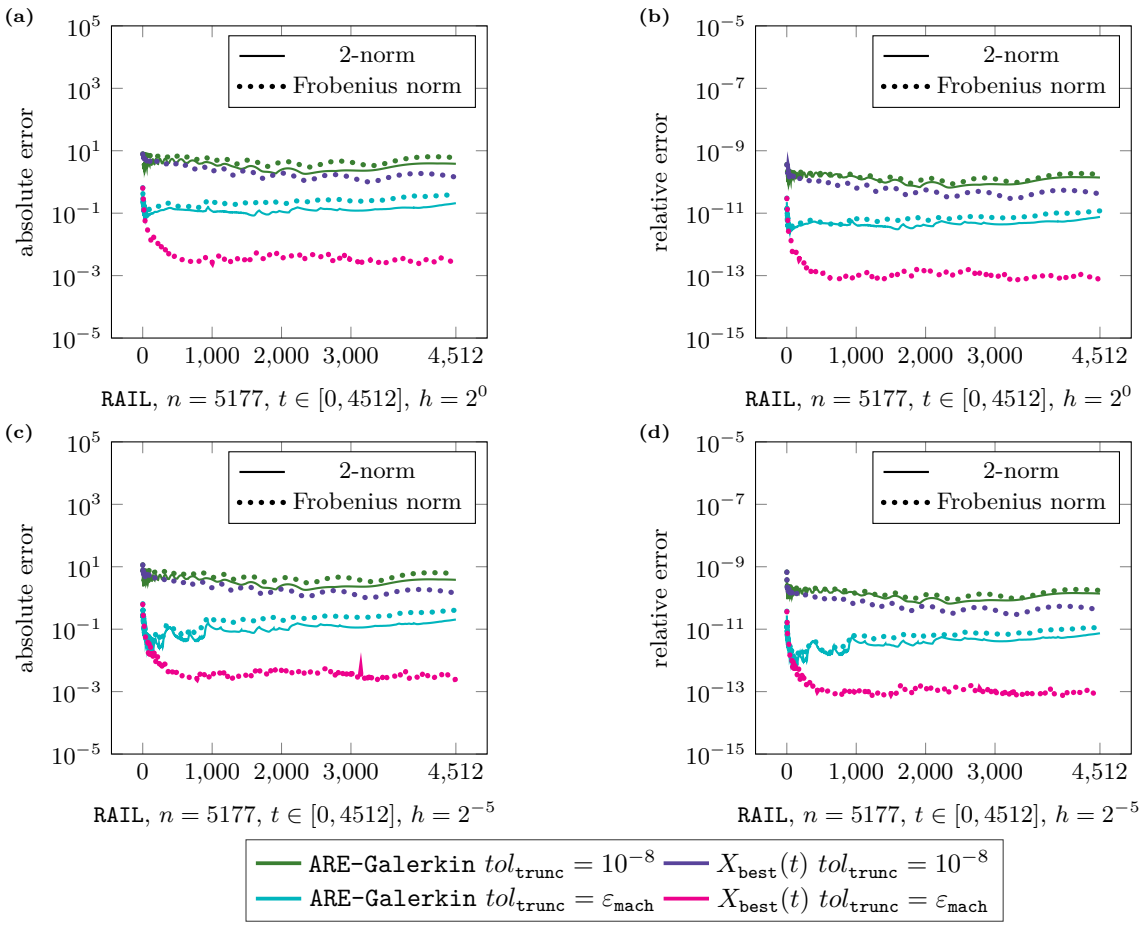


Fig. A.7. **(a), (c)** Absolute error of the `ARE-Galerkin` and Best approximation. **(b), (d)** Relative error of the `ARE-Galerkin` and Best approximation.

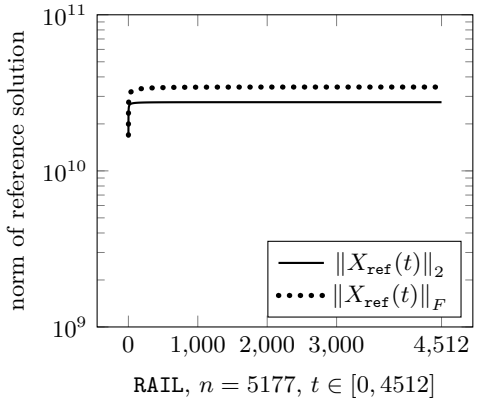


Fig. A.8. Norm of the reference solution.

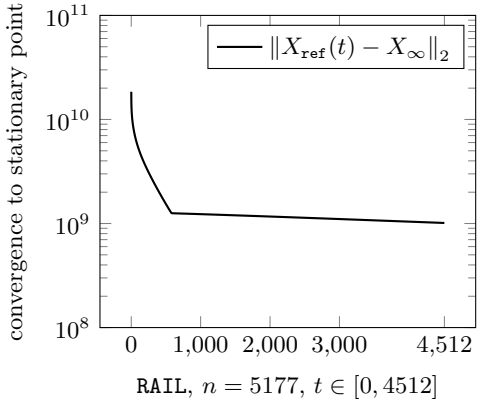


Fig. A.9. Convergence to the stationary point.

$\texttt{CONV\_DIFF}$, $n = 6400$, $\dot{X}(t) = A^\top X(t) + X(t)A - X(t)BB^\top X(t) + C^\top C,\ X(0) = 0_n$.

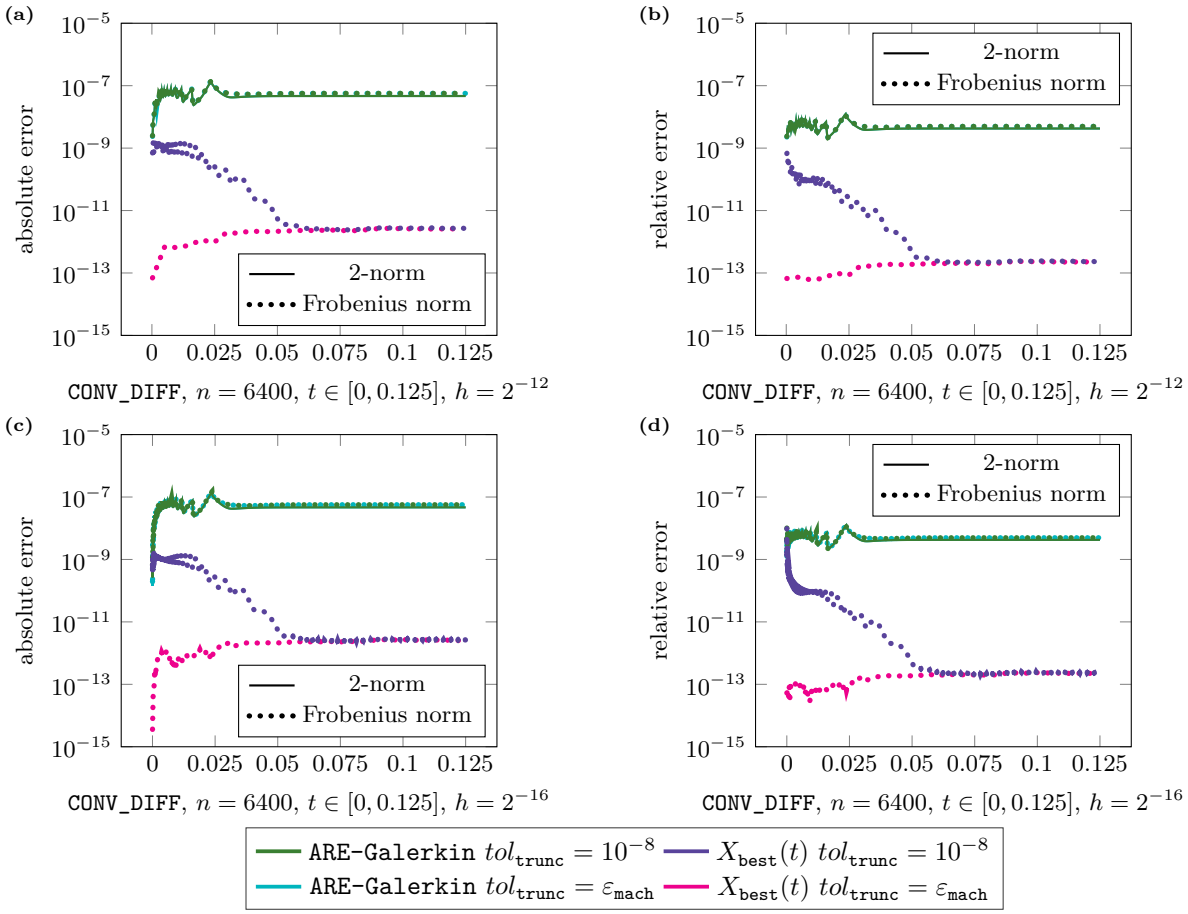


Fig. A.10. **(a), (c)** Absolute error of the `ARE-Galerkin` and Best approximation. **(b), (d)** Relative error of the `ARE-Galerkin` and Best approximation.

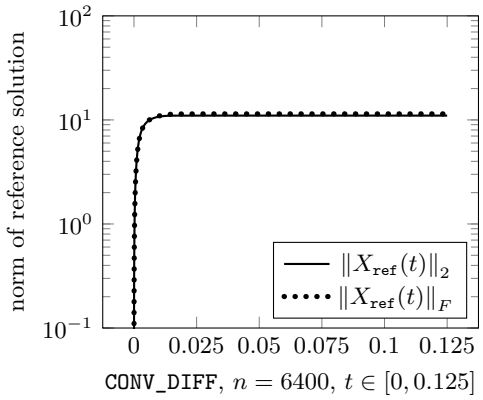


Fig. A.11. Norm of the reference solution.

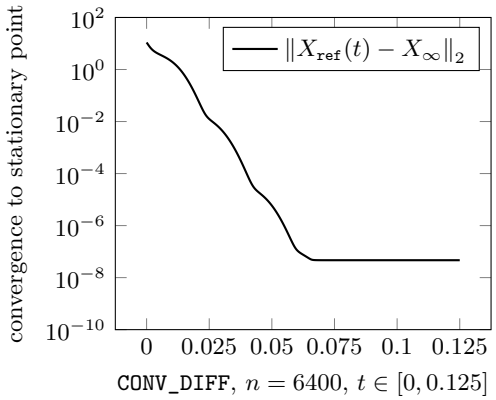


Fig. A.12. Convergence to the stationary point.

$\texttt{FLOW},\ n = 9669,\ M^\top \dot{X}(t) M = A^\top X(t) M + M^\top X(t) A - M^\top X(t) B B^\top X(t) M + C^\top C,\ X(0) = 0_n.$

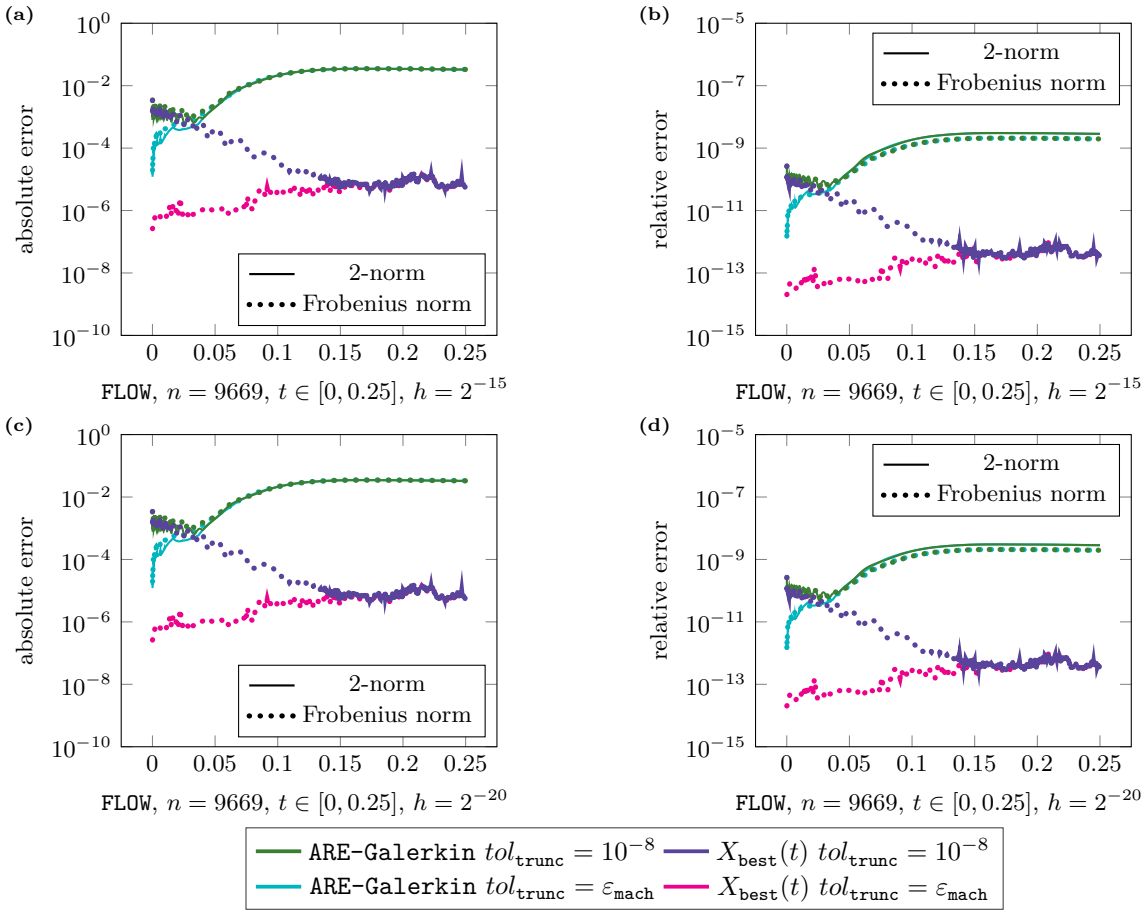


Fig. A.13. **(a), (c)** Absolute error of the `ARE-Galerkin` and Best approximation. **(b), (d)** Relative error of the `ARE-Galerkin` and Best approximation.

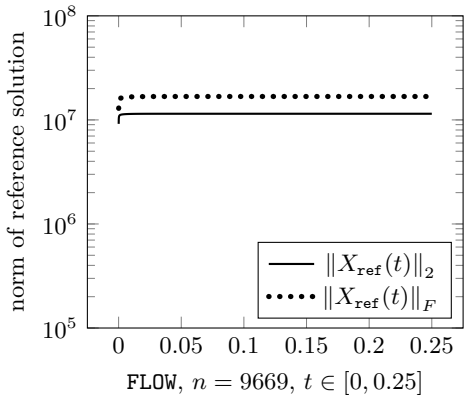


Fig. A.14. Norm of the reference solution.

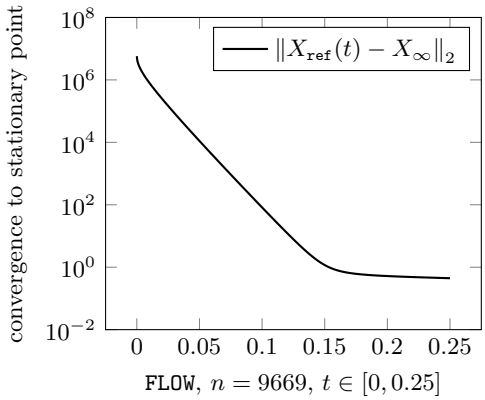


Fig. A.15. Convergence to the stationary point.

COOKIE, $n = 7488$, $M^\top \dot{X}(t) M = A^\top X(t) M + M^\top X(t) A - M^\top X(t) B B^\top X(t) M + C^\top C$, $X(0) = 0_n$.

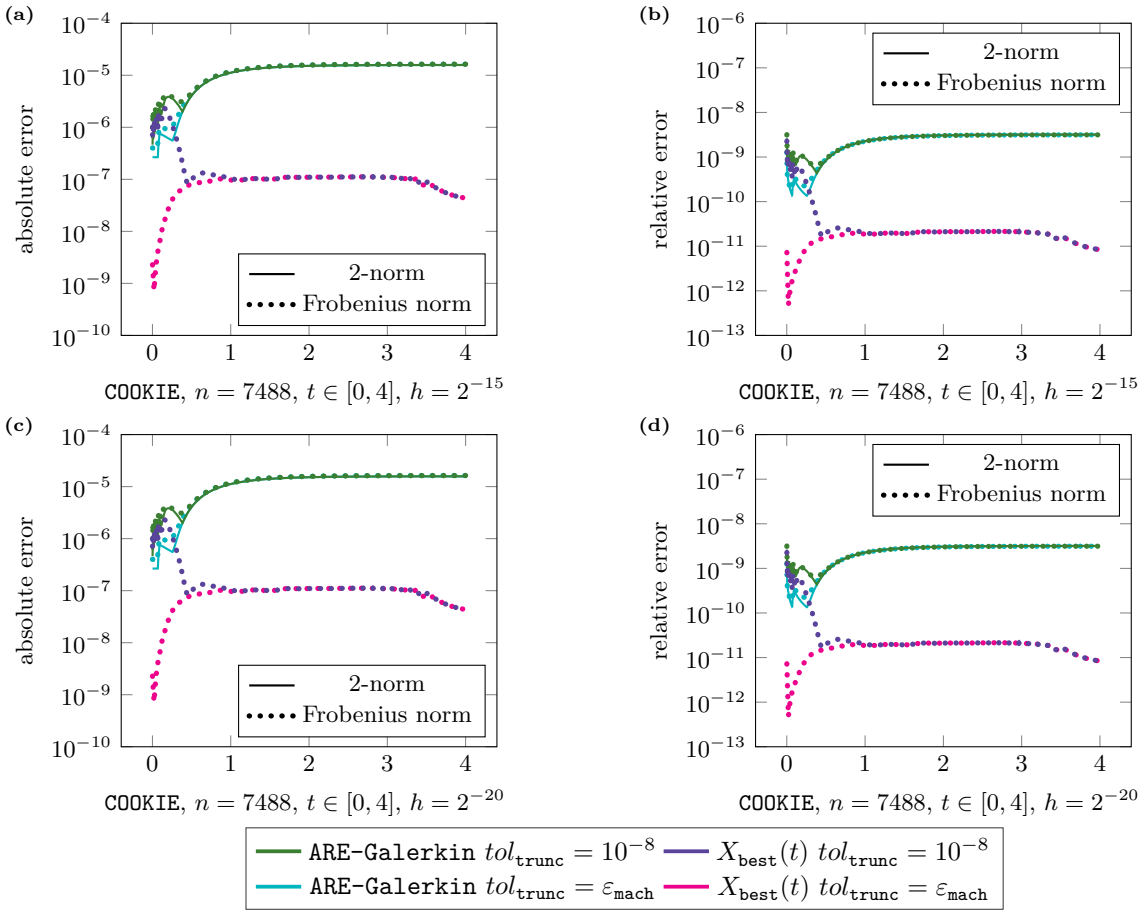


Fig. A.16. **(a), (c)** Absolute error of the ARE-Galerkin and Best approximation. **(b), (d)** Relative error of the ARE-Galerkin and Best approximation.

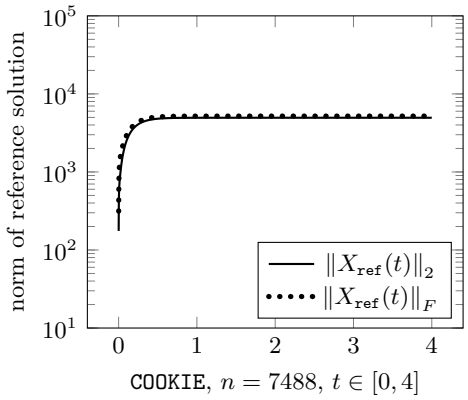


Fig. A.17. Norm of the reference solution.

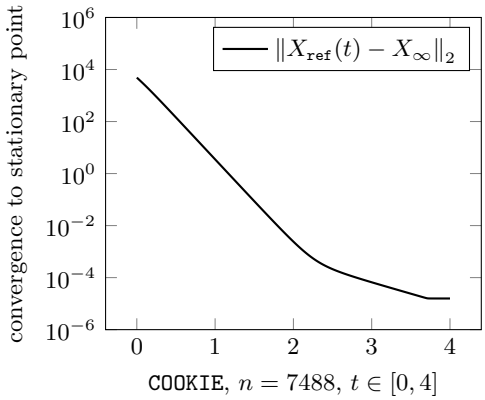


Fig. A.18. Convergence to the stationary point.

Appendix B. Numerical Results for Splitting Schemes

RAIL, $n = 5177$, $M^\top \dot{X}(t)M = A^\top X(t)M + M^\top X(t)A - M^\top X(t)BB^\top X(t)M + C^\top C,\ X(0) = 0_n$.

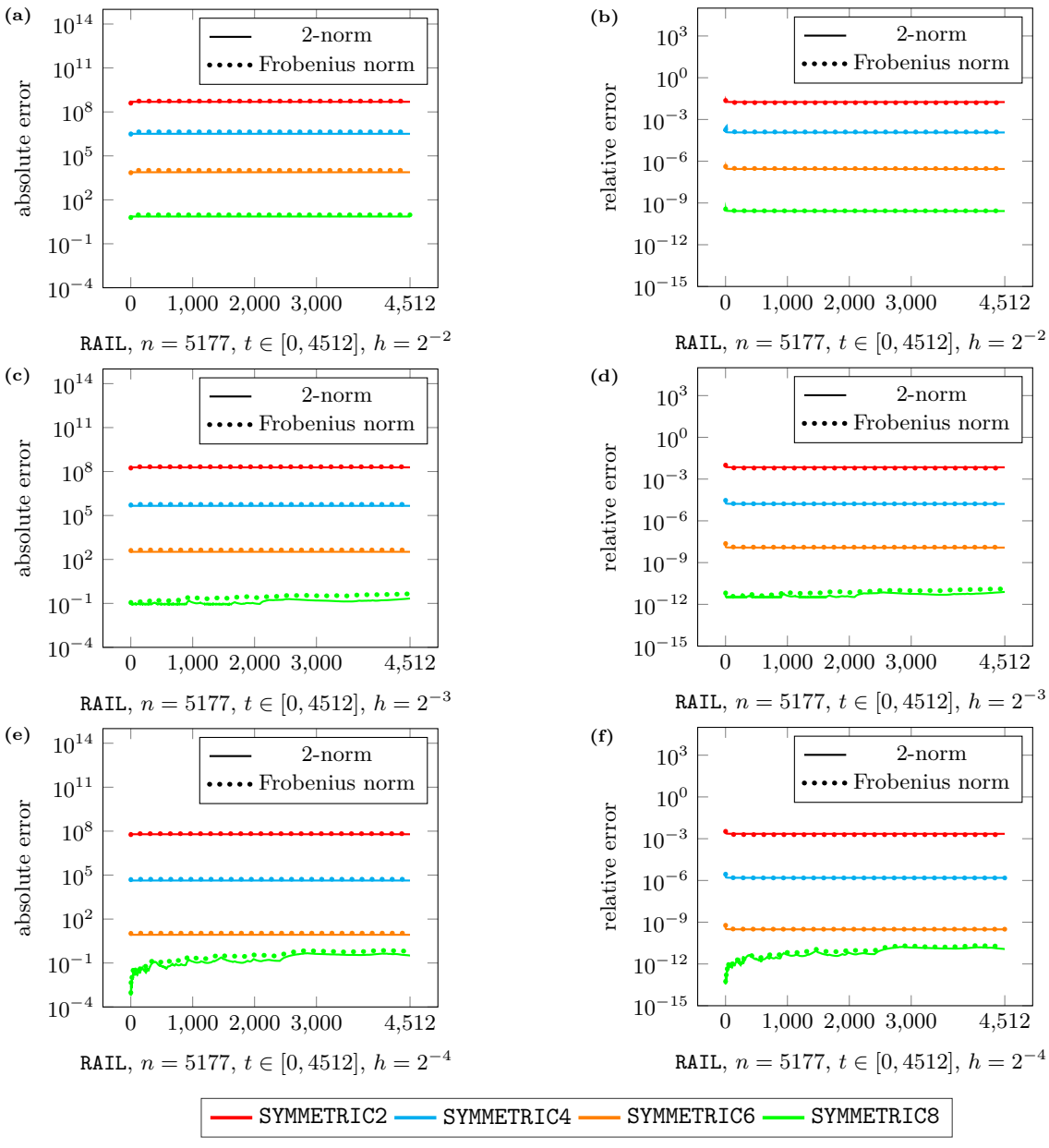


Fig. B.19. **(a), (c), (e)** Absolute error of the splitting scheme approximation. **(b), (d), (f)** Relative error of the splitting scheme approximation.

CONV_DIFF, $n = 6400$, $\dot{X}(t) = A^\top X(t) + X(t)A - X(t)BB^\top X(t) + C^\top C,\ X(0) = 0_n$.

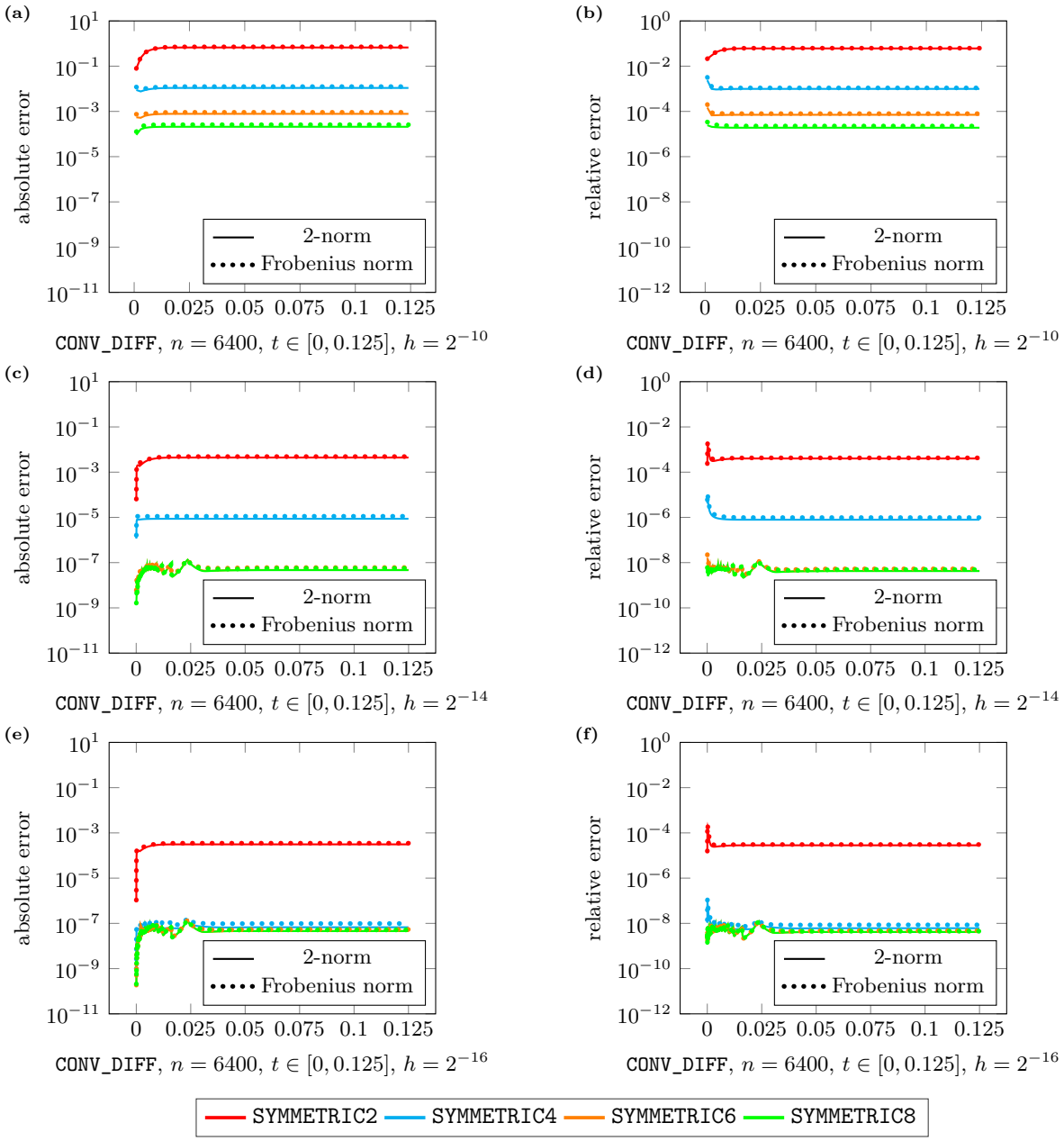


Fig. B.20. **(a), (c), (e)** Absolute error of the splitting scheme approximation. **(b), (d), (f)** Relative error of the splitting scheme approximation.

FLOW, $n = 9669$, $M^\top \dot{X}(t) M = A^\top X(t) M + M^\top X(t) A - M^\top X(t) B B^\top X(t) M + C^\top C$, $X(0) = 0_n$.

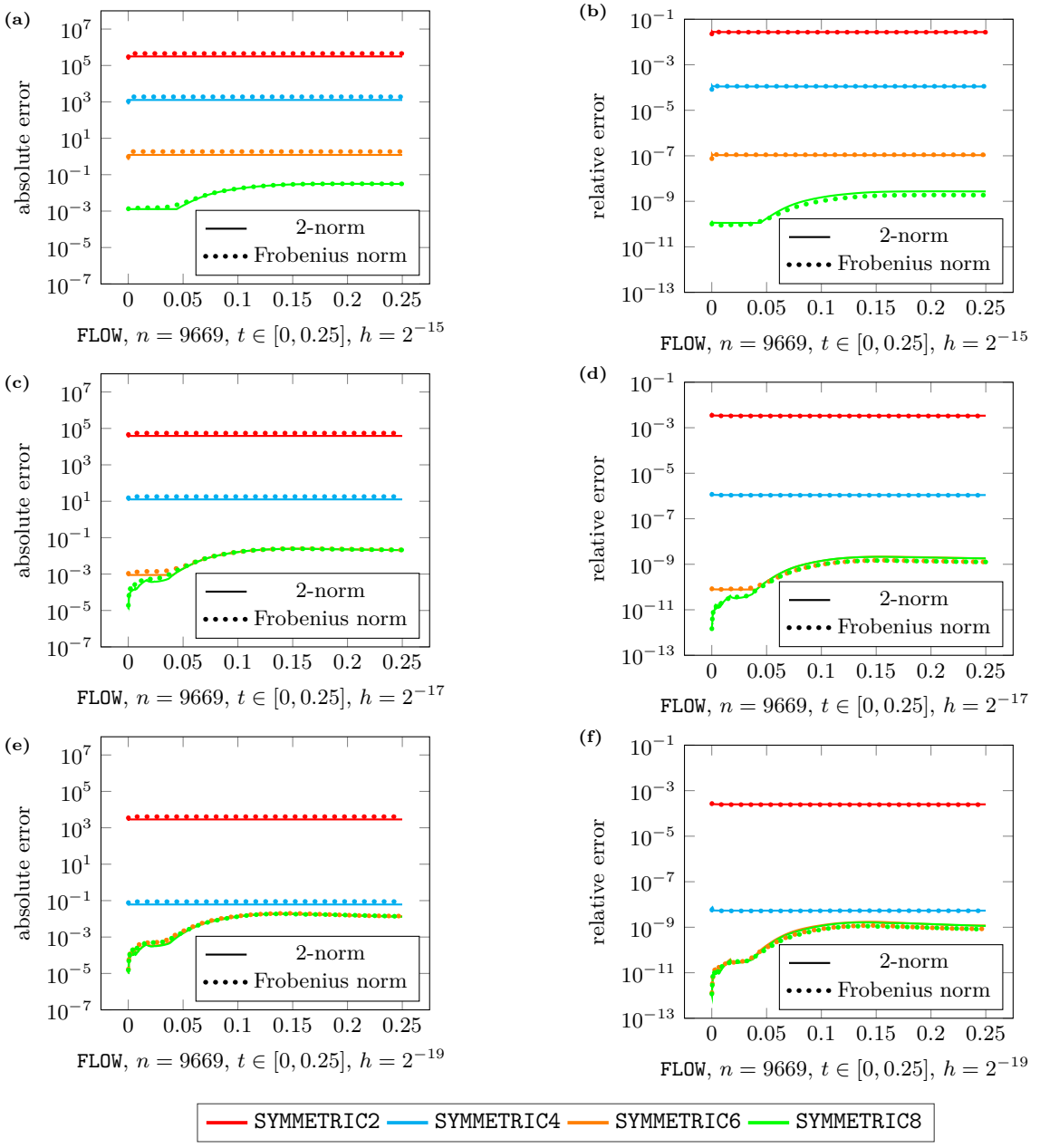


Fig. B.21. **(a), (c), (e)** Absolute error of the splitting scheme approximation. **(b), (d), (f)** Relative error of the splitting scheme approximation.

COOKIE, $n = 7488$, $M^\top \dot{X}(t) M = A^\top X(t) M + M^\top X(t) A - M^\top X(t) B B^\top X(t) M + C^\top C,\ X(0) = 0_n$.

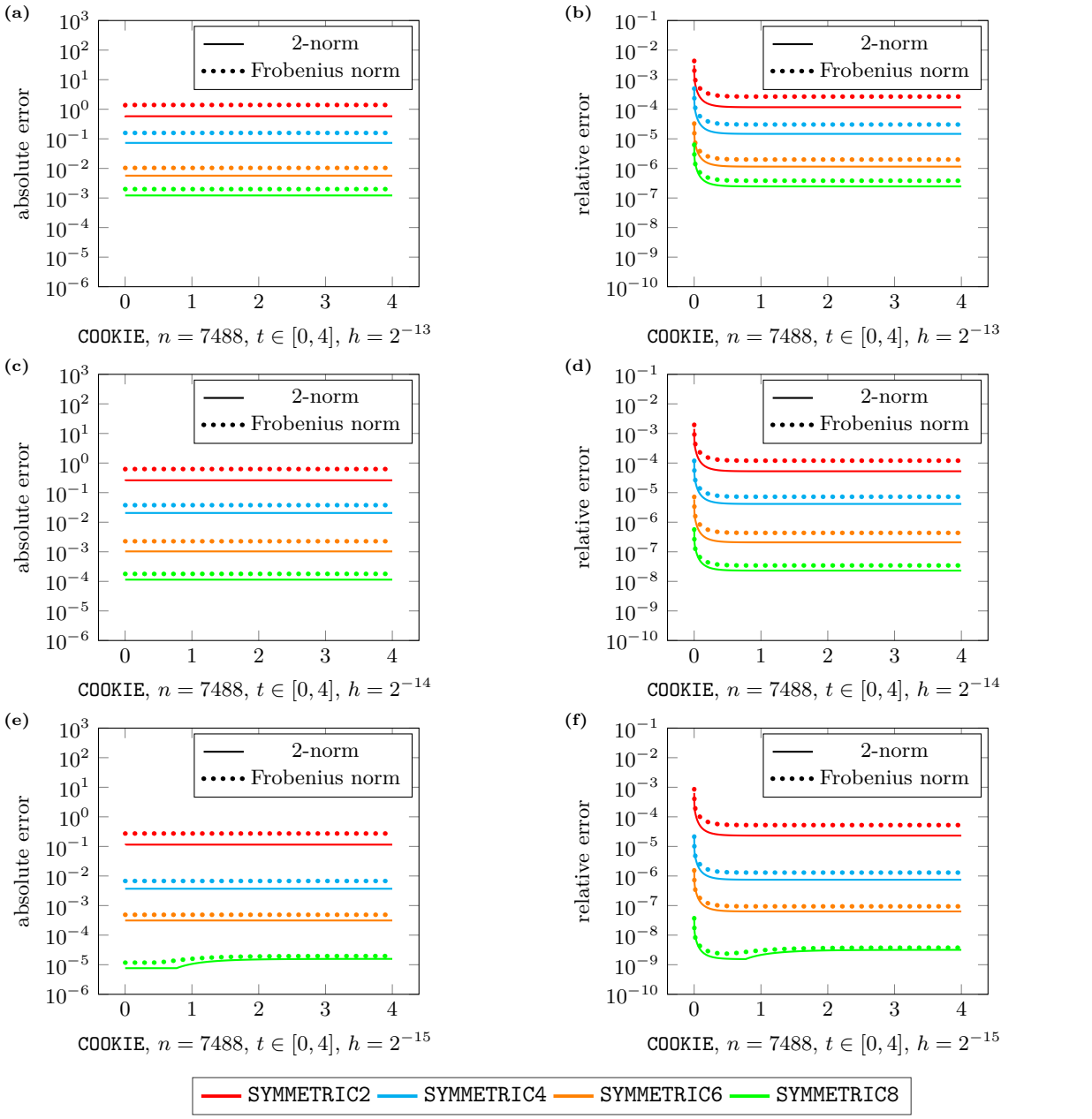


Fig. B.22. **(a), (c), (e)** Absolute error of the splitting scheme approximation. **(b), (d), (f)** Relative error of the splitting scheme approximation.