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

Maximilian Behr Peter Benner Jan Heiland

October 30, 2019

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. In view of memory efficient approximation, we review and extend known solution formulas and identify invariant subspaces for a possibly low-dimensional solution representation. Based on these theoretical findings, we propose a Galerkin projection onto a space related to a low-rank approximation of the algebraic Riccati equation. For the numerical implementation, we provide an alternative interpretation of the modified *Davison-Maki method* via the transformed flow of the differential Riccati equation, which enables us to rule out known stability issues of the method in combination with the proposed projection scheme. We present numerical experiments for large-scale autonomous differential Riccati equations and compare our approach with high-order splitting schemes.

Contents

1. Introduction

In this paper we consider the autonomous differential Riccati equation

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

The equation plays an important role in model order reduction, optimal control, differential games and stability analysis [1, 7, 15, 17, 32, 38, 40]. We focus in this work on the large-scale case. In this setting, 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 $N_t n^2$, where n is the dimension of the problem and N_t is the number of time steps. The approach of first discretizing in time and then focusing on efficient approximation of the resulting algebraic equations has been the main course of research on this problem setup, see, e.g., [13–15, 27, 33, 35, 36, 41, 42, 52, 54]. In all these approaches, the approximation of at least one large-scale algebraic equation has to be solved and stored for every time step so that the memory demands still scale with $N_t n$. 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 in the order of $N_t k$. In this respect, Krylov subspace methods have been proposed [4, 23–26, 31, 34] 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 demand 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 differential Lyapunov equations [10].

The paper is organized as follows. In Section 3 we introduce the algebraic and differential Riccati equations and review the relevant fundamental properties about their solutions. In Section 4 we review *Radon's Lemma* and work out its implication that the differential Riccati equation is connected to a flow on a Grassmanian manifold. Moreover, in Section 4.2, 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 4.3 Then we will use the connection to the Grassmanian manifold to derive the *modified Davison-Maki method* in a way that overcomes these instabilities. In Section 5, 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 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 6 and Appendices A and B.

2. Preliminaries

In this section we set the notation and review some basic results from linear algebra. The identity matrix and zero matrix of size n are written by I_n and 0. 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 quadratic 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 non-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, [28, Ch. 7.7]. 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.

3. Algebraic and Differential Riccati Equations

In this section we introduce the algebraic and differential Riccati equation (ARE/DRE) and the algebraic Lyapunov equation (ALE).
Consider $A, X_0 \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{c\times n}$ and $B \in \mathbb{R}^{n\times b}$. Throughout this paper, we assume that X_0 is a symmetric positive semidefinite matrix and consider the DRE

$$\dot{X}(t) = \mathcal{R}(X(t)) := A^T X(t) + X(t)A - X(t)BB^T X(t) + C^T C, \tag{1a}$$
$$X(0) = X_0. \tag{1b}$$

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

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

The linear version ($B = 0$) of the ARE is the ALE

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

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

Theorem 3.1 (Existence and Uniqueness of Solutions to the ALE (3), [1, Thm. 1.1.3, 1.1.7])**.**
If $\Lambda(A) \cap \Lambda(-A) = \emptyset$, then the ALE (3) *admits 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*

$$X_L = \int_0^\infty e^{tA^T} C^T C e^{tA} \mathrm{d}t. \tag{4}$$

Theorem 3.2 (Existence and Uniqueness of Solutions for the ARE (2), [1, Lem. 2.4.1, Cor. 2.4.3], [32, 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* $\mathcal{R}(X_\infty) = 0$ *and* $\Lambda(A - BB^TX_\infty) \subseteq \mathbb{C}^-$. *Moreover* X_∞ *is symmetric positive semidefinite and there is no other symmetric positive semidefinite solution of the* ARE (2).

Theorem 3.3 (Range of the Solution of the ARE (2), [9, 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:*

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

The inclusion $\mathcal{K}\left(A^T, C^T\right) \subseteq \operatorname{range}(X_\infty)$ in Theorem 3.3 is actually true for each symmetric solution of the ARE (2), cf. [1, Lemma 2.4.9]. In [2, Ch. 3.3] a Kalman decomposition is used to show that $\operatorname{rank}(X_\infty) = \dim\left(\mathcal{K}\left(A^T, C^T\right)\right)$. A connection between the space $\mathcal{K}\left(A^T, C^T\right)$ and a certain Krylov subspace generated by the associated Hamiltonian matrix, which can be used for numerical approximation of the solution of the ARE (2), was presented in [11, Thm. 10].
Typically, solutions of quadratic differential equations like the DRE (1) exhibit a finite-time escape phenomena. By means of comparison arguments and the fact that $-BB^T$ 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 ARE (2) and is, thus, bounded.

Theorem 3.4 (Existence and Uniqueness of Solutions of the DRE (1), [1, Thm. 4.1.6, 4.1.8], [32, Ch. 10])**.**
The DRE (1) *has a unique solution* $X\colon [0,\infty) \to \mathbb{R}^{n\times n}$. *The solution* X *has the following properties:*

- $X(t)$ *is symmetric positive semidefinite for all* $t \geq 0$.
- *If* $\dot{X}(0) = \mathcal{R}(X_0) \succcurlyeq 0$, *then* $t \mapsto X(t)$ *is monotonically non-decreasing 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 3.5 (Invariant Subspace of the Solution of the DRE (1), cp. [9, Thm. 3.1])**.**
Let the columns of $Q \in \mathbb{R}^{n\times p}$ *span an orthonormal basis of* $\mathcal{K}\left(A^T, C^T\right)$ *and define the linear space* $\mathcal{Q} := \left\{QYQ^T \mid Y \in \mathbb{R}^{p\times p}\right\} \subseteq \mathbb{R}^{n\times n}$ *or* $\mathcal{Q} := \{0\} \subseteq \mathbb{R}^{n\times n}$, *if* $C = 0$. *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* $X_0 = 0$.

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 for low-rank approximations, i.e. a $q \ll n$ so that the relation in Theorem 3.5 is still valid up to a given tolerance, to avoid overly demanding memory requirements. Therefore, the relevant literature features numerous contributions which study the decay rate of $\lambda_k^{\downarrow}(X)$ or $\frac{\lambda_k^{\downarrow}(X)}{\lambda_1^{\downarrow}(X)}$ for increasing k; see, e.g., [5, 6, 8, 21, 22, 44, 45, 51] on the eigenvalue decay of the solution of the ALE and [11, 44, 54] for results on the ARE and DRE.
For the autonomous DRE (1), one can derive estimates based on the monotonicity. Assume that $\mathcal{R}(X_0) \succcurlyeq 0$, then by Theorem 3.4 the function $t \mapsto X(t)$ is monotonically non-decreasing on $[0,\infty)$, where X is the unique solution of the DRE (1). A direct consequence of the Courant-Fischer-Weyl min-max principle [28, Cor. 7.7.4] implies that $t \mapsto \lambda_k^{\downarrow}(X(t))$ is also monotonically non-decreasing on $[0,\infty)$. Therefore the number of eigenvalues of $X(t)$ greater than or equal to a given threshold $\varepsilon > 0$ is non-decreasing over time.

Example 3.1 (Eigenvalue Decay)**.**
We illustrate this by an example in Figure 1. *We have chosen* $X_0 = 0$, $C = \begin{bmatrix} 1, \dots, 1 \end{bmatrix} = B^T$ *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 a 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 non-increasing order and plotted for* $t \in \{0.5, 1, \dots, 15\}$. *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.*

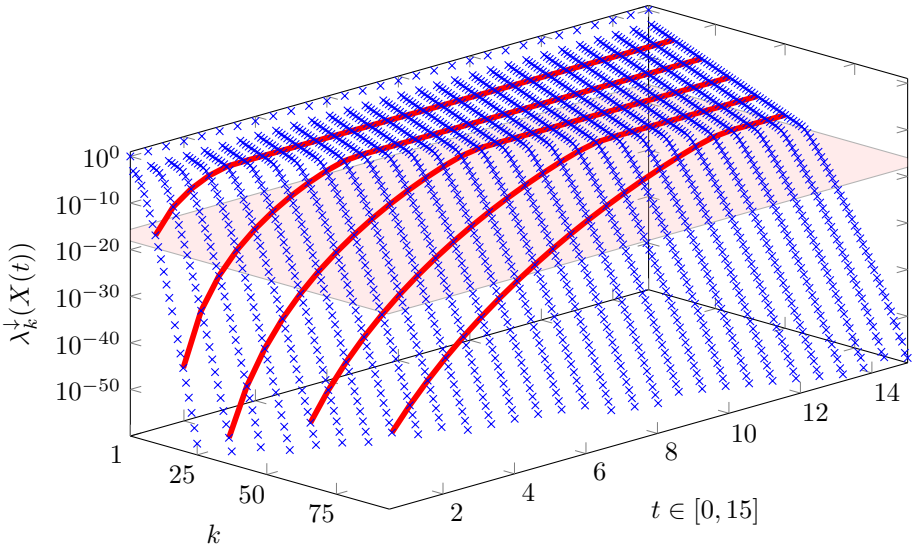


Fig. 1. Eigenvalues $\lambda_k^{\downarrow}(X(t))$ of the numerical solution of DRE (1).

4. Radon's Lemma

In this section, we consider the non-symmetric 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 (Thm. 4.1) has several consequences. In Section 4.1, we review the fact that the solution of the NDRE induces a flow on the Grassmanian manifold. This flow has a simpler structure as it is based on a matrix exponential. In Section 4.2 we show how solution formulas can be obtained by applying suitable linear transformations, which decouple the linear differential equation. Then, in view of numerical approximation, we review the *Davison-Maki method* and the *modified Davison-Maki method* in Section 4.3. We use the solution formula from Section 4.2 to explain, why the *Davison-Maki method* applied to the DRE usually suffers from numerical instabilities and show that an exploitation of the structure of the transformed flow on the Grassmanian manifold leads to a suitable modification of the *Davison-Maki method*.

Theorem 4.1 (Radon's Lemma, [1, 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

$$\dot{W}(t) = M_{22}W(t) - W(t)M_{11} - W(t)M_{12}(t)W(t) + M_{21}, \tag{5a}$$

$$W(0) = M_0. \tag{5b}$$

The following holds:

1. *Let $W\colon \mathbb{I} \to \mathbb{R}^{m\times n}$ be the solution of (5) and $U\colon \mathbb{I} \to \mathbb{R}^{n\times n}$ be the solution of the linear initial value problem*

$$\dot{U}(t) = (M_{11} + M_{12}W(t))\,U(t),\ U(0) = I_n. \tag{6}$$

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

$$\begin{bmatrix}\dot{U}(t)\\ \dot{V}(t)\end{bmatrix} = M\begin{bmatrix}U(t)\\ V(t)\end{bmatrix} := \begin{bmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{bmatrix}\begin{bmatrix}U(t)\\ V(t)\end{bmatrix},\ \begin{bmatrix}U(0)\\ V(0)\end{bmatrix} = \begin{bmatrix}I_n\\ M_0\end{bmatrix}. \tag{7}$$

2. *If $\begin{bmatrix}U\\ V\end{bmatrix}\colon \mathbb{I} \to \mathbb{R}^{(n+m)\times n}$ is a solution of (7) 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 (5).*

Radon's Lemma (Thm. 4.1) also holds for time-dependent continuous matrix valued functions as coefficients. Note that, usually, the solution of the NDRE (5) has finite time escape, while the solution of system (7) exists for all $t \in \mathbb{R}$. However, one can consider the solution W of the NDRE (5) on the interval of existence. As the function U is a solution of the linear initial value problem (6) 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. [57, §15]. Therefore as long as the solution of the NDRE (5) exists, it can be recovered from the solution of system (7).

4.1. Flow on the Grassmanian Manifold

In this section we review the fact that the solution of the NDRE (5) is locally equivalent to a flow on the Grassmanian manifold. This connection was first observed in [49] and the corresponding flow was further studied in [39, 50]. The content of this subsection is a summary of [50, §2]. One main observation from Radon's Lemma (Thm. 4.1) is that the solution W of the NDRE (5) depends only on the linear space spanned by $U(t)$ and $V(t)$. This can be seen by the following arguments. Let $\begin{bmatrix}U\\ V\end{bmatrix}\colon \mathbb{I} \to \mathbb{R}^{(n+m)\times n}$ be a solution of (7) and $t_0 \in \mathbb{I}$. Moreover assume that $\tilde{U} \in \mathbb{R}^{n\times n}, \tilde{V} \in \mathbb{R}^{m\times n}$ are such that

$$\operatorname{range}\left(\begin{bmatrix}\tilde{U}\\ \tilde{V}\end{bmatrix}\right) = \operatorname{range}\left(\begin{bmatrix}U(t_0)\\ V(t_0)\end{bmatrix}\right).$$

The linear spaces are equal, if and only if there is a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ such that

$$\begin{bmatrix}\tilde{U}\\ \tilde{V}\end{bmatrix} = \begin{bmatrix}U(t_0)\\ V(t_0)\end{bmatrix}T.$$

Since $U(t_0)$ is nonsingular, we have

$$\tilde{V}\tilde{U}^{-1} = V(t_0)TT^{-1}U(t_0)^{-1} = V(t_0)U(t_0)^{-1} = W(t_0).$$

Consequently, it is the linear subspace range $\left(\begin{bmatrix} U(t) \\ V(t) \end{bmatrix}\right) \subseteq \mathbb{R}^{n+m}$ that defines the solution $W(t)$, rather than the chosen basis $\begin{bmatrix} U(t) \\ V(t) \end{bmatrix}$ to represent the space. Since

$$\operatorname{range}\left(\begin{bmatrix} U(t) \\ V(t) \end{bmatrix}\right) = \operatorname{range}\left(e^{tM}\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right) = e^{tM}\operatorname{range}\left(\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right),$$

and the (nonsingular) matrix exponential is applied to an n-dimensional subspace range $\left(\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right)$, we obtain a time-dependent family of n-dimensional subspaces of $\mathbb{R}^{n+m}$. The Grassmanian manifold $G^n\left(\mathbb{R}^{n+m}\right)$ consists of all n-dimensional subspaces of $\mathbb{R}^{n+m}$. Therefore the flow associated to the NDRE (5) on $G^n\left(\mathbb{R}^{n+m}\right)$ is given by

$$\varphi : \mathbb{R} \times G^n\left(\mathbb{R}^{n+m}\right) \to G^n\left(\mathbb{R}^{n+m}\right), \quad (t, S) \mapsto e^{tM}S.$$

The flow exists for all $t \in \mathbb{R}$ and has the flow properties $\varphi(0, S) = S$ and $\varphi(t_2, \varphi(S, t_1)) = \varphi(t_1 + t_2, S)$ for all $S \in G^n\left(\mathbb{R}^{n+m}\right)$ and $t_1, t_2 \in \mathbb{R}$.
In addition it holds that $U(t)$ is nonsingular as long as W exists. This motivates us to consider the set of all graph subspaces of $G^n\left(\mathbb{R}^{n+m}\right)$

$$G_0^n\left(\mathbb{R}^{n+m}\right) := \left\{ \operatorname{range}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) \;\middle|\; U \in \mathbb{R}^{n\times n},\ V \in \mathbb{R}^{m\times n},\ \det U \neq 0 \right\} \subseteq G^n\left(\mathbb{R}^{n+m}\right),$$

together with the function

$$\psi : G_0^n\left(\mathbb{R}^{n+m}\right) \to \mathbb{R}^{m\times n}, \quad \operatorname{range}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) \mapsto VU^{-1}.$$

The function ψ is well defined, as it does not depend on the basis of the graph subspace. Thus, we have that

$$W(t) = \psi\left(\operatorname{range}\left(\begin{bmatrix} U(t) \\ V(t) \end{bmatrix}\right)\right) = \psi\left(\varphi\left(t, \operatorname{range}\left(\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right)\right)\right),$$

and

$$\begin{aligned}\psi^{-1}\left(W(t)\right) &= \operatorname{range}\left(\begin{bmatrix} I_n \\ W(t) \end{bmatrix}\right) = \operatorname{range}\left(\begin{bmatrix} I_n \\ V(t)U(t)^{-1} \end{bmatrix}\right) = \operatorname{range}\left(\begin{bmatrix} U(t) \\ V(t) \end{bmatrix}\right) \\ &= \varphi\left(t, \operatorname{range}\left(\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right)\right),\end{aligned}$$

as long as the solution W exists. Therefore the solution of the NDRE (5) induces a flow on the Grassmanian manifold. The solution W can be recovered from the flow by using ψ, and, vice versa, the flow can be obtained from the solution of the NDRE (5) using ψ^{-1}.

4.2. Solution Formulas

Radon's Lemma (Thm. 4.1) enables a certain solution representations for the DRE (1): Theorem 3.4 ensures that the DRE (1) has a unique solution for $t \geq 0$. By Radon's Lemma (Thm. 4.1) we have

that $U(t)$ is nonsingular for all $t \geq 0$.
Let $H := \begin{bmatrix} A & -BB^T \\ -C^TC & -A^T \end{bmatrix} \in \mathbb{R}^{2n\times 2n}$ be the Hamiltonian matrix corresponding to the DRE (1). The matrices $U(t)$ and $V(t)$ are determined by the linear initial value problem

$$\begin{bmatrix} \dot{U}(t) \\ \dot{V}(t) \end{bmatrix} = -H \begin{bmatrix} U(t) \\ V(t) \end{bmatrix}, \quad \begin{bmatrix} U(0) \\ V(0) \end{bmatrix} = \begin{bmatrix} I_n \\ X_0 \end{bmatrix}. \tag{8}$$

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 (8) decouples.

Theorem 4.2 (Solution representation I for DRE (1), [48])**.**
Let $X \in \mathbb{R}^{n\times n}$ be any solution of the ARE (2)*. Then the solution of the* DRE (1) *for $t \geq 0$ is given by*

$$\begin{aligned} X(t) &= X - e^{t(A-BB^TX^T)^T}\tilde{X}\left(I_n - \int_0^t e^{s(A-BB^TX)}BB^Te^{s(A-BB^TX^T)^T}\mathrm{d}s\tilde{X}\right)^{-1} e^{t(A-BB^TX)}, \\ \tilde{X} &:= X - X_0. \end{aligned}$$

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

$$\begin{aligned} T^{-1}HT &= \begin{bmatrix} I_n & 0 \\ -X & I_n \end{bmatrix} \begin{bmatrix} A & -BB^T \\ -C^TC & -A^T \end{bmatrix} \begin{bmatrix} I_n & 0 \\ X & I_n \end{bmatrix} \\ &= \begin{bmatrix} A-BB^TX & -BB^T \\ 0 & -(A-BB^TX^T)^T \end{bmatrix} =: \tilde{H}. \end{aligned}$$

This gives

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

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

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

By using the variation of constants formula [57, §18] we obtain that $\tilde{U}$ and $\tilde{V}$ are given by

$$\begin{aligned} \tilde{V}(t) &= -e^{t(A-BB^TX^T)^T}(X - X_0), \\ \tilde{U}(t) &= e^{-t(A-BB^TX)} + \int_0^t e^{-(t-s)(A-BB^TX)}BB^T\tilde{V}(s)\mathrm{d}s \end{aligned}$$

$$= e^{-t(A-BB^TX)} \left(I_n - \int_0^t e^{s(A-BB^TX)} BB^T e^{s(A-BB^TX^T)^T} \mathrm{d}s(X - X_0) \right).$$

Since $\tilde{U}(t) = U(t)$ is nonsingular for all $t \geq 0$ and the matrix exponential is nonsingular, the matrix in brackets is also nonsingular for all $t \geq 0$. Finally we obtain

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

□

The formula was presented in [48] without proof. Since the existence of the involved inverse is not trivially established, we provide a proof.

Theorem 4.3 (Solution Representation II for DRE (1), [18, Thm. 1], [47])**.**
Let (A, B) be stabilizable and (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 $\hat{A} := A - BB^TX_\infty$ and $X_L \in \mathbb{R}^{n\times n}$ be the unique symmetric positive semidefinite solution of the Lyapunov equation*

$$\hat{A}X_L + X_L\hat{A}^T + BB^T = 0. \tag{9}$$

Then the solution of the DRE (1) *for $t \geq 0$ is given by*

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

Proof. Similar to the proof of Theorem 4.2 we use similarity transformations to decompose the Hamiltonian matrix H. This is also known as a Riccati-Lyapunov transformation [1, Ch. 3.1.1.]. We obtain

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

We thus get

$$\begin{aligned} \begin{bmatrix} U(t) \\ V(t) \end{bmatrix} &= e^{-tH}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} = e^{-t(T\tilde{T})\hat{H}(T\tilde{T})^{-1}}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} = (T\tilde{T})e^{-t\hat{H}}(T\tilde{T})^{-1}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} \\ &= \begin{bmatrix} I_n & -X_L \\ X_\infty & I_n - X_\infty X_L \end{bmatrix}\begin{bmatrix} e^{-t\hat{A}} & 0 \\ 0 & e^{t\hat{A}^T} \end{bmatrix}\begin{bmatrix} I_n - X_LX_\infty & X_L \\ -X_\infty & I_n \end{bmatrix}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} \\ &= \begin{bmatrix} e^{-t\hat{A}}\left(I_n - X_LX_\infty\right) + X_Le^{t\hat{A}}X_\infty & e^{-t\hat{A}}X_L - X_Le^{t\hat{A}^T} \\ X_\infty e^{-t\hat{A}}\left(I_n - X_LX_\infty\right) - \left(I_n - X_\infty X_L\right)e^{t\hat{A}}X_\infty & X_\infty e^{-t\hat{A}}X_L + \left(I_n - X_\infty X_L\right)e^{t\hat{A}} \end{bmatrix}\begin{bmatrix} I_n \\ X_0 \end{bmatrix} \\ &= \begin{bmatrix} e^{-t\hat{A}}(I_n - X_L(X_\infty - X_0)) + X_Le^{t\hat{A}^T}(X_\infty - X_0) \\ X_\infty e^{-t\hat{A}}(I_n - X_L(X_\infty - X_0)) - (X_\infty X_L + I_n)e^{t\hat{A}^T}(X_\infty - X_0) \end{bmatrix}. \end{aligned} \tag{10}$$

Now observe that

$$U(t) = e^{-t\hat{A}}\left(I_n - (X_L - e^{t\hat{A}}X_Le^{t\hat{A}^T})(X_\infty - X_0)\right),$$

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

therefore

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

□

In [3, 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 [1, Thm. 3.2.1].

4.3. Davison-Maki Methods

The *Davison-Maki method* for the NDRE (5) was proposed in [20]. The method is based on first computing the matrix exponential e^{hM} for a given step size $h > 0$. According to Radon's Lemma (Thm. 4.1) 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

$$
\begin{bmatrix} U(kh) \\ V(kh) \end{bmatrix} = \left(e^{hM} \right)^k \begin{bmatrix} I_n \\ M_0 \end{bmatrix}, \quad W(kh) = V(kh)U(kh)^{-1}. \tag{11}
$$

Another variant of the *Davison-Maki method* updates U and V instead of the matrix exponential. The variant follows from

$$
\begin{bmatrix} U(kh) \\ V(kh) \end{bmatrix} = e^{khM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix} = e^{hM} e^{(k-1)hM} \begin{bmatrix} I_n \\ M_0 \end{bmatrix} = e^{hM} \begin{bmatrix} U((k-1)h) \\ V((k-1)h) \end{bmatrix}. \tag{12}
$$

Both variants of the method are given in Algorithm 1.

Algorithm 1 Davison-Maki method for the NDRE (5) [20, 30]

Assumption: The NDRE (5) 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 4.1, 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 e.g. by a scaling and squaring method:
$\Theta_h = \exp\left(h \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}\right)$;

`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* (Alg. 1) is applied to the DRE (1), usually numerical instabilities occur which are due to the fact that each block of e^{-tH} as well as $U(t)$ and $V(t)$ contains the matrix $e^{-t\hat{A}}$, cp. equation (10). Since $\hat{A} = A - BB^T X_\infty$ 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* (Alg. 1) was also pointed out in [19, 30, 37, 56]. Another reason is that the spectrum of a real Hamiltonian matrix comes in quadruples, that is $\Lambda(H) = \{\lambda_1, \ldots, \lambda_n, -\lambda_1, \ldots, -\lambda_n\}$ with $\mathrm{Re}(\lambda_i) \leq 0$. Therefore, usually, the spectrum of the Hamiltonian contains eigenvalues with positive real part and, thus, also it's matrix exponential grows [43, Prop. 2.3.1].
A suitable modification of the *Davison-Maki method* (Alg. 1) was proposed in [30], but the modified method originates back to [29, p. 9]. By Radon's Lemma (Thm. 4.1), as laid out in Section 4.1, we

have the identity

$$W(kh) = \psi\left(\operatorname{range}\left(\begin{bmatrix} U(kh) \\ V(kh) \end{bmatrix}\right)\right) = \psi\left(e^{khM}\operatorname{range}\left(\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right)\right) = \psi\left(e^{hM}\operatorname{range}\left(\begin{bmatrix} U((k-1)h) \\ V((k-1)h) \end{bmatrix}\right)\right)$$
$$= \psi\left(e^{hM}\operatorname{range}\left(\begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix}\right)\right) = \psi\left(\operatorname{range}\left(e^{hM}\begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix}\right)\right).$$

Therefore the iteration for the *modified Davison-Maki method* is given by

$$\begin{bmatrix} \tilde{U} \\ \tilde{V} \end{bmatrix} := e^{hM}\begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix}, \quad W(kh) = \tilde{V}\tilde{U}^{-1}. \tag{13}$$

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

Algorithm 2 Modified Davison-Maki method for the NDRE (5) [29, 30]

Assumption: The NDRE (5) 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 4.1, 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 $kh < t_f$.

1: $W_0 = M_0$;
2: $k = 1$;
% Compute matrix exponential e.g. by a scaling and squaring method:
3: $\Theta = \exp\left(h\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}\right)$;
% Check the norm of the matrix exponential:
4: **if** $\|\Theta\|_1 > tol_{\texttt{exp}}$ **then**
5: return Error(„1-Norm of the matrix exponential is too large, decrease the step size h“).
6: **end if**
7: Partition $\begin{matrix} \\ n \\ m \end{matrix}\begin{matrix} \begin{matrix} n & m \end{matrix} \\ \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12} & \Theta_{22} \end{bmatrix} \end{matrix} = \Theta$;
8: **while** $kh < t_f$ **do**
9: $U_{\texttt{mod\_dm}} = \Theta_{11} + \Theta_{12}W_{k-1}$;
10: $V_{\texttt{mod\_dm}} = \Theta_{21} + \Theta_{22}W_{k-1}$;
11: $W_k = V_{\texttt{mod\_dm}}U^{-1}_{\texttt{mod\_dm}}$;
12: $k = k + 1$;
13: **end while**

A decrease of the step size $h > 0$, does not improve the accuracy in general, because the iteration is exact. The accuracy is determined by the accuracy of the matrix exponential computation and the matrix inversion. The step size cannot be chosen arbitrary large as the matrix exponential may become too large in norm. In practice we suggest to compute the norm of the matrix exponential before the iteration starts. If the norm is too large, then the step size has to be decreased. In the k-th iteration of Algorithm 2 we have

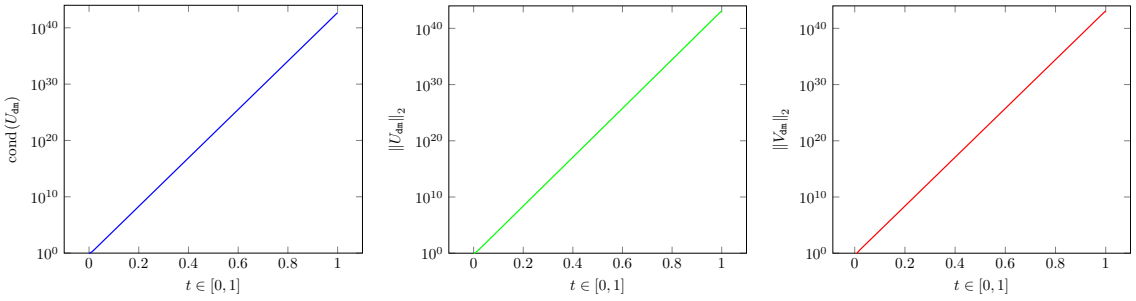
$$\begin{bmatrix} U_{\texttt{mod\_dm}} \\ V_{\texttt{mod\_dm}} \end{bmatrix} = e^{hM}\begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix} = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix}\begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix},$$

and the norm of the iterates can be bounded by

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

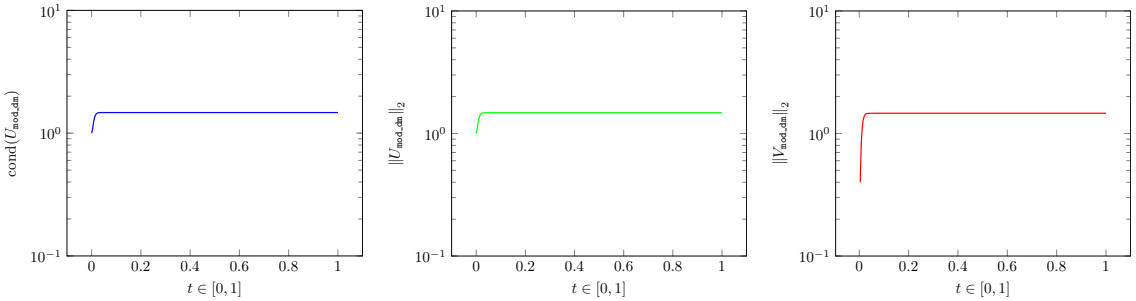
For small step sizes of $h > 0$ it holds $e^{hM} \approx I_{n+m} + hM$ and $\Theta_{11} \approx I_n + hM_{11}$, $\Theta_{12} \approx hM_{12}$, $\Theta_{21} \approx hM_{21}$ and $\Theta_{22} \approx I_m + hM_{22}$. Therefore for small enough step size and moderate norm of the solution $\|W(t)\|$, the norm of the iterates cannot grow heavily in contrast to Algorithm 1. If the norm of the iterates becomes too large during iteration, the step size should be decreased. Assume that the matrix exponential in line 3 of Algorithm 2 was approximated by using the scaling and squaring method, then the intermediates of the squaring phase can be used and the matrix exponential needs not be recomputed from scratch.

Example 4.1 (Exponential Growth Davison-Maki method)**.**
We applied the Davison-Maki method (Alg. 1) *with step size* $h = 2^{-8}$ *to a* DRE *with the same matrices* A, B, C *and* X_0 *as for* Example 3.1. *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* (Alg. 2)*.*



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

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



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

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

If a symmetric solution is expected, then line 11 in Algorithm 2 should be altered with $W_k = \frac{1}{2}\left(W_k + W_k^T\right)$, because due to numerical errors the symmetry will be lost after some iterations. Any computational 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 less matrix-matrix products are needed by time step, compare Algorithm 2 line 8-13 with Algorithm 1 line 5-12 and line 16-21.
The computational cost apart from matrix exponential computation grows linearly with the time step size h, compare Algorithm 2 line 8-13.
The intermediates $U_{\mathtt{dm}}, V_{\mathtt{dm}}$ from Algorithm 2 and $U_{\mathtt{mod\_dm}}, V_{\mathtt{mod\_dm}}$ from Algorithm 1 are usually different. The next lemma shows the connection.

Lemma 4.1.
In the k-th iteration of Algorithm 1 *and* Algorithm 2*, the iterates* $U_{\mathtt{dm}}, V_{\mathtt{dm}}$ *and* $U_{\mathtt{mod\_dm}}, V_{\mathtt{mod\_dm}}$ *fulfill*

$$\operatorname{range}\left(\begin{bmatrix} U_{\mathtt{dm}} \\ V_{\mathtt{dm}} \end{bmatrix}\right) = \operatorname{range}\left(\begin{bmatrix} U_{\mathtt{mod\_dm}} \\ V_{\mathtt{mod\_dm}} \end{bmatrix}\right).$$

Proof. From equation (11) it follows

$$\operatorname{range}\left(\begin{bmatrix} U_{\mathtt{dm}} \\ V_{\mathtt{dm}} \end{bmatrix}\right) = \operatorname{range}\left(e^{hkM}\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right).$$

Equation (13) gives

$$\begin{aligned}\operatorname{range}\left(\begin{bmatrix} U_{\mathtt{mod\_dm}} \\ V_{\mathtt{mod\_dm}} \end{bmatrix}\right) &= \operatorname{range}\left(e^{hM}\begin{bmatrix} I_n \\ W((k-1)h) \end{bmatrix}\right) = \operatorname{range}\left(e^{hM}\begin{bmatrix} I_n \\ V((k-1)h)U((k-1)h)^{-1} \end{bmatrix}\right) \\ &= \operatorname{range}\left(e^{hM}\begin{bmatrix} U((k-1)h) \\ V((k-1)h) \end{bmatrix}\right) = \operatorname{range}\left(e^{khM}\begin{bmatrix} I_n \\ M_0 \end{bmatrix}\right).\end{aligned}$$

□

5. Galerkin Approach for Large-Scale Differential Riccati Equations

In this section we develop a feasible numerical approach for large-scale differential Riccati equations. We consider the DRE (1) and assume that $X_0 = 0$. We develop the Galerkin approach based on two theoretical considerations. First we use the solution formula of Theorem 4.3. We show that the range of the solution X_∞ of the ARE is invariant under the action of the closed-loop matrix $A - BB^T X_\infty$. 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 ansatz 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, a quick decay of the eigenvalues of the solution of the ARE implies a decent approximation of the solution $X(t)$ of the DRE.
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.

5.1. 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 $\left(A - BB^T X_\infty\right)^T$.

Lemma 5.1.
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 $(A - BB^T X_\infty)^T$–invariant.*

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

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

By Theorem 3.3, the columns of Q_∞ are also an orthonormal basis for $\mathcal{K}\left(A^T, C^T\right)$. The space $\mathcal{K}\left(A^T, C^T\right)$ is A^T–invariant. We obtain

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

Finally, we have

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

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

According to Theorem 4.3 the solution of the DRE (1) is for $t \geq 0$ given by

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

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

$$\begin{aligned}X(t) = X_\infty &- e^{t\hat{A}^T} X_\infty e^{t\hat{A}} \\ &- e^{t\hat{A}^T} X_\infty \Big(I_n - \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^T}\right) X_\infty\Big)^{-1} \left(X_L - e^{t\hat{A}} X_L e^{t\hat{A}^T}\right) X_\infty e^{t\hat{A}}.\end{aligned} \tag{14}$$

Derivation by using the exact solution X_∞ of the ARE
By Lemma 5.1 it holds that range (X_∞) is invariant under $\hat{A}^T$. Assume now that X_∞ is given in factorized form, this means that $X_\infty = Z_\infty Z_\infty^T$ 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 $X_\infty = 0$ as well as the solution $X(t)$. Now it holds that range$(X_\infty) =$ range(Z_∞) and consequently range(Z_∞) is invariant under $\hat{A}^T$.
By means of 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 $Z_\infty = Q_\infty S_\infty V_\infty^T$, range (Q_∞) = range (Z_∞) and

$$Z_\infty Z_\infty^T = Q_\infty S_\infty^2 Q_\infty^T.$$

Because of the invariance we get

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

Now observe that

$$e^{t\hat{A}^T} X_\infty = e^{t\hat{A}^T} Z_\infty Z_\infty^T = e^{t\hat{A}^T} Q_\infty S_\infty^2 Q_\infty^T = Q_\infty e^{tQ_\infty^T \hat{A}^T Q_\infty} S_\infty^2 Q_\infty^T. \tag{15}$$

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

$$X(t) = X_\infty - Q_\infty \tilde{X}(t) Q_\infty^T. \tag{16}$$

We use the DRE (1) and equation (16) and get a differential equation for $\tilde{X}(t)$

$$\dot{\tilde{X}}(t) = Q_\infty^T \hat{A}^T Q_\infty \tilde{X}(t) + \tilde{X}(t) Q_\infty^T \hat{A} Q_\infty + \tilde{X} Q_\infty^T B B^T Q_\infty \tilde{X}(t), \tag{17a}$$

$$\tilde{X}(0) = Q_\infty^T X_\infty Q_\infty. \tag{17b}$$

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

Let now $Z_N Z_N^T = X_N \approx X_\infty$ be a low-rank approximation obtained by a numerical method. We replace X_∞ by X_N in formula 14 and obtain

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

Let $Z_N = Q_N S_N V_N^T$ be the compact singular value decomposition of the low-rank factor. According to formula 15, we propose to approximate the action of the matrix exponential by

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

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

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

We assume that the numerical low-rank approximation is accurate enough such that $\mathcal{R}(X_N) \approx 0$. Then it holds:

$$\left\| Q_N^T \mathcal{R}(X_N) Q_N \right\|_2 \leq \|\mathcal{R}(X_N)\|_2 \approx 0.$$

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

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

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

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

This means that the diagonal matrix S_∞^2 contains all non-zero eigenvalues of X_∞ in a non-increasing fashion. We have that range (X_∞) = range (Z_∞) = range (Q_∞). Because of Theorem 3.3 it holds that range $(Q_\infty) = \mathcal{K}\left(A^T, C^T\right)$. According to Theorem 3.5 we can represent the solution in the following form

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

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

Theorem 5.1.
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:*

$$\left|q_i^T X(t) q_j\right| \leq \sqrt{\lambda_i^\downarrow(X_\infty)\, \lambda_j^\downarrow(X_\infty)}, \tag{18}$$

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

Proof. According to Theorem 3.1 the inequality $0 \preccurlyeq X(t) \preccurlyeq X_\infty$ holds for all $t \geq 0$. By multiplying the inequality with q_i^T from the left and q_i from the right we obtain

$$0 \leq q_i^T X(t) q_i \leq q_i^T X_\infty q_i = \lambda_i^\downarrow(X_\infty)\, q_i^T q_i = \lambda_i^\downarrow(X_\infty).$$

Now let $i \neq j$ and $\alpha, \beta \in \mathbb{R}$. Again by multiplying the inequality with $\alpha q_i + \beta q_j$ we obtain

$$0 \leq (\alpha q_i + \beta q_j)^T X(t) (\alpha q_i + \beta q_j) \leq (\alpha q_i + \beta q_j)^T X_\infty (\alpha q_i + \beta q_j).$$

Since $X(t)$ and X_∞ are symmetric, it applies that

$$\alpha^2 q_i^T X(t) q_i + 2\alpha\beta q_i^T X(t) q_j + \beta^2 q_j^T X(t) q_j \leq \alpha^2 q_i^T X_\infty q_i + 2\alpha\beta q_i^T X_\infty q_j + \beta^2 q_j^T X_\infty q_j.$$

As q_i and q_j are different orthonormal eigenvectors of X_∞, we obtain for the right hand side

$$\begin{aligned}\alpha^2 q_i^T X_\infty q_i + 2\alpha\beta q_i^T X_\infty q_j + \beta^2 q_j^T X_\infty q_j &= \alpha^2 \lambda_i^\downarrow(X_\infty) + 2\alpha\beta\lambda_j^\downarrow(X_\infty)\, q_i^T q_j + \beta^2 \lambda_i^\downarrow(X_\infty) \\ &= \alpha^2 \lambda_i^\downarrow(X_\infty) + \beta^2 \lambda_i^\downarrow(X_\infty).\end{aligned}$$

As $X(t)$ is symmetric positive semidefinite, the following inequality holds for the left hand side.

$$\alpha^2 q_i^T X(t) q_i + 2\alpha\beta q_i^T X(t) q_j + \beta^2 q_j^T X(t) q_j \geq 2\alpha\beta q_i^T X(t) q_j.$$

Now we have

$$0 \leq \alpha^2 \lambda_i^\downarrow(X_\infty) - 2\alpha\beta q_i^T X(t) q_j + \beta^2 \lambda_j^\downarrow(X_\infty) = \begin{bmatrix}\alpha & \beta\end{bmatrix} \begin{bmatrix}\lambda_i^\downarrow(X_\infty) & -{q_i}^T X(t) q_j \\ -{q_i}^T X(t) q_j & \lambda_j^\downarrow(X_\infty)\end{bmatrix} \begin{bmatrix}\alpha \\ \beta\end{bmatrix}.$$

Since this holds for all $\alpha, \beta \in \mathbb{R}$ the matrix

$$\begin{bmatrix}\lambda_i^\downarrow(X_\infty) & -{q_i}^T X(t) q_j \\ -{q_i}^T X(t) q_j & \lambda_j^\downarrow(X_\infty)\end{bmatrix}$$

is symmetric positive semidefinite. Therefore its determinant must be non-negative,

$$0 \leq \lambda_i^{\downarrow}(X_\infty)\,\lambda_j^{\downarrow}(X_\infty) - \left({q_i}^T X(t) q_j\right)^2.$$

Finally this leads to

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

□

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 (18) from Theorem 5.1, the values $\left|q_i^T X(t) q_j\right|$ also decay for $i+j$ increasing. We have that

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

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

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

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

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

together with the orthogonal projection

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

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

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

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

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

and for the approximation error we obtain

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

$$\leq \sqrt{\sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{p} \lambda_i^{\downarrow}(X_\infty)\,\lambda_j^{\downarrow}(X_\infty)}.$$

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

$$\|X(t) - \mathcal{P}_{\infty,k}(X(t))\|_F \leq \sqrt{\sum_{\substack{i,j=1 \\ i>k \vee j>k}}^{n} \lambda_i^{\downarrow}(X_\infty)\,\lambda_j^{\downarrow}(X_\infty)}. \tag{19}$$

We propose therefore to setup 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. Then 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. The small singular values can be safely truncated from the singular value decomposition by virtue of Thm. 5.1. This reduces also the dimension of the trial space. Let

$$Z_N = Q_N S_N V_N^T.$$

be the truncated reduced singular value decomposition of the low-rank approximation. With that, the trial space for the Galerkin approach is given by

$$\left\{ Q_N \tilde{X} Q_N^T \mid \tilde{X} \in \mathbb{R}^{p\times p} \right\},$$

and, as $X(t)$ converges to X_∞ and $X_\infty \approx Z_N Z_N^T$, we propose the Galerkin ansatz

$$X(t) \approx Z_N Z_N^T - Q_N \tilde{X}(t) Q_N^T.$$

Example 5.1 (Decay of Absolute Values of Entries)**.**
We illustrate the decay of $\left|q_i X(t) q_j^T\right|$ *in* Figures 4-8. *We have chosen the same matrices as for the* Example 3.1. *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 9.

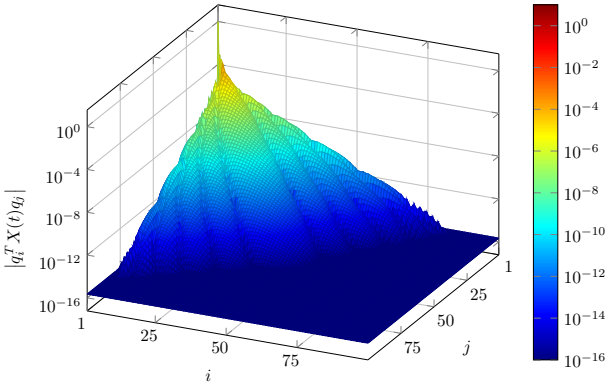


Fig. 4. Decay of $\left|q_i^T X(t) q_j\right|$ for $t = 1$.

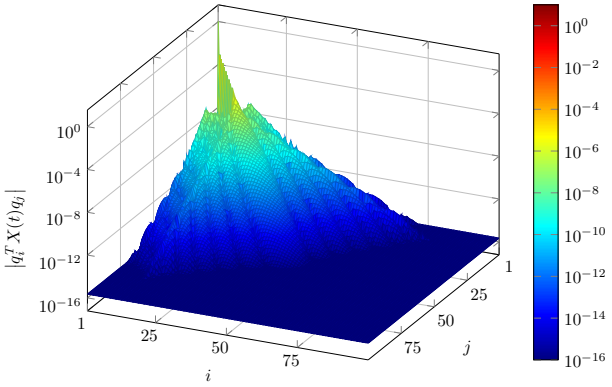


Fig. 5. Decay of $\left|q_i^T X(t) q_j\right|$ for $t = 3$.

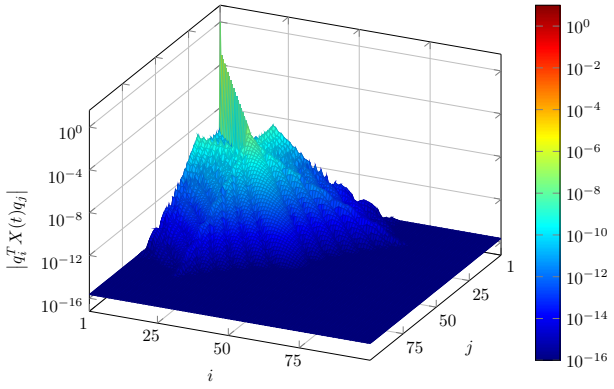


Fig. 6. Decay of $\left|q_i^T X(t) q_j\right|$ for $t = 5$.

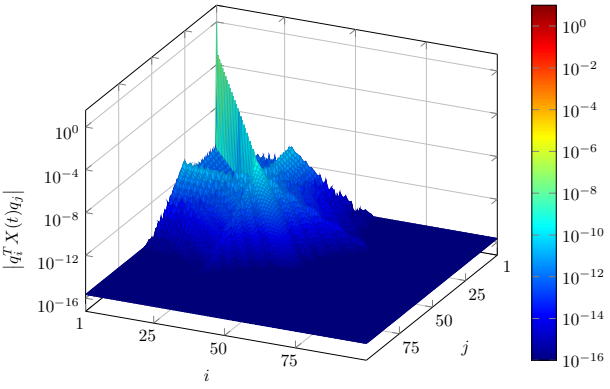


Fig. 7. Decay of $\left|q_i^T X(t) q_j\right|$ for $t = 7$.

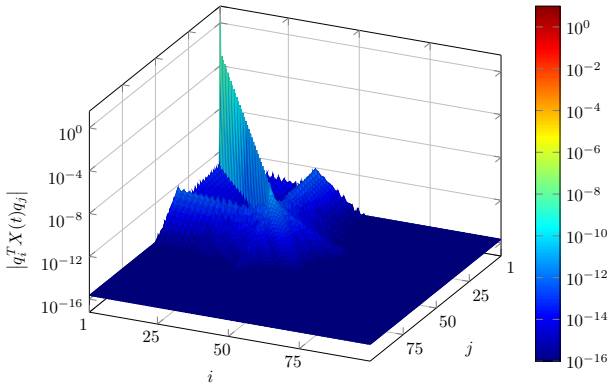


Fig. 8. Decay of $\left|q_i^T X(t) q_j\right|$ for $t = 9$.

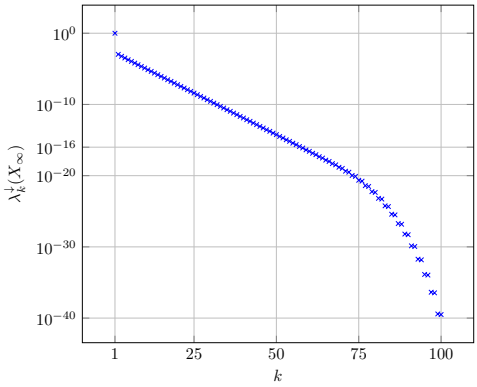


Fig. 9. The eigenvalue decay of X_∞.

Remark 5.1.
With minor adjustments, all arguments also hold for the generalized DRE

$$M^T \dot{X}(t) M = A^T X(t) M + M^T X(t) A - M^T X(t) B B^T X(t) M + C^T C, \tag{20a}$$

$$X(0) = 0, \tag{20b}$$

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 reads as written down in Algorithm 3.

Algorithm 3 Galerkin approach for the generalized DRE (20) (`ARE-Galerkin`)

Assumption: (AM^{-1}, B) is stabilizable and (AM^{-1}, CM^{-1}) is detectable.
Input: $M,\ A \in \mathbb{R}^{n\times n},\ B \in \mathbb{R}^{n\times b},\ C \in \mathbb{R}^{c\times n}$.
Output: $X(t) \approx Z_\infty Z_\infty^T - Q_\infty \tilde{X}(t) Q_\infty^T$ that approximates the solution to

$M^T\dot{X}(t)M = A^TX(t)M + M^TX(t)A - M^TX(t)BB^TX(t)M + C^TC,\ X(0) = 0.$

% Solve the ARE:

$A^TX_\infty M + M^TX_\infty A - M^TX_\infty BB^TX_\infty M + C^TC = 0$ for $X_\infty \approx Z_\infty Z_\infty^T$ and $Z_\infty \in \mathbb{R}^{n\times q}$;

% Compute compact singular value decomposition:

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

% Set tolerance to largest singular value times machine epsilon:

$tol = \varepsilon_{\text{machine}} \cdot S_\infty(1,1)$;

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

$idx = \texttt{diag}(S_\infty) \geq tol$;
$S_\infty = S_\infty(idx, idx)$;
$Q_\infty = Q_\infty(:, idx)$;
$Z_\infty = Q_\infty S_\infty$;

% Compute matrices:

$A_F = Q_\infty^T \left(AM^{-1} - BB^TZ_\infty Z_\infty^T\right) Q_\infty$;
$B_F = Q_\infty^T B$;

% Solve the differential equation using Algorithm 2:

$\dot{\tilde{X}}(t) = A_F^T\tilde{X}(t) + \tilde{X}(t)A_F + \tilde{X}(t)B_FB_F^T\tilde{X}(t),\ \tilde{X}(0) = S_\infty^2$;

6. Numerical Experiments

To quantify the performance of Algorithm 3, we consider a number of differential Riccati equations that are used to define optimal controls. Concretely, we consider the generalized differential Riccati equation

$$M^T\dot{X}(t)M = A^TX(t)M + M^TX(t)A - M^TX(t)BB^TM + C^TC, \tag{21a}$$
$$X(0) = 0. \tag{21b}$$

and their realizations. First, we consider the `RAIL` benchmark example, that is a finite element discretization of a heat equation; see [16] for the model description. The second example, `CONV_DIFF`, derives from a finite-differences discretized heat equation with convection on the unit square with homogenous Dirichlet boundary conditions,

$$\frac{\partial}{\partial t}x(\xi,t) - \Delta x(\xi,t) - v\cdot\nabla x(\xi,t) = f(\xi)u(t) \quad \text{in } \Omega\times(0,T)$$

where $\Omega = (0,1)^2$ and $v = [10,100]^T$; see [46].
On both examples, we compare the proposed method with the splitting methods developed in [52,53]. The 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 a matrix exponential by means of Krylov subspace methods. We used the MATLAB implementation DREsplit [55] of the splitting methods for our experiments. In the tests, we employed the *Lie* and *Strang* splitting of order 1 and 2 respectively, as well as the symmetric splitting of order 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 constant time step size h. The basic information about the setup of the benchmark problems are given in Table 1.

problem	n	matrices	interval	reference solution
`RAIL`	5177	M symmetric positive definite, A symmetric, $M^{-1}A$ stable, $B\in\mathbb{R}^{n\times 6}$, $C\in\mathbb{R}^{7\times n}$	$[0,464]$	`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`, $h=2^{-18}$

Table 1: Information about benchmark problems.

All computations are carried out on a machine with 2× Xeon® Skylake Silver 4110 @ 2.10GHz CPU with 8 cores, 192 GB Ram and MATLAB 2018a. We have used the low-rank Newton ADI iteration implemented in MEX-M.E.S.S.[12] to solve the algebraic Riccati equations; as required for our approach as laid out in Algorithm 3.
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 $\|X_{\texttt{ref}}(t)\|$ as well as 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 and be found in Appendices A and B. The computational costs for both methods are given in Section 6.2. Also, we evaluate the best approximation in the trial space of the reference solution, which is given by

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

where Q_∞ is the matrix from Algorithm 3 Line 8.
The code of the implementation and the precomputed reference solution are available as mentioned in Figure 10.

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

`doi:10.5281/zenodo.2629737`
`https://gitlab.mpi-magdeburg.mpg.de/behr/behbh19_dre_are_galerkin_code`

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

Fig. 10. Link to code and data.

6.1. Galerkin Approach and Splitting Schemes

The initial step of Algorithm 3 requires the solution to the associated ARE. For this task we call MEX-M.E.S.S. that iteratively computes the numerical solution to the following absolute and relative residuals

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

and

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

The achieved values for the different test setups as well as the number of columns of the corresponding Z_∞ after truncation (see Step 5 of Algorithm 3), that define the dimension of the reduced model, are listed in Table 2.

instance	n	size of Galerkin system	absolute residual	relative residual
`RAIL`	5,177	319	$5.068 \cdot 10^{-14}$	$4.223 \cdot 10^{-15}$
`CONV_DIFF`	6,400	56	$1.922 \cdot 10^{-10}$	$4.291 \cdot 10^{-14}$

Table 2: Residuals for the ARE $0 = A^T X M + M X A - M^T X B B^T X M + C^T C$.

The 1-norm bound for the matrix exponential $tol_{\texttt{exp}}$ from Algorithm 2 was set to $1 \cdot 10^{10}$. The

resulting step sizes are given in Table 3.

Instance	n	Step sizes h
RAIL	5,177	$\{2^0, 2^{-1}, \ldots, 2^{-5}\}$
CONV_DIFF	6,400	$\{2^{-12}, 2^{-13}, \ldots, 2^{-16}\}$

Table 3: Step sizes h for modified Davison-Maki method Algorithm 2.

We plot the numerical errors in Figures 15–18 and 21–24. Figures 19, 20, 25 and 26 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: Firstly, the accuracy of the *modified Davison-Maki method*; cf. Figure 16 and 18 is independent of the step size, as discussed in Section 4.3. 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 solve 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., 42s ($h = 2^{-3}$) and 77s ($h = 2^{-4}$) (see Figure 11), one infers that most of the time is spent to solve the dense DRE. Conversely, for the `CONV_DIFF` benchmark problem, most of the time (45s) was used to solve the ARE. As the resulting Galerkin projected DRE system is of size 56 only, the computational costs for the time integration are vanishingly small. Accordingly, the differences in the effort caused by finer time grids are hardly visible; see Figure 13.
The reference solution for the `RAIL` problem is large in norm what makes the absolute error comparatively large; see the Figure 19 in Appendix A.
In both examples, in terms of accuracy, the `ARE-Galerkin` approximation is nearly at the same level as the high order splitting schemes, cf. Figures 16, 32 and Figures 22, 38. We note, however, 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 clearly outperforms the splitting methods in terms of computational time versus accuracy in all test examples.

6.2. Computational Time

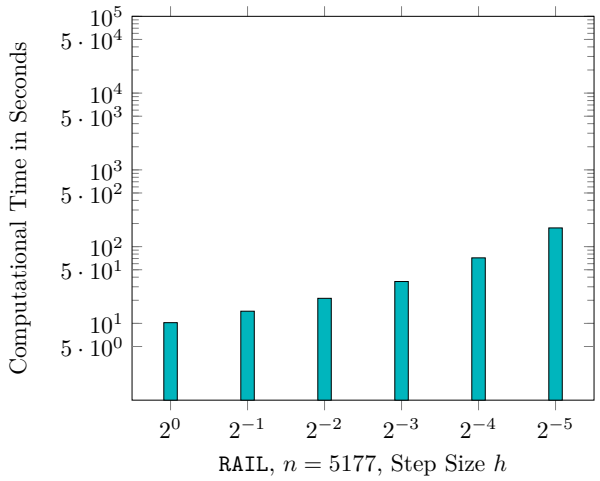


Fig. 11. Timing for `ARE-Galerkin`.

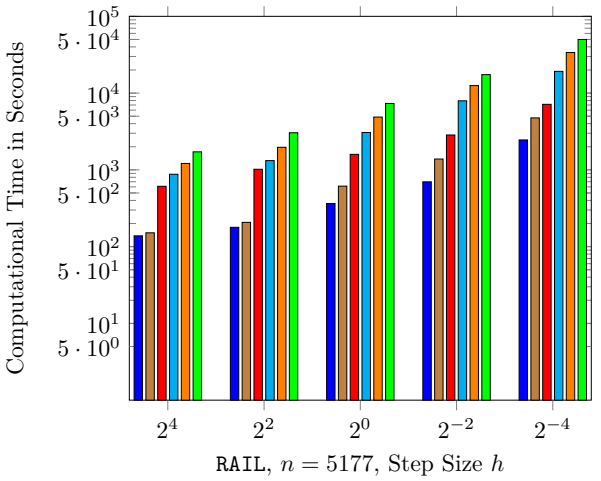


Fig. 12. Timing for splitting schemes.

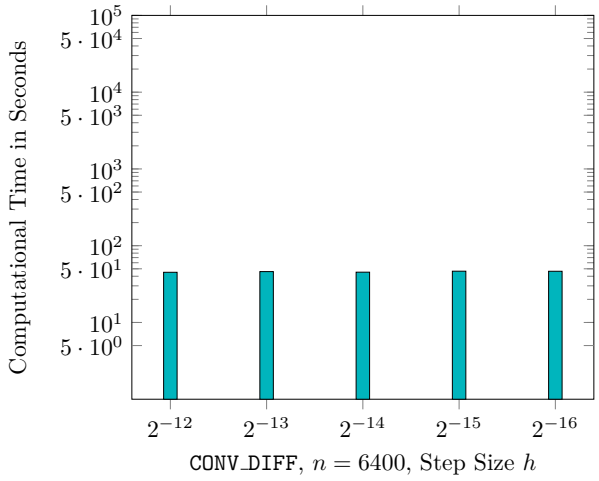


Fig. 13. Timing for ARE-Galerkin.

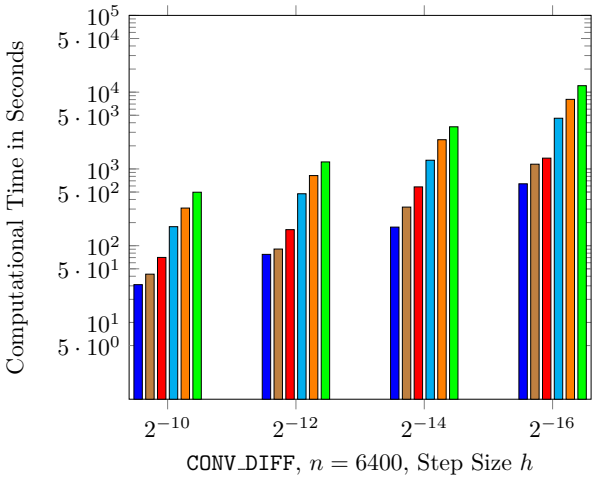
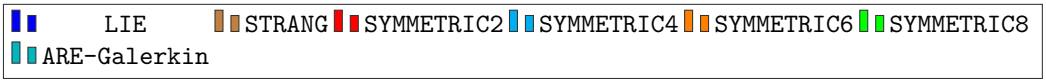


Fig. 14. Timing for splitting schemes.



7. Conclusion

We have reviewed and extended fundamental properties of the solution to the differential and algebraic Riccati equation 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 existing 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 differential Riccati equations.

Our proposed Galerkin ansatz bases on a low-rank approximation of the associated algebraic Riccati equation (ARE) for which there are efficient solvers. Moreover, the information on the residual and on 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 the more that the use of the *Davison-Maki methods* leads to an *exact* time discretization.

Future work will deal with the treatment of nonzero initial conditions. While the formulas are easily extended to this case, the invariance properties and the eigenvalue comparisons, that were the backbone of our numerical approach, are no longer given in general. For the (not so) special case that the initial condition X_0 writes as $X_0 = C^T W C$ with a symmetric positive definite weighting matrix W, the flow invariance as established in Section 5.1 still holds so that the presented algorithm can be applied without modification. For a general low-rank initial condition $X_0 = Z_0 Z_0^T$ the flow invariance can be achieved by taking the columns of the solution to the ARE (2) with $C^T C$ replaced by $\left[C^T, Z_0\right]\left[C^T, Z_0\right]^T$ as the Galerkin ansatz space. In any case, the inequality $0 \preccurlyeq X(t) \preccurlyeq X_\infty$, where X_∞ is the solution of the ARE, does not hold anymore and, thus, approximation quality cannot be assured as in (19). However, if one can find a matrix $\tilde{X}$ for which the comparison $0 \preccurlyeq X(t) \preccurlyeq \tilde{X}$ holds, all arguments of Section 5.2 apply accordingly.

References

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank. *Matrix Riccati Equations in Control and Systems Theory*. Birkhäuser, Basel, Switzerland, 2003.

[2] L. Amodei and J.-M. Buchot. An invariant subspace method for large-scale algebraic Riccati equation. *Appl. Numer. Math.*, 60(11):1067–1082, 2010.

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

[4] V. Angelova, M. Hached, and K. Jbilou. Approximate solutions to large nonsymmetric differential Riccati problems with applications to transport theory. Technical Report arXiv:1801.01291v2, arXiv, 2019. math.NA.

[5] A. C. Antoulas, D. C. Sorensen, and Y. Zhou. On the decay rate of Hankel singular values and related issues. *Syst. Cont. Lett.*, 46(5):323–342, 2002.

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

[7] M. Beck and S. J. A. Malham. Computing the Maslov index for large systems. *Proc. Amer. Math. Soc.*, 143(5):2159–2173, 2015.

[8] B. Beckermann and A. Townsend. On the singular values of matrices with displacement structure. *SIAM J. Matrix Anal. Appl.*, 38(4):1227–1248, 2017.

[9] 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(1), 2018.

[10] M. Behr, P. Benner, and J. Heiland. Solution formulas for differential Sylvester and Lyapunov equations. e-print 1811.08327, arXiv, November 2018. math.NA.

[11] P. Benner and Z. Bujanović. On the solution of large-scale algebraic Riccati equations by using low-dimensional invariant subspaces. *Linear Algebra Appl.*, 488:430–459, 2016.

[12] P. Benner, M. Köhler, and J. Saak. M.E.S.S. – Matrix Equations Sparse Solver. `https://www.mpi-magdeburg.mpg.de/projects/mess`.

[13] P. Benner and H. Mena. BDF methods for large-scale differential Riccati equations. In B. De Moor, B. Motmans, J. Willems, P. Van Dooren, and V. Blondel, editors, *Proc. 16th Intl. Symp. Mathematical Theory of Network and Systems, MTNS 2004*, 2004.

[14] P. Benner and H. Mena. Rosenbrock methods for solving Riccati differential equations. *IEEE Trans. Autom. Control*, 58(11):2950–2957, 2013.

[15] P. Benner and H. Mena. Numerical solution of the infinite-dimensional LQR-problem and the associated differential Riccati equations. *J. Numer. Math.*, 26(1):1–20, 2018.

[16] P. Benner and J. Saak. A semi-discretized heat transfer model for optimal cooling of steel profiles. In P. Benner, V. Mehrmann, and D. Sorensen, editors, *Dimension Reduction of Large-Scale Systems*, volume 45 of *Lect. Notes Comput. Sci. Eng.*, pages 353–356. Springer-Verlag, Berlin/Heidelberg, Germany, 2005.

[17] R. A. Brocket. *Finite Dimensional Linear Systems.* Wiley, New York, 1970.

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

[19] C. H. Choi. A survey of numerical methods for solving matrix Riccati differential equations. In *IEEE Proceedings on Southeastcon*, pages 696–700 vol.2, 1990.

[20] E. J. Davison and M. C. Maki. The numerical solution of the matrix Riccati differential equation. *IEEE Trans. Autom. Control*, 18:71–73, 1973.

[21] 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):371–389, 2004.

[22] L. Grubišić and D. Kressner. On the eigenvalue decay of solutions to operator Lyapunov equations. *Syst. Cont. Lett.*, 73:42–47, 2014.

[23] Y. Güldoğan, M. Hached, K. Jbilou, and M. Kurulay. Low-rank approximate solutions to large-scale differential matrix Riccati equations. *Applicationes Mathematicae*, 45(2):233–254, 2018.

[24] M. Hached and K. Jbilou. Computational Krylov-based methods for large-scale differential Sylvester matrix problems. *Numer. Lin. Alg. Appl.*, 25(5):e2187, 14, 2018.

[25] M. Hached and K. Jbilou. Numerical methods for differential linear matrix equations via Krylov subspace methods. Technical Report arXiv:1805.10192v1, arXiv, 2018. math.NA.

[26] M. Hached and K. Jbilou. Numerical solutions to large-scale differential Lyapunov matrix equations. *Numer. Algorithms*, 2018.

[27] J. Heiland. *Decoupling and Optimization of Differential-Algebraic Equations with Application in Flow Control*. Dissertation, TU Berlin, 2014.

[28] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

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

[30] C. Kenney and R. B. Leipnik. Numerical integration of the differential matrix Riccati equation. *IEEE Trans. Autom. Control*, 30:962–970, 1985.

[31] G. Kirsten and V. Simoncini. Order reduction methods for solving large-scale differential matrix Riccati equations. e-print 1905.12119, arXiv, 2019. math.NA.

[32] H. W. Knobloch and H. Kwakernaak. *Lineare Kontrolltheorie*. Springer-Verlag, Berlin, 1985. In German.

[33] M. Köhler, N. Lang, and J. Saak. Solving differential matrix equations using Parareal. *Proc. Appl. Math. Mech.*, 16(1):847–848, 2016.

[34] A. Koskela and H. Mena. A structure preserving Krylov subspace method for large scale differential Riccati equations. e-print arXiv:1705.07507, arXiv, 2017. math.NA.

[35] N. Lang. *Numerical Methods for Large-Scale Linear Time-Varying Control Systems and related Differential Matrix Equations*. Dissertation, Technische Universität Chemnitz, Germany, 2017.

[36] N. Lang, H. Mena, and J. Saak. On the benefits of the LDL^T factorization for large-scale differential matrix equation solvers. *Linear Algebra Appl.*, 480:44–71, 2015.

[37] A. J. Laub. Schur techniques for Riccati differential equations. In D. Hinrichsen and A. Isidori, editors, *Feedback Control of Linear and Nonlinear Systems*, pages 165–174. Springer-Verlag, New York, 1982.

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

[39] C. Martin. Grassmannian manifolds, Riccati Equations and Feedback Invariants of Linear Systems. In *Geometrical Methods for the Theory of Linear Systems*, pages 195–211. Springer, 1980.

[40] T. McCauley. Computing the Maslov index from singularities of a matrix Riccati equation. *J. Dyn. Diff. Equat.*, 29(4):1487–1502, 2017.

[41] H. Mena. *Numerical Solution of Differential Riccati Equations Arising in Optimal Control of Partial Differential Equations*. Dissertation, Escuela Politécnica Nacional, Ecuador, 2007.

[42] H. Mena, L.-M. Pfurtscheller, and T. Stillfjord. GPU acceleration of splitting schemes applied to differential matrix equations. *Numer. Algorithms*, 2019.

[43] K. R. Meyer and D. C. Offin. *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, volume 90 of *Applied Mathematical Sciences*. Springer, Cham, third edition, 2017.

[44] M. Opmeer. Decay of singular values of the Gramians of infinite-dimensional systems. In *Proceedings 2015 European Control Conference (ECC)*, pages 1183–1188, Linz, Austria, 2015. IEEE.

[45] T. Penzl. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Syst. Cont. Lett.*, 40:139–144, 2000.

[46] T. Penzl. LYAPACK Users Guide. Technical Report SFB393/00-33, Sonderforschungsbereich 393 *Numerische Simulation auf massiv parallelen Rechnern*, TU Chemnitz, 09107 Chemnitz, Germany, 2000. Available from `http://www.tu-chemnitz.de/sfb393/sfb00pr.html`.

[47] V. Radisavljevic. Improved Potter-Anderson-Moore algorithm for the differential Riccati equation. *Appl. Math. Comput.*, 218(8):4641–4646, 2011.

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

[49] C. R. Schneider. Global aspects of the matrix Riccati equation. *Math. Systems Theory*, 7(3):281–286, 1973.

[50] M. A. Shayman. Phase portrait of the matrix Riccati equation. *SIAM J. Control Optim.*, 24(1):1–65, 1986.

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

[52] T. Stillfjord. Low-rank second-order splitting of large-scale differential Riccati equations. *IEEE Trans. Autom. Control*, 60(10):2791–2796, 2015.

[53] T. Stillfjord. Adaptive high-order splitting schemes for large-scale differential Riccati equations. *Numer. Algorithms*, 78:1129–1151, 2018.

[54] T. Stillfjord. Singular value decay of operator-valued differential Lyapunov and Riccati equations. *SIAM J. Control Optim.*, 56:3598–3618, 2018.

[55] T. Stilljford. DREsplit. `www.tonystillfjord.net/DREsplit.zip`.

[56] D. R. Vaughan. A negative exponential solution for the matrix Riccati equation. *IEEE Trans. Autom. Control*, 14:72–75, 1969.

[57] W. Walter. *Ordinary differential equations*, volume 182 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

A. Numerical Results for Galerkin Approach

RAIL, $n = 5177$ and $M^T\dot{X}(t)M = A^TX(t)M + M^TX(t)A - M^TX(t)BB^TX(t)M + C^TC,\ X(0) = 0.$

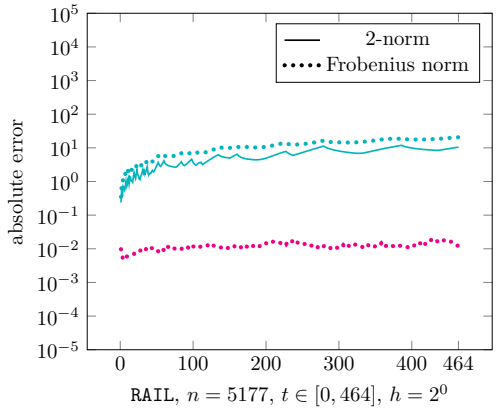


Fig. 15. Absolute error of the Galerkin and Best approximation.

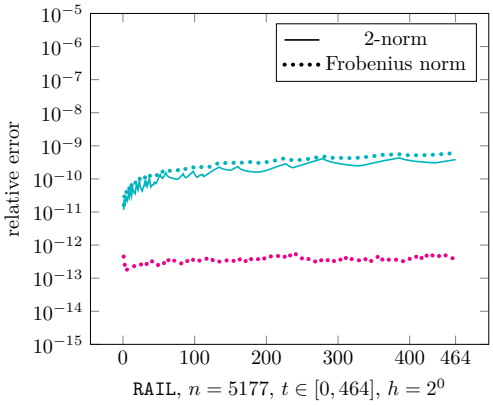


Fig. 16. Relative error of the Galerkin and Best approximation.

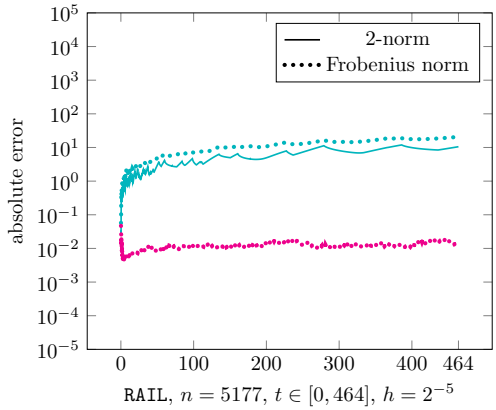


Fig. 17. Absolute error of the Galerkin and Best approximation.

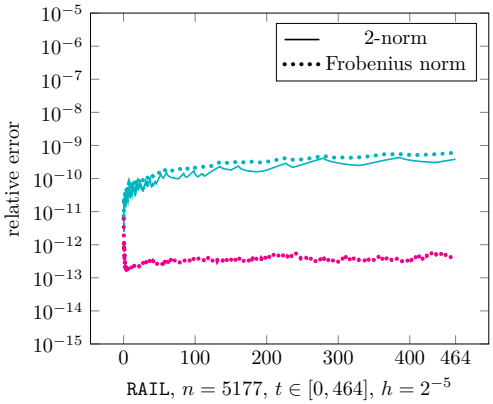


Fig. 18. Relative error of the Galerkin and Best approximation.

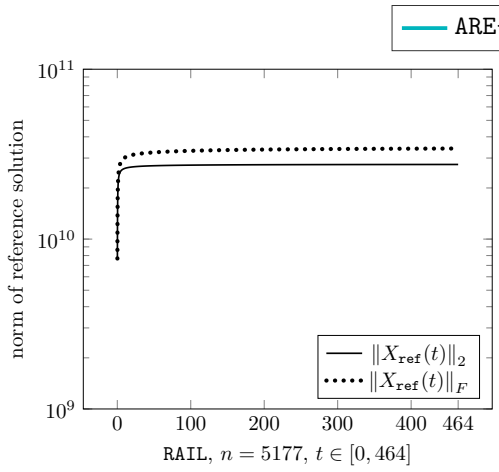


Fig. 19. Norm of the reference solution.

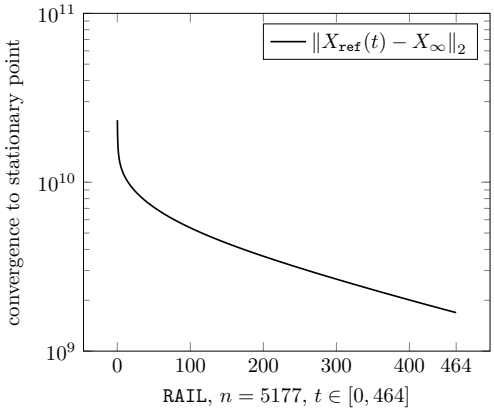


Fig. 20. Convergence to the stationary point.

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

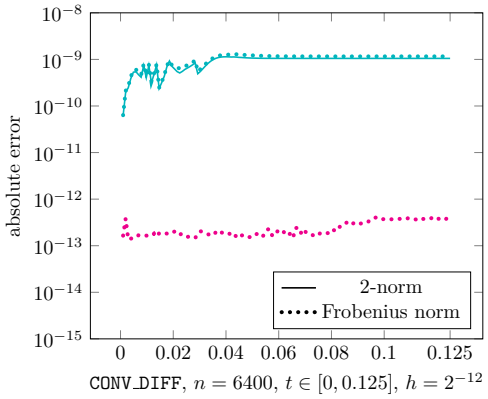


Fig. 21. Absolute error of the Galerkin and Best approximation.

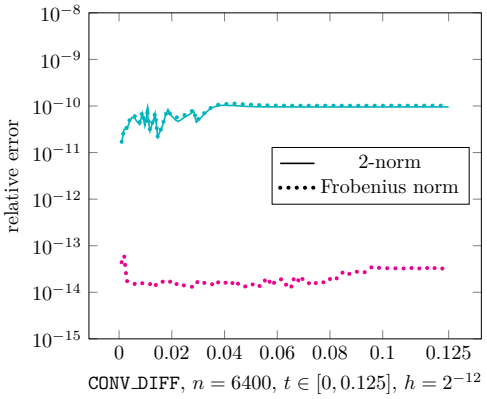


Fig. 22. Relative error of the Galerkin and Best approximation.

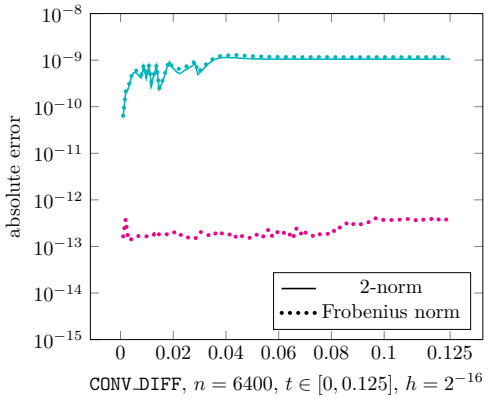


Fig. 23. Absolute error of the Galerkin and Best approximation.

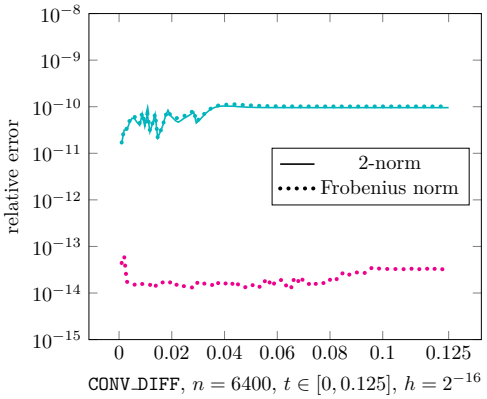


Fig. 24. Relative error of the Galerkin and Best approximation.

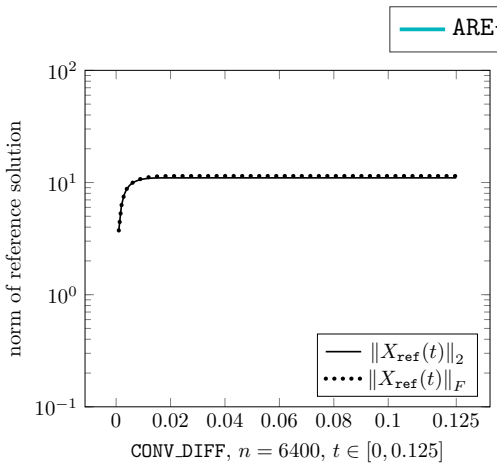


Fig. 25. Norm of the reference solution.

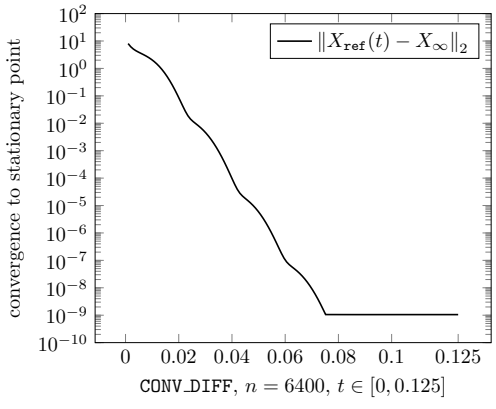


Fig. 26. Convergence to the stationary point.

B. Numerical Results for Splitting Schemes

RAIL, $n = 5177$ and $M^T\dot{X}(t)M = A^TX(t)M + M^TX(t)A - M^TX(t)BB^TX(t)M + C^TC,\ X(0) = 0.$

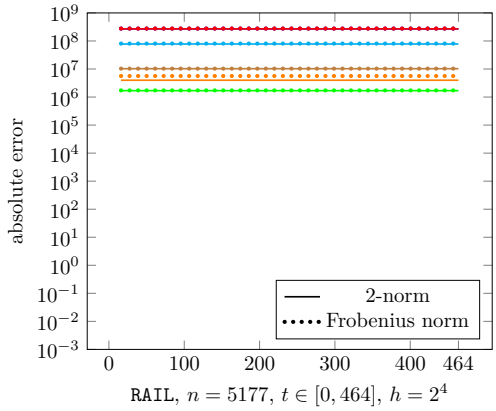


Fig. 27. Absolute error of the splitting scheme approximation.

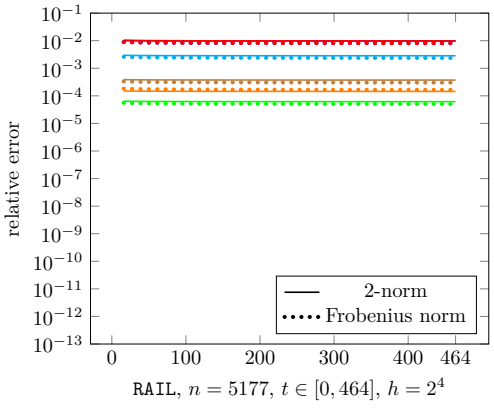


Fig. 28. Relative error of the splitting scheme approximation.

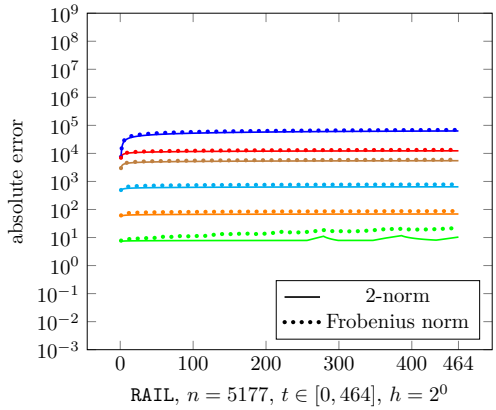


Fig. 29. Absolute error of the splitting scheme approximation.

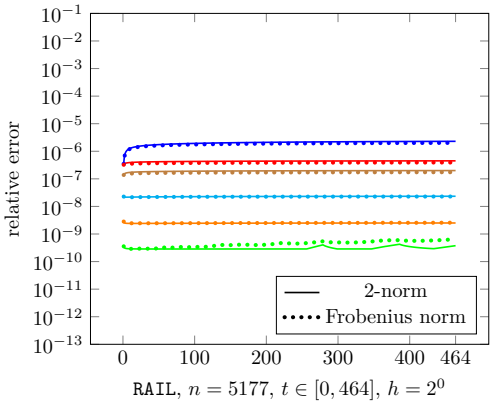


Fig. 30. Relative error of the splitting scheme approximation.

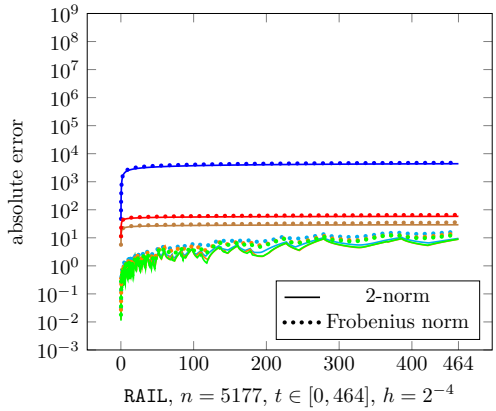


Fig. 31. Absolute error of the splitting scheme approximation.

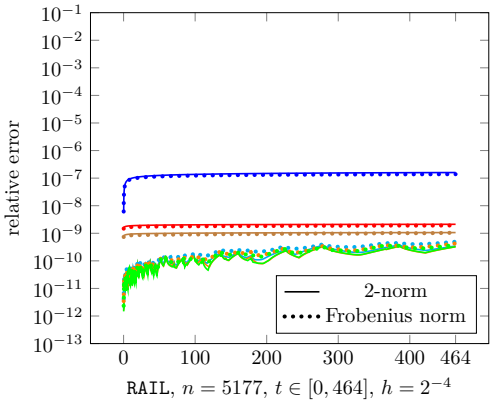
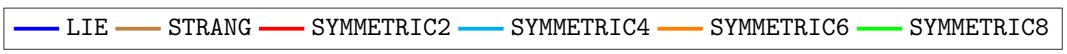


Fig. 32. Relative error of the splitting scheme approximation.



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

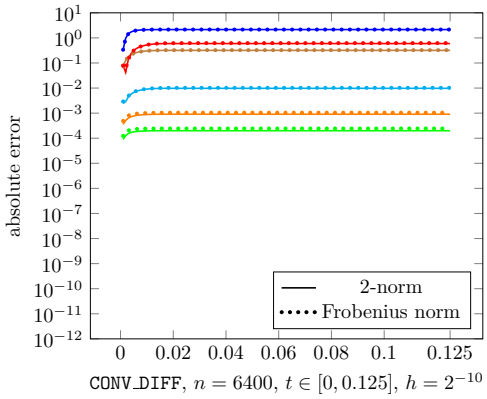


Fig. 33. Absolute error of the splitting scheme approximation.

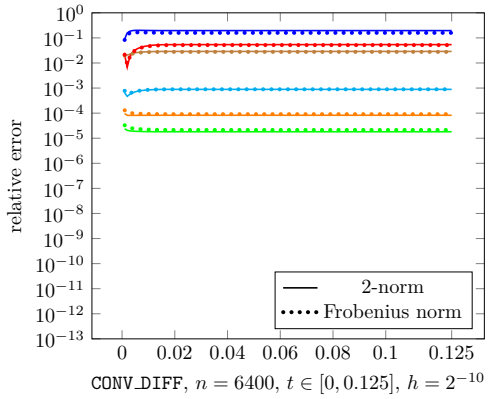


Fig. 34. Relative error of the splitting scheme approximation.

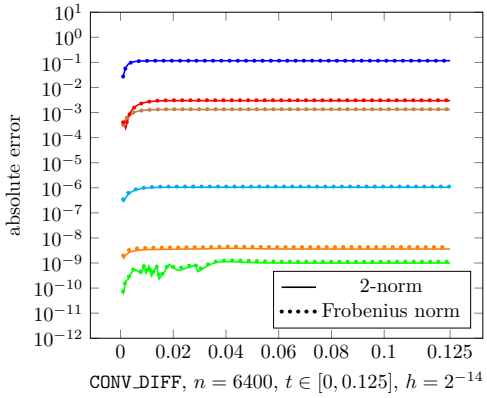


Fig. 35. Absolute error of the splitting scheme approximation.

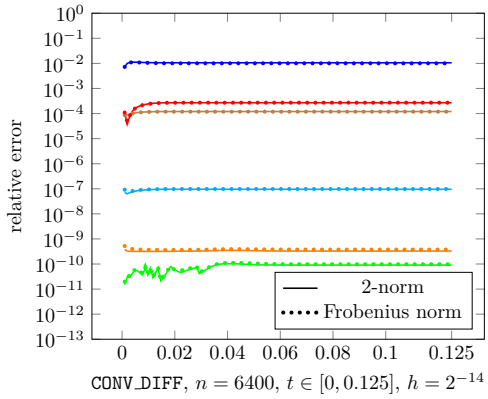


Fig. 36. Relative error of the splitting scheme approximation.

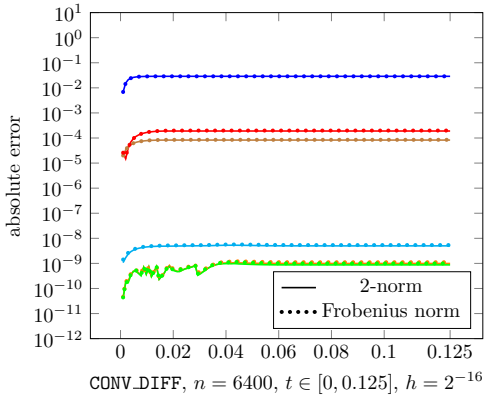


Fig. 37. Absolute error of the splitting scheme approximation.

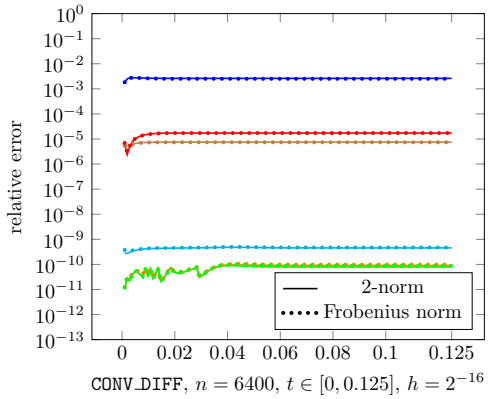


Fig. 38. Relative error of the splitting scheme approximation.