Hamburger Beiträge
zur Angewandten Mathematik

The ADI method for algebraic Riccati equations

Arash Massoudi, Mark R. Opmeer, Timo Reis

Nr. 2014-16
September 2014

THE ADI METHOD FOR THE ALGEBRAIC RICCATI EQUATION

ARASH MASSOUDI*‡, MARK R. OPMEER†§, AND TIMO REIS*¶

Abstract. We consider the recently published ADI method for algebraic Riccati equations. We present a new perspective on this algorithm in terms of the underlying linear-quadratic optimal control problem. This gives rise to a convergence proof. We also consider the rational Krylov–Galerkin method from the viewpoint of linear-quadratic optimal control. Thereby we can compare the approximate solutions computed by ADI and rational Krylov–Galerkin in terms of semi-definiteness.

Key words. algebraic Riccati equation, ADI iteration, numerical method in control theory, linear-quadratic optimal control

AMS subject classifications. 15A24, 49N10, 47J20, 65F30, 49M30, 93B52, 65K10

1. Introduction. We consider an algorithm for the approximation of the unique nonnegative definite solution of the algebraic Riccati equation

$$A^*X + XA + C^*C - XBB^*X = 0, \tag{1.1}$$

where $A \in \mathbb{C}^{n\times n}$ is stable (i.e. all its eigenvalues are in the open left half-plane), $B \in \mathbb{C}^{n\times m}$ and $C \in \mathbb{C}^{p\times n}$.

This algorithm is equivalent to the one recently obtained in [7]. However, our derivation of the algorithm is very different and this new perspective gives important properties of the algorithm not obtained in [7]. We show in particular that this algorithm has a descriptive interpretation in terms of the underlying linear-quadratic optimal control problem. This gives rise to a convergence proof (which was lacking in [7]). We also consider the relation between this algorithm and the rational Krylov–Galerkin method for approximating the solution of the algebraic Riccati equation (1.1), also obtaining some results not provided in [7].

The considered algorithm is iterative in nature and at step k produces an approximate solution of the form $X_k = S_k T_k^{-1} S_k^*$, where $S_k \in \mathbb{C}^{n\times kp}$ and $T_k \in \mathbb{C}^{kp\times kp}$ is positive definite. The main computational cost in the algorithm consists of, at each iteration step, solving a linear system of the form $(\alpha - A)x = v$, where $v \in \mathbb{C}^{n\times p}$ and $\alpha \in \mathbb{C}$ with $\mathrm{Re}(\alpha) > 0$. These features make this algorithm attractive for the case where n is large, p is small and A is sparse. This situation arises for example when considering discretizations of partial differential equations. In fact, our analysis extends to the case where the coefficients in the algebraic Riccati equation are operators with suitable properties and includes the case of partial differential equations rather than only their discretizations. However, in this introduction we will restrict ourselves to the matrix case.

In the case where the algebraic Riccati equation (1.1) reduces to a Lyapunov equation (i.e. when $B = 0$), the considered algorithm reduces to the Alternating Direction Implicit (ADI) method in its factored algorithmic form [6]. We will therefore refer to this algorithm as the *Riccati-ADI* method.

*Fachbereich Mathematik, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany.

†Department of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom.

‡arash.massoudi@uni-hamburg.de

§m.opmeer@maths.bath.ac.uk

¶timo.reis@uni-hamburg.de

The algorithm for this Riccati-ADI method is given in Algorithm 1 in Section 7. We now describe our motivation behind this method and give the important properties of this method. To this end, we first relate the unique nonnegative definite solution of the algebraic Riccati equation (1.1) to an optimal control problem and then give an explicit formula for this solution.

It is well-known that the algebraic Riccati equation (1.1) is intimately connected to the following optimal control problem: for $x_0 \in \mathbb{C}^n$ find

$$\inf_{u\in L^2(0,\infty;\mathbb{C}^m)} \int_0^\infty \|u(t)\|^2 + \|y(t)\|^2\, dt, \tag{1.2}$$

where

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad y(t) = Cx(t), \tag{1.3}$$

as the above infimum is given by $\langle Xx_0, x_0\rangle$, where X is the unique nonnegative definite solution of the algebraic Riccati equation (1.1). We note that our analysis of the Riccati-ADI method is completely based on the optimal control problem and not on the algebraic Riccati equation (1.1).

We will also make extensive use of an explicit formula for X. To give that explicit formula, we first associate the following maps to the dynamical system (1.3):

- the *output map* $\Psi : \mathbb{C}^n \to L^2(0,\infty;\mathbb{C}^p)$ which maps the initial state x_0 to the output y (for control $u = 0$),

  $$\Psi x_0 = t \mapsto Ce^{At}x_0, \tag{1.4}$$

  with adjoint $\Psi^* : L^2(0,\infty;\mathbb{C}^p) \to \mathbb{C}^n$ given by

  $$\Psi^* y = \int_0^\infty e^{A^*\tau} C^* y(\tau)\, d\tau,$$

- the *input-output map* $\mathbb{F} : L^2(0,\infty;\mathbb{C}^m) \to L^2(0,\infty;\mathbb{C}^p)$ which maps the input u to the output y (for initial condition $x_0 = 0$),

  $$\mathbb{F}u = t \mapsto \int_0^t Ce^{A(t-\tau)}Bu(\tau)\, d\tau, \tag{1.5}$$

  with adjoint $\mathbb{F}^* : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^m)$ given by

  $$\mathbb{F}^* y = t \mapsto \int_t^\infty B^* e^{A^*(\tau-t)} C^* y(\tau)\, d\tau,$$

- the *complimentary Popov operator* $\mathcal{R}_c : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ defined by

  $$\mathcal{R}_c := I + \mathbb{F}\mathbb{F}^* = \begin{bmatrix} I & \mathbb{F} \end{bmatrix} \begin{bmatrix} I \\ \mathbb{F}^* \end{bmatrix}. \tag{1.6}$$

We note that the complimentary Popov operator is bounded, self-adjoint, positive definite, and has a bounded inverse. We have that the unique nonnegative definite solution of the algebraic Riccati equation (1.1) is given by (see Lemma 4.1)

$$X = \Psi^* \mathcal{R}_c^{-1} \Psi, \tag{1.7}$$

and that the (unique) optimal control for (1.2) & (1.3) is given by

$$u^{\mathrm{opt}} = -\mathbb{F}^* \mathcal{R}_c^{-1} \Psi x_0.$$

To relate the matrix calculated by Riccati-ADI to a similar optimal control problem and a similar explicit formula, we need to introduce a subspace of $L^2(0,\infty)$ and give some of its properties. For a sequence $(\alpha_j)_{j=1}^{\infty}$ with $\alpha_j \in \mathbb{C}$ with $\mathrm{Re}(\alpha_j) > 0$ define for $k \in \mathbb{N}$

$$\mathscr{V}_k := \mathrm{span}\{t \mapsto \mathrm{e}^{-\alpha_1 t}, \ldots, t \mapsto \mathrm{e}^{-\alpha_k t}\}. \tag{1.8}$$

In this introduction we assume for notational simplicity that the "shift parameters" α_j are distinct (in the main part of the article we drop this assumption; the definition of $\mathscr{V}_k$ has to be modified in case of non-distinct parameters). We show that $\mathscr{V}_k$ is not just any subspace of $L^2(0,\infty)$, but a rational Krylov subspace (Lemma 2.8). The image of $\mathscr{V}_k \otimes \mathbb{C}^p$ under the adjoint of the output map is also a rational Krylov subspace (of $\mathbb{C}^n$):

$$\mathscr{X}_k := \Psi^*(\mathscr{V}_k \otimes \mathbb{C}^p) = \sum_{j=1}^{k} \mathrm{ran}(\alpha_j - A^*)^{-1} C^*, \tag{1.9}$$

(see Lemma 2.12 and Remark 2.13).

Let $P_k : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ denote the orthogonal projection onto $\mathscr{V}_k \otimes \mathbb{C}^p$. The operator computed by Riccati-ADI gives the optimal cost for the optimal control problem (see Lemma 4.2): for $x_0 \in \mathbb{C}^n$ find

$$\inf_{u \in L^2(0,\infty;\mathbb{C}^m)} \int_0^\infty \|u(t)\|^2 + \|(P_k y)(t)\|^2 \, dt, \tag{1.10}$$

subject to (1.3) and is explicitly given by

$$X_k = \Psi_k^* \mathcal{R}_{c,k}^{-1} \Psi_k, \tag{1.11}$$

where

$$\Psi_k : \mathbb{C}^n \to L^2(0,\infty;\mathbb{C}^p), \qquad \Psi_k = P_k \Psi, \tag{1.12}$$

$$\mathbb{F}_k : L^2(0,\infty;\mathbb{C}^m) \to L^2(0,\infty;\mathbb{C}^p), \qquad \mathbb{F}_k = P_k \mathbb{F}, \tag{1.13}$$

$$\mathcal{R}_{c,k} : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p), \qquad \mathcal{R}_{c,k} = I + \mathbb{F}_k \mathbb{F}_k^* = \begin{bmatrix} I & \mathbb{F}_k \end{bmatrix} \begin{bmatrix} I \\ \mathbb{F}_k^* \end{bmatrix}. \tag{1.14}$$

The (unique) optimal control for (1.10) & (1.3) is given by

$$u_k^{\mathrm{opt}} = -\mathbb{F}_k^* \mathcal{R}_{c,k}^{-1} \Psi_k x_0, \tag{1.15}$$

from which we in particular conclude (Corollary 4.3) that $u_k^{\mathrm{opt}} \in \mathscr{V}_k \otimes \mathbb{C}^m$. This implies that in (1.10) we can equivalently infimize over $\mathscr{V}_k \otimes \mathbb{C}^m$ rather than all of $L^2(0,\infty;\mathbb{C}^m)$.

It follows from $\mathscr{V}_k \subset \mathscr{V}_{k+1}$ that (Theorem 4.4)

$$X_k \le X_{k+1}, \quad X_k \le X,$$

i.e. $(X_k)_{k=1}^{\infty}$ is a non-decreasing sequence bounded from above by X. It follows that $(X_k)_{k=1}^{\infty}$ converges, but the limit may not necessarily equal X.

From the explicit formula for X_k, we obtain (Theorem 5.1) that Riccati-ADI converges to X, i.e.

$$\lim_{k\to\infty} X_k = X, \tag{1.16}$$

provided that the sequence of parameters satisfies the *non-Blaschke condition*

$$\sum_{j=1}^{\infty} \frac{\operatorname{Re}(\alpha_j)}{1+|\alpha_j|^2} = \infty. \tag{1.17}$$

We note that this non-Blaschke condition is for example satisfied if the parameters all belong to a fixed compact set contained in the open right half-plane. This convergence result was previously obtained for the special case of the Lyapunov equation in [9]. When the state space is infinite-dimensional rather than $\mathbb{C}^n$, we have to specify in what topology the convergence (1.16) takes place; such results are presented in Theorem 5.2.

As noted above, the approximate solution X_k obtained using the Riccati-ADI method is identical to that obtained in [7]. That the sequence X_k is non-decreasing is also obtained in [7, Theorem 4.2] by using very different arguments. Convergence of X_k to X is not obtained in [7]. Some estimates for the distance between X_k and X in the gap metric were considered in [7]. These are based on the shift parameters and the eigenvalues of the associated Hamiltonian matrix.

1.1. Comparison with the rational Krylov–Galerkin method. The dynamical system (1.3) can be approximated by using a Galerkin method. Let $\mathscr{W}_q \subset \mathbb{C}^n$ be a q-dimensional subspace and let $W_q \in \mathbb{C}^{n\times q}$ be such that the columns of W_q span $\mathscr{W}_q$ and $W_q^* W_q = I$. Define

$$A_q := W_q^* A W_q, \quad B_q := W_q^* B, \quad C_q := C W_q.$$

We then consider the infimization problem (1.2), where now the dynamical system is

$$\dot{x}_q(t) = A_q x_q(t) + B_q u_q(t), \quad x_q(0) = W_q^* x_0, \quad y_q(t) = C_q x_q(t). \tag{1.18}$$

The approximation to the state x of the original system (1.3) is $W_q x_q$. As before, the optimal control problem is related to the algebraic Riccati equation (1.1), but now with coefficients A_q, B_q and C_q. If A is dissipative (i.e. $A + A^* < 0$), then its Galerkin approximation A_q is also dissipative and therefore stable. The algebraic Riccati equation associated to (1.2) & (1.18) therefore has a unique nonnegative definite solution which as before gives the optimal cost for (1.2) & (1.18). If q is small, then this algebraic Riccati equation can be solved using a direct method. Denote this solution by $X_q^G \in \mathbb{C}^{q\times q}$. Then $\widetilde{X}_q := W_q^* X_q^G W_q \in \mathbb{C}^{n\times n}$ is an approximation of X. Trivially, $\widetilde{X}_n = X$.

The Galerkin space $\mathscr{W}_q$ can be chosen to be the rational Krylov subspace $\mathscr{X}_k$ from (1.9). The resulting method has been extensively studied, especially in the special case where the algebraic Riccati equation reduces to a Lyapunov equation (see for example the review articles [1] and [11]).

We compare the methods Riccati-ADI and rational Krylov–Galerkin. We show (Lemma 6.1) that $P_k y = P_k y_q$ for initial condition x_0 in the rational Krylov subspace $\mathscr{X}_k$ and input $u \in L^2(0,\infty;\mathbb{C}^m)$, where y is the output of the original system (1.3),

y_q is the output of the rational Krylov–Galerkin approximation (1.18) and P_k is the orthogonal projection introduced earlier. From this it follows that

$$X_k \leq \widetilde{X}_k,$$

i.e., the rational Krylov–Galerkin approximation is always larger than or equal to the Riccati-ADI approximation with the same shift parameters (Theorem 6.3).

We now consider when $X_k = \widetilde{X}_k$, i.e. when the rational Krylov–Galerkin approximation and the Riccati-ADI approximation with the same shift parameters coincide. This happens if and only if for all $x_0 \in \mathscr{X}_k$ we have that y_q^{opt} belongs to $\mathscr{V}_k \otimes \mathbb{C}^p$. By classical linear-quadratic optimal control theory (e.g., cf. [15, Chapter 14]), the optimal output of (1.2) & (1.18) for $x_0 \in \mathscr{X}_k$ is given by

$$y_q^{\text{opt}}(t) = C_q e^{A_q^{\text{opt}} t} x_0,$$

where $A_q^{\text{opt}} := A_q - B_q B_q^* X_q^G$. It follows that if the eigenvalues of $-A_q^{\text{opt}}$ are shift parameters, then the optimal output belongs to $\mathscr{V}_k \otimes \mathbb{C}^p$. That this condition is also necessary (if X_q^G is positive definite) follows from considering a (generalized) eigenvector of A_q^{opt} as x_0. See Theorem 6.4 for the details.

That $X_k = \widetilde{X}_k$ if and only if the eigenvalues of $-A_q^{\text{opt}}$ are shift parameters is also obtained in [7, Theorem 4.4] (using a very different argument and with the additional assumption that $p = 1$). The inequality $X_k \leq \widetilde{X}_k$ is not obtained in [7].

The remainder of this article is organized as follows. Section 2 considers the canonical rational Krylov subspace (the appropriate generalization of the space $\mathscr{V}_k$ from (1.8)) and how the operators Ψ^* and $\mathbb{F}^*$ act on various bases of this space. This is used in Section 3 to determine matrix representations of Ψ_k^* and $\mathbb{F}_k^*$. Section 4 relates the operators Ψ^*, $\mathbb{F}^*$, Ψ_k^* and $\mathbb{F}_k^*$ to the already mentioned optimal control problems. Using this connection, convergence of Riccati-ADI is shown in Section 5. Section 6 considers the connection between Riccati-ADI and the rational Krylov–Galerkin method. The algorithm for Riccati-ADI, together with some remarks regarding its implementation, is given in Section 7. Finally, Section 8 illustrates the obtained results using two numerical examples: one arising from a convection-diffusion equation and one arising from an Euler–Bernoulli beam equation.

2. Rational Krylov subspaces. In this section we consider rational Krylov subspaces. In particular, we study what we call the "canonical rational Krylov subspace" for a given sequence of parameters $(\alpha_j)_{j=1}^{\infty}$. The space $\mathscr{V}_k$ from (1.8) is the particular instance of this canonical rational Krylov subspace when the α_j are distinct. To define this space as a rational Krylov subspace we need to consider an unbounded operator on an infinite-dimensional space rather than just rational Krylov subspaces originating from matrices.

REMARK 2.1. *Below we will consider a densely defined closed linear operator with non-empty resolvent set $T : D(T) \subset \mathscr{Z} \to \mathscr{Z}$ on a Hilbert space $\mathscr{Z}$. For $k \in \mathbb{N}_0$, the domain of T^k with the graph norm is a Hilbert space which we will denote by $\mathscr{Z}_{(k)}$. The operator T restricts to an operator $T_k : \mathscr{Z}_{(k)} \to \mathscr{Z}_{(k)}$ with domain $\mathscr{Z}_{(k+1)}$. The operator T_k is densely defined, closed and has the same resolvent set as T. If we identify $\mathscr{Z}'$ with $\mathscr{Z}$, then the dual of $\mathscr{Z}_{(k)}$ is a Hilbert space which we will denote by $\mathscr{Z}_{(-k)}$. The operator T extends by continuity to a bounded operator $T_{-k-1} : \mathscr{Z}_{(-k)} \to \mathscr{Z}_{(-k-1)}$. Considered as an unbounded operator on $\mathscr{Z}_{(-k-1)}$, the operator T_{-k-1} again has the same resolvent set as T. For $\alpha \in \rho(T)$ and $k \in \mathbb{Z}$, the*

operator $\alpha - T_k$ is a bijection $\mathscr{Z}_{(k+1)} \to \mathscr{Z}_{(k)}$. See e.g. [12, Section 3.6] or [2, Section II.5.a] for details.

We are specifically interested in two cases. In the first case $\mathscr{Z} = \mathbb{C}^n$ in which case $\mathscr{Z}_{(k)} = \mathbb{C}^n$ for all $k \in \mathbb{Z}$. In the second case $\mathscr{Z} = L^2(\mathbb{R})$ and T is the first derivative operator. Then $\mathscr{Z}_{(k)}$ equals the Sobolev space $H^k(\mathbb{R})$ for $k \in \mathbb{Z}$, and T_k also equals the first derivative.

DEFINITION 2.2. *Let $\mathscr{Z}$ be a Hilbert space. Let $T : D(T) \subset \mathscr{Z} \to \mathscr{Z}$ be a densely defined closed linear operator with non-empty resolvent set. Let $b \in \mathscr{Z}_{-1}$ and $(\alpha_j)_{j=1}^{\infty}$ be such that $\alpha_j \in \rho(T)$. The corresponding sequence of* rational Krylov subspaces *(of $\mathscr{Z}$) is defined for $k \in \mathbb{N}$ by*

$$\mathscr{K}_k(T, b, \alpha) := \operatorname{span}\left\{ \left(\prod_{j=1}^{\ell} (\alpha_j - T_{-1})^{-1} \right) b : \ \ell \in \{1, \dots, k\} \right\}.$$

DEFINITION 2.3. *Let $(\alpha_j)_{j=1}^{\infty}$ be such that $\operatorname{Re}(\alpha_j) > 0$ for all j. For $k \in \mathbb{N}$, we define the* canonical rational Krylov subspace *as*

$$\mathscr{K}_k(\alpha) := \mathscr{K}_k(-D, \delta, \alpha),$$

where $D : H^1(\mathbb{R}) \subset L^2(\mathbb{R}) \to L^2(\mathbb{R})$ is the first derivative operator and $\delta \in H^{-1}(\mathbb{R})$ is the Dirac delta.

We will consider two bases of the canonical rational Krylov subspace: one orthonormal and one not orthonormal. We now first introduce these bases and show in Lemma 2.8 that they are indeed bases for the canonical rational Krylov subspace.

DEFINITION 2.4. *Let $(\alpha_j)_{j=1}^{\infty}$ be such that $\operatorname{Re}(\alpha_j) > 0$ for all j. We define the corresponding* convolution system *$(\varphi_j)_{j=1}^{\infty}$, $\varphi_j \in L^2(0, \infty)$ by*

$$\begin{aligned} \varphi_1 &:= t \mapsto \mathrm{e}^{-\alpha_1 t}, \\ \varphi_j &:= \mathrm{e}^{-\alpha_j \cdot} * \varphi_{j-1}, \end{aligned}$$

where $*$ denotes the convolution product, i.e. $(g * h)(t) = \int_0^t g(t-\tau)h(\tau)\, d\tau$.

REMARK 2.5. *Let $(\alpha_j)_{j=1}^{k}$ be a tuple of numbers in the open right complex half plane, let $(\varphi_j)_{j=1}^{k}$ be the corresponding convolution system. Let $\widehat{\varphi}_i$ be the Laplace transform of φ_i.*

a) Since the Laplace transform turns convolution into multiplication, we obtain

$$\widehat{\varphi}_1(s) = \tfrac{1}{s+\alpha_1}, \qquad \widehat{\varphi}_j(s) = \tfrac{1}{s+\alpha_j} \cdot \widehat{\varphi}_{j-1}(s),$$

and therefore

$$\widehat{\varphi}_j(s) = \prod_{\ell=1}^{j} \frac{1}{s + \alpha_\ell}. \tag{2.1}$$

b) Assume that the numbers $q_1, \dots, q_J$ are pairwise different with $\{q_1, \dots, q_J\} = \{\alpha_1, \dots, \alpha_k\}$. Further, let ℓ_j be the number of times in which q_j appears in $(\alpha_j)_{j=1}^{k}$ (thus $k = \ell_1 + \ldots + \ell_J$). Then

$$\operatorname{span}\{\varphi_1, \dots, \varphi_k\} = \bigoplus_{j=1}^{J} \operatorname{span}\left\{ \ t \mapsto t^l \mathrm{e}^{-q_j t} \ \middle| \ l = 0, \dots, \ell_j - 1 \ \right\}.$$

The easiest way to see this is by considering $(\widehat{\varphi}_j)_{j=1}^k$ *and using partial fractions. In particular, if the numbers* $\alpha_1, \dots, \alpha_k$ *are distinct, then*

$$\operatorname{span}\{\varphi_1, \dots, \varphi_k\} = \operatorname{span}\{e^{-\alpha_1 \cdot}, \dots, e^{-\alpha_k \cdot}\}.$$

c) *It follows from b) that, if* $(\tilde{\alpha}_j)_{j=1}^k$ *is a permutation of* $(\alpha_j)_{j=1}^k$ *and* $(\tilde{\varphi}_j)_{j=1}^k$ *and* $(\varphi_j)_{j=1}^k$ *are the corresponding convolution systems, then*

$$\operatorname{span}\{\tilde{\varphi}_1, \dots, \tilde{\varphi}_k\} = \operatorname{span}\{\varphi_1, \dots, \varphi_k\}.$$

DEFINITION 2.6. *Let* $(\alpha_j)_{j=1}^\infty$ *be such that* $\operatorname{Re}(\alpha_j) > 0$ *for all* $j \in \mathbb{N}$. *We define the corresponding* Takenaka–Malmquist system $(\psi_j)_{j=1}^\infty$, $\psi_j \in L^2(0,\infty)$ *by*

$$\begin{aligned} &\phi_1 = t \mapsto \mathrm{e}^{-\alpha_1 t}, && \psi_1 = \sqrt{2\operatorname{Re}(\alpha_1)}\cdot \phi_1, \\ &\phi_j = \phi_{j-1} - (\alpha_j + \overline{\alpha_{j-1}}) \cdot (\mathrm{e}^{-\alpha_j \cdot} * \phi_{j-1}), && \psi_j = \sqrt{2\operatorname{Re}(\alpha_j)}\cdot \phi_j, \end{aligned} \tag{2.2}$$

where $*$ *denotes the convolution product, i.e.* $(g*h)(t) = \int_0^t g(t-\tau)h(\tau)\,d\tau$.

REMARK 2.7.

a) *The Takenaka–Malmquist system is orthonormal (see e.g. [9, Appendix B] for a proof).*

b) *Laplace transformation of* (2.2) *yields that for all* $s \in \mathbb{C}$ *with* $\operatorname{Re}(s) > 0$ *there holds*

$$\begin{aligned} &\widehat{\phi}_1(s) = \tfrac{1}{s+\alpha_1}, && \widehat{\psi}_1(s) = \sqrt{2\operatorname{Re}(\alpha_1)}\cdot \widehat{\phi}_1(s), \\ &\widehat{\phi}_j(s) = \widehat{\phi}_{j-1}(s) - (\alpha_j + \overline{\alpha_{j-1}}) \cdot \tfrac{1}{s+\alpha_j} \cdot \widehat{\phi}_{j-1}(s), && \widehat{\psi}_j(s) = \sqrt{2\operatorname{Re}(\alpha_j)}\cdot \widehat{\phi}_j(s). \end{aligned} \tag{2.3}$$

Therefore we obtain by induction that

$$\widehat{\psi}_j(s) = \frac{\sqrt{2\operatorname{Re}(\alpha_j)}}{(s+\alpha_j)} \cdot \prod_{\ell=1}^{j-1} \frac{s-\overline{\alpha_\ell}}{s+\alpha_\ell}. \tag{2.4}$$

LEMMA 2.8. *Let* $(\alpha_j)_{j=1}^\infty$ *be such that* $\operatorname{Re}(\alpha_j) > 0$ *for all* j. *Let* $(\varphi_j)_{j=1}^\infty$ *be the corresponding convolution system, let* $(\psi_j)_{j=1}^\infty$ *be the corresponding Takenaka–Malmquist system and let* $\mathscr{K}_k(\alpha)$ *be the corresponding sequence of canonical rational Krylov subspaces. Then*

$$\mathscr{K}_k(\alpha) = \operatorname{span}\{\varphi_1, \dots, \varphi_k\} = \operatorname{span}\{\psi_1, \dots, \psi_k\},$$

where we view a function in $L^2(0,\infty)$ *as an element of* $L^2(\mathbb{R})$ *by defining it to be zero on* $(-\infty, 0)$.

Proof. We note that the extension $D_{-1} : L^2(\mathbb{R}) \to H^{-1}(\mathbb{R})$ of the operator defining the canonical rational Krylov subspace is also the first derivative operator.

We have

$$f = (\mu + D_{-1})^{-1}\delta \quad \Longleftrightarrow \quad \mu f + f' = \delta.$$

The unique solution of the latter ordinary differential equation is

$$f = t \mapsto e^{-\mu t} 1_{(0,\infty)}(t) \in L^2(\mathbb{R}).$$

Therefore $\varphi_1 = (\alpha_1 + D_{-1})^{-1}\delta$ and so $\operatorname{span}\{\varphi_1\} = \mathscr{K}_1(\alpha)$. We have

$$f_j = (\alpha_j + D_{-1})^{-1} f_{j-1} \quad \Longleftrightarrow \quad \alpha_j f_j + f_j' = f_{j-1}.$$

Using that, as just shown, $t \mapsto e^{-\alpha_j t} 1_{(0,\infty)}(t)$ is the fundamental solution of this differential equation we have

$$f_j = e^{-\alpha_j \cdot} 1_{(0,\infty)} * f_{j-1}.$$

Hence $\varphi_j = f_j$ and we conclude that $\operatorname{span}\{\varphi_1, \ldots, \varphi_k\} = \mathscr{K}_k(\alpha)$.
The relation $\operatorname{span}\{\varphi_1, \ldots, \varphi_k\} = \operatorname{span}\{\psi_1, \ldots, \psi_k\}$ follows most readily by considering the partial fraction expansions of their Laplace transforms, see (2.1) and (2.4). □

Now we determine how the operators Ψ^* and $\mathbb{F}^*$ act on the considered bases for the canonical rational Krylov subspaces. We first define the following two operators (for $t \geq 0$)

$$\Phi^t : L^2(0,\infty;\mathbb{C}^p) \to \mathbb{C}^n, \quad \Phi^t z := \int_t^\infty \mathrm{e}^{A^*(\tau - t)} C^* z(\tau)\, d\tau, \tag{2.5}$$

$$\Lambda : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^n), \quad \Lambda z := t \mapsto \int_t^\infty \mathrm{e}^{A^*(\tau - t)} C^* z(\tau)\, d\tau. \tag{2.6}$$

The significance of these operators is that $\Psi^* = \Phi^0$, $\mathbb{F}^* = B^*\Lambda$ and $\Lambda z = t \mapsto \Phi^t z$.

The following lemma is the crucial technical result in determining how the operators Ψ^* and $\mathbb{F}^*$ act on the considered bases for the canonical rational Krylov subspaces.

LEMMA 2.9. *Let $A \in \mathbb{C}^{n\times n}$ be stable, let $C \in \mathbb{C}^{p\times n}$ and define for $t \geq 0$ the operator Φ^t by (2.5). Then for $\mu \in \mathbb{C}$ with $\operatorname{Re}(\mu) > 0$, $v \in \mathbb{C}^p$ and $z \in L^2(0,\infty;\mathbb{C}^p)$ there holds*

$$\Phi^t(\mathrm{e}^{-\mu\cdot} v) = (\mu - A^*)^{-1} C^* v \mathrm{e}^{-\mu t}, \tag{2.7}$$

and

$$\Phi^t(\mathrm{e}^{-\mu\cdot} * z) = (\mu - A^*)^{-1} C^* (\mathrm{e}^{-\mu\cdot} * z)(t) + (\mu - A^*)^{-1} \Phi^t(z). \tag{2.8}$$

Proof. We first consider (2.7). We have by the change of variables $\theta := \tau - t$

$$\begin{aligned}\Phi^t(\mathrm{e}^{-\mu\cdot} v) = \int_t^\infty \mathrm{e}^{A^*(\tau-t)} C^* v \mathrm{e}^{-\mu\tau}\, d\tau &= \int_0^\infty \mathrm{e}^{A^*\theta} C^* v \mathrm{e}^{-\mu\theta} \mathrm{e}^{-\mu t}\, d\theta \\ &= \mathrm{e}^{-\mu t} \int_0^\infty \mathrm{e}^{(A^* - \mu)\theta} C^* v\, d\theta,\end{aligned}$$

and elementary integration then gives the result.

We now consider (2.8). We have

$$\begin{aligned}\Phi^t(\mathrm{e}^{-\mu\cdot} * z) &= \int_t^\infty \mathrm{e}^{A^*(\tau-t)} C^* \int_0^\tau \mathrm{e}^{-\mu(\tau-\sigma)} z(\sigma)\, d\sigma d\tau \\ &= \int_t^\infty \int_0^\tau \mathrm{e}^{(\mu - A^*)(t-\tau)} C^* \mathrm{e}^{-\mu(t-\sigma)} z(\sigma)\, d\sigma d\tau.\end{aligned}$$

Interchanging the order of integration gives that the above equals

$$\begin{aligned}
&\int_0^t \int_t^\infty \mathrm{e}^{(\mu-A^*)(t-\tau)} C^* \mathrm{e}^{-\mu(t-\sigma)} z(\sigma)\, d\tau d\sigma \\
&\qquad + \int_t^\infty \int_\sigma^\infty \mathrm{e}^{(\mu-A^*)(t-\tau)} C^* \mathrm{e}^{-\mu(t-\sigma)} z(\sigma)\, d\tau d\sigma \\
&= \int_0^t \Big[-(\mu - A^*)^{-1} \mathrm{e}^{(\mu-A^*)(t-\tau)} C^* \mathrm{e}^{-\mu(t-\sigma)} z(\sigma) \Big]_{\tau=t}^{\infty} d\sigma \\
&\qquad + \int_t^\infty \Big[-(\mu - A^*)^{-1} \mathrm{e}^{(\mu-A^*)(t-\tau)} C^* \mathrm{e}^{-\mu(t-\sigma)} z(\sigma) \Big]_{\tau=\sigma}^{\infty} d\sigma \\
&= \int_0^t (\mu - A^*)^{-1} C^* \mathrm{e}^{-\mu(t-\sigma)} z(\sigma)\, d\sigma + \int_t^\infty (\mu - A^*)^{-1} \mathrm{e}^{(\mu-A^*)(t-\sigma)} C^* \mathrm{e}^{-\mu(t-\sigma)} z(\sigma)\, d\sigma \\
&= (\mu - A^*)^{-1} C^* \int_0^t \mathrm{e}^{-\mu(t-\sigma)} z(\sigma)\, d\sigma + (\mu - A^*)^{-1} \int_t^\infty \mathrm{e}^{A^*(\sigma-t)} C^* z(\sigma)\, d\sigma \\
&= (\mu - A^*)^{-1} C^* \ (\mathrm{e}^{-\mu \cdot} * z)(t) + (\mu - A^*)^{-1} \Phi^t(z),
\end{aligned}$$

as claimed. □

COROLLARY 2.10. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, $(\alpha_j)_{j=1}^\infty$ such that $\mathrm{Re}(\alpha_j) > 0$ for all j, $(\varphi_j)_{j=1}^\infty$ as in Definition 2.4 and $v \in \mathbb{C}^p$.*
a) Let $t \geq 0$. With Φ^t as in (2.5) *there holds*

$$\begin{aligned}
\Phi^t(\varphi_1 v) &= (\alpha_1 - A^*)^{-1} C^* v \varphi_1(t), \\
\Phi^t(\varphi_j v) &= (\alpha_j - A^*)^{-1} C^* v \varphi_j(t) + (\alpha_j - A^*)^{-1} \Phi^t(\varphi_{j-1} v).
\end{aligned}$$

b) With Ψ as in (1.4) *there holds*

$$\begin{aligned}
\Psi^*(\varphi_1 v) &= (\alpha_1 - A^*)^{-1} C^* v, \\
\Psi^*(\varphi_j v) &= (\alpha_j - A^*)^{-1} \Psi^*(\varphi_{j-1} v).
\end{aligned}$$

c) With Λ as in (2.6) *there holds*

$$\begin{aligned}
\Lambda(\varphi_1 v) &= (\alpha_1 - A^*)^{-1} C^* v \varphi_1, \\
\Lambda(\varphi_j v) &= (\alpha_j - A^*)^{-1} C^* v \varphi_j + (\alpha_j - A^*)^{-1} \Lambda(\varphi_{j-1} v).
\end{aligned}$$

Proof. We first prove part a). The first formula follows directly from (2.7) with $\mu := \alpha_1$. The second formula follows from multiplying the iterative definition of $(\varphi_j)_{j=1}^\infty$ from Definition 2.4 by v, applying Φ^t to the result and using that by Lemma 2.9,

$$\Phi^t(\mathrm{e}^{-\alpha_j \cdot} * \varphi_{j-1} v) = (\alpha_j - A^*)^{-1} C^* v \varphi_j(t) + (\alpha_j - A^*)^{-1} \Phi^t(\varphi_{j-1} v).$$

Part b) follows from part a) by using that $\Psi^* = \Phi^0$, $\varphi_1(0) = 1$ and $\varphi_j(0) = 0$ for $j > 1$. Part c) follows from part a) using that $\Lambda z = t \mapsto \Phi^t z$. □

COROLLARY 2.11. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, $(\alpha_j)_{j=1}^\infty$ such that $\mathrm{Re}(\alpha_j) > 0$ for all j, $(\phi_j)_{j=1}^\infty$ and $(\psi_j)_{j=1}^\infty$ as in Definition 2.6 and $v \in \mathbb{C}^p$.*
a) With Ψ as in (1.4) *there holds*

$$\begin{aligned}
\Psi^*(\phi_1 v) &= (\alpha_1 - A^*)^{-1} C^* v, \\
\Psi^*(\phi_j v) &= \Psi^*(\phi_{j-1} v) - (\alpha_j + \overline{\alpha_{j-1}})(\alpha_j - A^*)^{-1} \Psi^*(\phi_{j-1} v).
\end{aligned}$$

b) For $j > 1$ and with Λ as in (2.6) and $\gamma_j := \frac{\operatorname{Re}(\alpha_j)}{\operatorname{Re}(\alpha_{j-1})}$ there holds

$$\Lambda(\psi_j v) = \gamma_j \Lambda(\psi_{j-1} v) - \gamma_j(\alpha_j + \overline{\alpha_{j-1}}) \cdot \left[(\alpha_j - A^*)^{-1} C^* v \mathrm{e}^{-\alpha_j \cdot} * \psi_{j-1} + (\alpha_j - A^*)^{-1} \Lambda(\psi_{j-1} v)\right].$$

Proof. We first prove part a). The first equation follows from (2.7) with $\mu := \alpha_1$ using that $\Psi^* = \Phi^0$. The second equation is obtained by multiplying (2.2) by v, applying Ψ^* to the result and using that by Lemma 2.9 (using that $\Psi^* = \Phi^0$),

$$\Psi^*(\mathrm{e}^{-\alpha_j \cdot} * \phi_{j-1} v) = (\alpha_j - A^*)^{-1} \Psi^*(\phi_{j-1} v). \tag{2.9}$$

We now prove part b). From (2.2) we obtain

$$\Lambda(\psi_j v) = \gamma_j \Lambda(\psi_{j-1} v) - \gamma_j(\alpha_j + \overline{\alpha_{j-1}}) \Lambda(\mathrm{e}^{-\alpha_j \cdot} * \psi_{j-1} v).$$

From Lemma 2.9 we obtain that

$$\Lambda(\mathrm{e}^{-\alpha_j \cdot} * \psi_{j-1} v) = (\alpha_j - A^*)^{-1} C^* v \mathrm{e}^{-\alpha_j \cdot} * \psi_{j-1} + (\alpha_j - A^*)^{-1} \Lambda(\psi_{j-1} v),$$

and the desired result follows. □

The following lemma and the subsequent remark show in what sense the canonical rational Krylov subspace is canonical: the other rational Krylov subspaces with the same shift parameters can be obtained from it.

LEMMA 2.12. *Let $(\alpha_j)_{j=1}^{\infty}$ be such that $\operatorname{Re}(\alpha_j) > 0$ for all j and let $k \in \mathbb{N}$. Let $\mathscr{K}_k(\alpha)$ be the canonical rational Krylov subspace from Definition 2.3. Let $T \in \mathbb{C}^{n \times n}$ be stable, $b \in \mathbb{C}^n$ and let $\mathscr{K}_k(T, b, \alpha)$ be the rational Krylov subspace from Definition 2.2. With*

$$\Phi : L^2(0, \infty) \to \mathbb{C}^n, \quad \Phi u := \int_0^\infty e^{Tt} b u(t) dt,$$

there holds

$$\Phi\left(\mathscr{K}_k(\alpha)\right) = \mathscr{K}_k(T, b, \alpha).$$

Proof. From Corollary 2.10 b) with $A = T^*$, $v = 1$ and $C^* = b$, we have that for $(\varphi_j)_{j=1}^k$ as in Definition 2.4 there holds

$$\Phi(\varphi_1) = (\alpha_1 - T)^{-1} b,$$
$$\Phi(\varphi_j) = (\alpha_j - T)^{-1} \Phi(\varphi_{j-1}).$$

From this and the definition of rational Krylov subspace (Definition 2.2) we have

$$\mathscr{K}_k(T, b, \alpha) = \Phi(\operatorname{span}\{\varphi_1, \ldots, \varphi_k\}).$$

By Lemma 2.8 we have that

$$\operatorname{span}\{\varphi_1, \ldots, \varphi_k\} = \mathscr{K}_k(\alpha),$$

so that we obtain the desired conclusion. □

REMARK 2.13. *More generally, for $b \in \mathbb{C}^{n \times m}$, the operator $\Phi : L^2(0, \infty; \mathbb{C}^m) \to \mathbb{C}^n$ defined as in Lemma 2.12 satisfies*

$$\Phi(\mathscr{K}_k(\alpha) \otimes \mathbb{C}^m) = \operatorname{span}\{\mathscr{K}_k(T, bv, \alpha) : v \in \mathbb{C}^m\}.$$

As in the proof of Lemma 2.12, we obtain for $v \in \mathbb{C}^m$

$$\mathscr{K}_k(T, bv, \alpha) = \Phi(\operatorname{span}\{\varphi_1 v, \ldots, \varphi_k v\}),$$

from which the result follows. In the case where the α_j *are distinct, this gives* (1.9) *from the introduction.*

PROPOSITION 2.14. *Let* $A \in \mathbb{C}^{n\times n}$ *be stable,* $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$, $(\alpha_j)_{j=1}^{\infty}$ *such that* $\operatorname{Re}(\alpha_j) > 0$ *for all* j *and* $\mathscr{K}_k(\alpha)$ *the sequence of canonical rational Krylov subspaces from Definition 2.3. For* $k \in \mathbb{N}$, *define*

$$\mathscr{X}_k := \Psi^*(\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p), \tag{2.10}$$

where Ψ *is as in* (1.4).
a) The following holds for all $j \in \{1, \ldots, k\}$ *and* $v \in \mathbb{C}^p$*:*

$$(\alpha_j - A^*)^{-1} C^* v \in \mathscr{X}_k.$$

b) The following holds for all $k \in \mathbb{N}$*:*

$$(\alpha_{k+1} - A^*)^{-1} \mathscr{X}_k \subset \mathscr{X}_{k+1}.$$

c) With $t \geq 0$ *and* Φ^t *as in* (2.5)*, the following holds for all* $k \in \mathbb{N}$*:*

$$\Phi^t\left(\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p\right) \subset \mathscr{X}_k.$$

d) With $\mathbb{F}$ *as in* (1.5) *the following holds for all* $k \in \mathbb{N}$*:*

$$\mathbb{F}^*\left(\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p\right) \subset \mathscr{K}_k(\alpha) \otimes \mathbb{C}^m.$$

Proof. We first prove a). We have by (2.7) that $(\alpha_j - A^*)^{-1} C^* v = \Psi^*(e^{-\alpha_j \cdot} v)$. Since by Remark 2.5 b) $e^{-\alpha_j \cdot} \in \mathscr{K}_k(\alpha)$, it follows that $(\alpha_j - A^*)^{-1} C^* v \in \mathscr{X}_k$.

We now prove b). Since $(\varphi_j)_{j=1}^k$ is a basis for $\mathscr{K}_k(\alpha)$, every element of $\mathscr{X}_k$ can be written as a linear combination of elements of the form

$$\Psi^*(\varphi_\ell v),$$

for $v \in \mathbb{C}^p$ and $\ell \in \{1, \ldots, k\}$. We show that

$$(\alpha_{k+1} - A^*)^{-1} \Psi^*(\varphi_\ell v) \subset \mathscr{X}_{k+1}. \tag{2.11}$$

Let $(\tilde{\alpha}_j)_{j=1}^{k+1}$ be a permutation of $(\alpha_j)_{j=1}^{k+1}$ such that $\tilde{\alpha}_j = \alpha_j$ for $j \in \{1, \ldots, \ell\}$ and $\tilde{\alpha}_{\ell+1} = \alpha_{k+1}$. Define for $j = 1, \ldots, k+1$

$$\widetilde{\mathscr{X}_j} := \Psi^*(\mathscr{K}_j(\tilde{\alpha}) \otimes \mathbb{C}^p).$$

By Remark 2.5 c) we have $\mathscr{K}_{k+1}(\alpha) = \mathscr{K}_{k+1}(\tilde{\alpha})$, so that $\widetilde{\mathscr{X}_{k+1}} = \mathscr{X}_{k+1}$. Define $(\tilde{\varphi}_j)_{j=1}^{k+1}$ as in Definition 2.4, but with parameters $\tilde{\alpha}_j$ rather than α_j. Since $\tilde{\alpha}_j = \alpha_j$ for $j \in \{1, \ldots, \ell\}$ we also have $\tilde{\varphi}_j = \varphi_j$ for $j \in \{1, \ldots, \ell\}$. Therefore (2.11) is equivalent to

$$(\tilde{\alpha}_{\ell+1} - A^*)^{-1} \Psi^*(\tilde{\varphi}_\ell v) \subset \mathscr{X}_{k+1}.$$

From Corollary 2.10 b) we have

$$\Psi^*(\tilde{\varphi}_{\ell+1} v) = (\tilde{\alpha}_{\ell+1} - A^*)^{-1} \Psi^*(\tilde{\varphi}_\ell v).$$

Since the left-hand side is in $\mathscr{X}_{k+1}$, the right-hand side is as well.

We prove part c) by induction on k. Let $(\varphi_j)_{j=1}^{\infty}$ be as in Definition 2.4 and let $v \in \mathbb{C}^p$. For $k = 1$ we have by Corollary 2.10 parts a) and b)

$$\Phi^t(\varphi_1 v) = \Psi^*(\varphi_1 v)\varphi_1(t),$$

which proves the case $k = 1$.

For $k > 1$ we have by Corollary 2.10 a) for $j \in \{1, \dots, k\}$

$$\Phi^t(\varphi_j v) = (\alpha_j - A^*)^{-1} C^* v \varphi_j(t) + (\alpha_j - A^*)^{-1} \Phi^t(\varphi_{j-1} v).$$

By the induction hypothesis we have $\Phi^t(\varphi_{j-1}v) \in \mathscr{X}_{j-1}$. Part b leads to $(\alpha_j - A^*)^{-1}\Phi^t(\varphi_{j-1}v) \in \mathscr{X}_j \subset \mathscr{X}_k$. From part a we obtain that $(\alpha_j - A^*)^{-1}C^*v \in \mathscr{X}_j \subset \mathscr{X}_k$. This proves that $\Phi^t(\varphi_j v) \in \mathscr{X}_k$.

We now prove part d). From Corollary 2.10 c), using that $(\varphi_j)_{j=1}^k$ is a basis for $\mathscr{K}_k(\alpha)$, we obtain by induction $\Lambda\left(\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p\right) \subset \mathscr{K}_k(\alpha) \otimes \mathbb{C}^n$. Using that $\mathbb{F}^* = B^*\Lambda$ then gives the desired result. □

3. Matrix representations. In this section we develop matrix representations of the operators Ψ_k and $\mathbb{F}_k$ from the introduction with respect to the Takenaka–Malmquist system from Definition 2.6.

DEFINITION 3.1. *Let $(\alpha_j)_{j=1}^{\infty}$ be such that $\operatorname{Re}(\alpha_j) > 0$ for all $j \in \mathbb{N}$. Let $(\psi_j)_{j=1}^{\infty}$, $\psi_j \in L^2(0,\infty)$ be the corresponding Takenaka–Malmquist system* (2.2)*. For $k \in \mathbb{N}$, the mapping ι_k is defined by*

$$\begin{aligned} \iota_k : \quad & \mathbb{C}^k \to L^2(0,\infty), \\ & x \mapsto \sum_{j=1}^{k} x_j \cdot \psi_j. \end{aligned} \tag{3.1}$$

Further, for the identity matrix $I_p \in \mathbb{C}^{p\times p}$, we identify $\iota_k : \mathbb{C}^{kp} \to L^2(0,\infty;\mathbb{C}^p)$ with the tensor product $\iota_k \otimes I_p$. We omit an additional subindex for sake of brevity.

It follows immediately from the orthonormality of the Takenaka–Malmquist system that ι_k defines an isometric embedding. In particular, the operator

$$P_k = \iota_k \iota_k^* : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$$

is the orthogonal projection onto $\mathscr{V}_k \otimes \mathbb{C}^p$. With operators Ψ and $\mathbb{F}$ as in (1.4) and (1.5), we define the matrices

$$S_k = \iota_k^* \Psi \in \mathbb{C}^{kp\times n}, \tag{3.2}$$

$$F_k = \iota_k^* \mathbb{F} \iota_k \in \mathbb{C}^{kp\times km}, \tag{3.3}$$

$$R_{c,k} = \iota_k^*(I + \mathbb{F}\mathbb{F}^*)\iota_k \in \mathbb{C}^{kp\times kp}. \tag{3.4}$$

It follows from (1.12) that

$$\Psi_k = P_k \Psi = \iota_k \iota_k^* \Psi = \iota_k S_k.$$

We conclude that the matrix S_k as in (3.2) is the matrix representation of $\Psi_k : \mathbb{C}^n \to \mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$ with respect to the basis given by the tensor product of $\{\psi_1, \dots, \psi_k\}$ and the canonical basis of $\mathbb{C}^p$.

With the matrix F_k as in (3.3) and $\mathbb{F}_k$ as in (1.13) we have

$$\iota_k F_k = P_k \mathbb{F} \iota_k = \mathbb{F}_k \iota_k,$$

which shows that F_k is the matrix representation of $\mathbb{F}_k|_{\mathscr{K}_k(\alpha)\otimes\mathbb{C}^m} : \mathscr{K}_k(\alpha)\otimes\mathbb{C}^m \to \mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$ with respect to the basis given by the tensor product of $\{\psi_1,\dots,\psi_k\}$ and the canonical basis of $\mathbb{C}^m$ (respectively, $\mathbb{C}^p$).

By Proposition 2.14 d) we have

$$P_k \mathbb{F}^* \iota_k = \mathbb{F}^* \iota_k.$$

It follows that

$$\begin{aligned} R_{c,k} &= I_{kp} + (\mathbb{F}^*\iota_k)^* \cdot (\mathbb{F}^*\iota_k) \\ &= I_{kp} + (\mathbb{F}^*\iota_k)^* P_k (\mathbb{F}^*\iota_k) \\ &= I_{kp} + (\mathbb{F}^*\iota_k)^* \iota_k (\iota_k^* \mathbb{F}^* \iota_k) \\ &= I_{kp} + (\iota_k^*\mathbb{F}^*\iota_k)^* \cdot (\iota_k^*\mathbb{F}^*\iota_k) \\ &= I_{kp} + F_k F_k^*. \end{aligned} \tag{3.5}$$

Note that, by

$$P_k (I + \mathbb{F}_k \mathbb{F}_k^*)^{-1} P_k = \iota_k (I + F_k F_k^*)^{-1} \iota_k^*,$$

we see that X_k as in (1.11) can be written as

$$X_k = \Psi_k^* \mathcal{R}_{c,k}^{-1} \Psi_k = S_k^* (I + F_k F_k^*)^{-1} S_k. \tag{3.6}$$

By the same argumentation, we see that u_k^{opt} as in (1.15) reads

$$u_k^{\mathrm{opt}} = -\mathbb{F}_k^* \mathcal{R}_{c,k}^{-1} \Psi_k x_0 = -\iota_k F_k^* R_{c,k}^{-1} S_k x_0. \tag{3.7}$$

Further details on the connection to the optimal control problem will be given in Section 4.

We now proceed to develop an algorithm for efficiently computing S_k and F_k. The algorithm for computation of S_k is straightforward. Corollary 2.11 a) together with the definition of the Takeneka-Malmquist system implies that

$$S_k = \begin{bmatrix} \sqrt{2\mathrm{Re}(\alpha_1)} \cdot V_1 & \dots & \sqrt{2\mathrm{Re}(\alpha_k)} \cdot V_k \end{bmatrix}^*,$$

where the sequence (V_k) is recursively defined by

$$V_1 = (\alpha_1 - A^*)^{-1} C^*, \qquad V_k = V_{k-1} - (\alpha_k + \overline{\alpha_{k-1}}) \cdot (\alpha_k - A^*)^{-1} V_{k-1}. \tag{3.8}$$

We note that this was already established in [9], where the case $B = 0$ (for which the Riccati equation becomes a Lyapunov equation) was considered.

Using that, by Proposition 2.14 d), the invariance $\mathbb{F}^* (\mathscr{K}_{k-1}(\alpha)\otimes\mathbb{C}^p) \subset \mathscr{K}_{k-1}(\alpha)\otimes \mathbb{C}^m$ holds true, we see that

$$\iota_k \mathbb{F}^* \iota_{k-1} = \begin{bmatrix} F_{k-1}^* \\ 0 \end{bmatrix}.$$

Thus we obtain that F_k has the form

$$F_k = \begin{bmatrix} [F_{k-1}, 0] \\ N_k \end{bmatrix}, \tag{3.9}$$

for some $N_k \in \mathbb{C}^{p\times km}$. Note that N_k is determined by $\mathbb{F}^*(\psi_k v)$ for $v \in \mathbb{C}^p$ and that this in turn is determined by $\Lambda(\psi_k v)$. Therefore we first express $\Lambda(\psi_k v)$ in an appropriate form.

LEMMA 3.2. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, $(\alpha_j)_{j=1}^\infty$ such that $\mathrm{Re}(\alpha_j) > 0$ for all j, $(\varphi_j)_{j=1}^\infty$ as in Definition 2.4, $v \in \mathbb{C}^p$, Ψ as in (1.4) and Λ as in (2.6). Then, for each $k \in \mathbb{N}$, there exists some $L_k \in \mathbb{C}^{k\times k}$ such that*

$$\Lambda(\psi_k v) = \sum_{j=1}^{k} \Psi^*(\psi_j v) \sum_{\ell=1}^{k} (L_k)_{j\ell}\psi_\ell \qquad \forall v \in \mathbb{C}^p. \tag{3.10}$$

Moreover, the matrix L_k can be calculated as in Algorithm 1.

Proof. We prove this by induction. For $k = 1$ we have by Corollary 2.10 c) that $\Lambda(\psi_1 v) = \psi_1(\alpha_1 - A^*)^{-1}C^* v$ and by Corollary 2.10 b) that $\Psi^*(\psi_1 v) = \sqrt{2\mathrm{Re}(\alpha_1)}(\alpha_1 - A^*)^{-1}C^* v$. Hence for $k = 1$, (3.10) is satisfied with $L_1 = \frac{1}{\sqrt{2\mathrm{Re}(\alpha_1)}}$.

With $\gamma_k := \sqrt{\frac{\mathrm{Re}(\alpha_k)}{\mathrm{Re}(\alpha_{k-1})}}$ we obtain from Corollary 2.11 b) that

$$\begin{aligned}\Lambda(\psi_k v) = {} & \gamma_k \Lambda(\psi_{k-1} v) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \\ & \cdot \left((\alpha_k - A^*)^{-1}C^* v \cdot (\mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}) + (\alpha_k - A^*)^{-1}\Lambda(\psi_{k-1} v)\right).\end{aligned} \tag{3.11}$$

From (2.7) with $\mu := \alpha_k$ and $t = 0$ (noting that $\Psi^* = \Phi^0$) and (2.9), we have

$$\begin{aligned}(\alpha_k - A^*)^{-1}C^* v &= \Psi^*(\mathrm{e}^{-\alpha_k \cdot} v), \\ (\alpha_k - A^*)^{-1}\Psi^*(\psi_j v) &= \Psi^*(\mathrm{e}^{-\alpha_k \cdot} * \psi_j v) \qquad \forall v \in \mathbb{C}^p, \qquad j = 1, \dots, k-1.\end{aligned} \tag{3.12}$$

By inserting (3.10) and (3.12) in (3.11), we obtain

$$\begin{aligned}\Lambda(\psi_k v) = {} & \gamma_k \Lambda(\psi_{k-1} v) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \\ & \cdot \left(\Psi^*(\mathrm{e}^{-\alpha_k \cdot} v) \cdot (\mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}) + \sum_{j=1}^{k-1} \Psi^*(\mathrm{e}^{-\alpha_k \cdot} * \psi_j v) \sum_{\ell=1}^{k-1} (L_{k-1})_{j\ell} \cdot \psi_\ell\right).\end{aligned} \tag{3.13}$$

Utilizing the bases

$$\begin{aligned}(z_1, \dots, z_{k-1}, z_k) &:= (\psi_1, \dots, \psi_{k-1}, \mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}), \\ (x_1, \dots, x_{k-1}, x_k) &:= (\mathrm{e}^{-\alpha_k \cdot} * \psi_1, \dots, \mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}, \mathrm{e}^{-\alpha_k \cdot}),\end{aligned}$$

this can be written as

$$\Lambda(\psi_k v) = \gamma_k \Lambda(\psi_{k-1} v) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \sum_{j=1}^{k} \Psi^*(x_j v) \sum_{\ell=1}^{k} (\widetilde{L}_{k-1})_{j\ell} z_\ell, \tag{3.14}$$

where

$$\widetilde{L}_{k-1} := \begin{bmatrix} L_{k-1} & 0 \\ 0 & 1 \end{bmatrix}.$$

At this point, we need a change of coordinates between the bases $(\psi_1, \cdots, \psi_k)$ and $(\mathrm{e}^{-\alpha_k \cdot} * \psi_1, \cdots, \mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}, \mathrm{e}^{-\alpha_k \cdot})$, as well as a transformation between the bases

$(\psi_1, \cdots, \psi_{k-1}, \mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1})$ and $(\psi_1, \cdots, \psi_k)$. Applying Laplace transform, this problem reduces to the determination of invertible matrices $T_k, M_k \in \mathbb{C}^{k\times k}$ with

$$T_k \begin{bmatrix} \widehat{\psi}_1(s) & \dots & \widehat{\psi}_{k-1}(s) & \widehat{\psi}_k(s) \end{bmatrix}^\top = \begin{bmatrix} \widehat{\psi}_1(s) & \dots & \widehat{\psi}_{k-1}(s) & \frac{\widehat{\psi}_{k-1}(s)}{s+\alpha_k} \end{bmatrix}^\top \tag{3.15}$$

$$\begin{bmatrix} \widehat{\psi}_1(s) & \dots & \widehat{\psi}_{k-1}(s) & \widehat{\psi}_k(s) \end{bmatrix} M_k = \begin{bmatrix} \frac{\widehat{\psi}_1(s)}{s+\alpha_k} & \dots & \frac{\widehat{\psi}_{k-1}(s)}{s+\alpha_k} & \frac{1}{s+\alpha_k} \end{bmatrix}. \tag{3.16}$$

We can immediately conclude from the recursion formula (2.3) that

$$T_k = \begin{bmatrix} I_{k-1} & 0 \\ \begin{bmatrix} 0 & \frac{1}{\alpha_k+\overline{\alpha_{k-1}}} \end{bmatrix} & \frac{-1}{\gamma_k(\alpha_k+\overline{\alpha_{k-1}})} \end{bmatrix}. \tag{3.17}$$

To obtain the matrix M_k, we use an argumentation similar to the proof of [7, Proposition 3.2.]. Namely, we show that

$$M_k = (M_{k,5} M_{k,4} M_{k,3} M_{k,2} M_{k,1})^{-1}$$

with

$$M_{k,1} := \begin{bmatrix} \frac{1}{\sqrt{2\mathrm{Re}(\alpha_1)}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{2\mathrm{Re}(\alpha_k)}} \end{bmatrix}, \quad M_{k,2} := \begin{bmatrix} \overline{\alpha_1}+\alpha_k & & \\ \alpha_1-\alpha_k & \overline{\alpha_2}+\alpha_k & \\ & \ddots & \\ & \alpha_{k-1}-\alpha_k & \overline{\alpha_k}+\alpha_k \end{bmatrix},$$

$$M_{k,3} := \begin{bmatrix} 1 & \dots & 1 \\ & \ddots & \vdots \\ & & 1 \end{bmatrix}, \quad M_{k,4} := \begin{bmatrix} 0 & I \\ 1 & 0 \end{bmatrix}, \quad M_{k,5} := \begin{bmatrix} -\sqrt{2\mathrm{Re}(\alpha_1)} & & \\ & \ddots & \\ & & -\sqrt{2\mathrm{Re}(\alpha_{k-1})} \\ & & 1 \end{bmatrix}.$$

We have, by (2.4),

$$\begin{aligned} E_k &:= \begin{bmatrix} \frac{\widehat{\psi}_1(s)}{s+\alpha_k} & \dots & \frac{\widehat{\psi}_{k-1}(s)}{s+\alpha_k} & \frac{1}{s+\alpha_k} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\sqrt{\mathrm{Re}(\alpha_1)}}{(s+\alpha_k)(s+\alpha_1)}, & \dots, & \frac{\sqrt{2\mathrm{Re}(\alpha_{k-1})}}{(s+\alpha_k)(s+\alpha_{k-1})} \prod_{\ell=1}^{k-2} \frac{s-\overline{\alpha_\ell}}{s+\alpha_\ell}, & \frac{1}{s+\alpha_k} \end{bmatrix}. \end{aligned}$$

Consecutive application of the matrices $(M_{k,j})_{j=1}^5$ to E_k results in

$$\begin{aligned} E_k M_{k,5} &= \begin{bmatrix} \frac{-2\mathrm{Re}(\alpha_1)}{(s+\alpha_k)(s+\alpha_1)}, & \dots, & \frac{-2\mathrm{Re}(\alpha_{k-1})}{(s+\alpha_k)(s+\alpha_{k-1})} \prod_{\ell=1}^{k-2} \frac{s-\overline{\alpha_\ell}}{s+\alpha_\ell}, & \frac{1}{s+\alpha_k} \end{bmatrix}. \\ E_k M_{k,5} M_{k,4} &= \begin{bmatrix} \frac{1}{s+\alpha_k}, & \frac{-2\mathrm{Re}(\alpha_1)}{(s+\alpha_k)(s+\alpha_1)}, & \dots, & \frac{-2\mathrm{Re}(\alpha_{k-1})}{(s+\alpha_k)(s+\alpha_{k-1})} \prod_{\ell=1}^{k-2} \frac{s-\overline{\alpha_\ell}}{s+\alpha_\ell} \end{bmatrix}. \\ E_k M_{k,5} M_{k,4} M_{k,3} &= \begin{bmatrix} \frac{1}{s+\alpha_k}, & \frac{s-\overline{\alpha_1}}{(s+\alpha_k)(s+\alpha_1)}, & \dots, & \frac{1}{(s+\alpha_k)} \prod_{\ell=1}^{k-1} \frac{s-\overline{\alpha_\ell}}{s+\alpha_\ell} \end{bmatrix}. \\ E_k M_{k,5} M_{k,4} M_{k,3} M_{k,2} &= \begin{bmatrix} \frac{2\mathrm{Re}(\alpha_1)}{s+\alpha_1}, & \frac{2\mathrm{Re}(\alpha_2)(s-\overline{\alpha_1})}{(s+\alpha_2)(s+\alpha_1)}, & \dots, & \frac{2\mathrm{Re}(\alpha_k)}{(s+\alpha_k)} \prod_{\ell=1}^{k-1} \frac{s-\overline{\alpha_\ell}}{s+\alpha_\ell} \end{bmatrix}. \\ E_k M_k^{-1} &= \begin{bmatrix} \frac{\sqrt{2\mathrm{Re}(\alpha_1)}}{s+\alpha_1}, & \frac{\sqrt{2\mathrm{Re}(\alpha_2)}(s-\overline{\alpha_1})}{(s+\alpha_2)(s+\alpha_1)}, & \dots, & \frac{\sqrt{2\mathrm{Re}(\alpha_k)}}{(s+\alpha_k)} \prod_{\ell=1}^{k-1} \frac{s-\overline{\alpha_\ell}}{s+\alpha_\ell} \end{bmatrix}, \\ &= \begin{bmatrix} \widehat{\psi}_1(s) & \dots & \widehat{\psi}_{k-1}(s) & \widehat{\psi}_k(s) \end{bmatrix}, \end{aligned}$$

which establishes (3.16).

Denote by (e_ℓ) the standard basis in $\mathbb{C}^p$ and in $\mathbb{C}^n$ (which space is intended will be clear from the context). Define the tensors $R_k, W_k \in \mathbb{C}^{n\times k\times p}$ by

$$\Lambda(\psi_k \otimes e_q) = \sum_{i=1}^{n}\sum_{j=1}^{k}(R_k)_{ijq}\psi_j \otimes e_i, \qquad q = 1,\dots,p, \tag{3.18}$$

$$\Psi^*(\psi_j \otimes e_q) = \sum_{i=1}^{n}(W_k)_{ijq}e_i, \quad j = 1,\dots,k, \quad q = 1,\dots,p. \tag{3.19}$$

In terms of these tensors, the induction hypothesis (3.10) can be written as

$$(R_k)_{ijq} = \sum_{\ell=1}^{k}(W_k)_{i\ell q}(L_k)_{\ell j}. \tag{3.20}$$

We now write all the terms in (3.14) in terms of these tensors.

We have by (3.18)

$$\Lambda(\psi_{k-1} \otimes e_q) = \sum_{i=1}^{m}\sum_{j=1}^{k-1}(R_{k-1})_{ijq}\psi_j \otimes e_i, \qquad q = 1,\dots,p,$$

which by (3.20) can be written as

$$\Lambda(\psi_{k-1} \otimes e_q) = \sum_{i=1}^{m}\sum_{j=1}^{k-1}\sum_{\ell=1}^{k-1}(W_{k-1})_{i\ell q}(L_{k-1})_{\ell j}\psi_j \otimes e_i, \qquad q = 1,\dots,p.$$

Defining

$$\widehat{L}_{k-1} = \begin{bmatrix} L_{k-1} & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{C}^{k\times k},$$

and using that $(W_{k-1})_{i\ell q} = (W_k)_{i\ell q}$ for $i = 1,\dots,n$, $q = 1,\dots,p$ and $\ell = 1,\dots,k-1$, we then have

$$\Lambda(\psi_{k-1} \otimes e_q) = \sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{\ell=1}^{k}(W_k)_{i\ell q}(\widehat{L}_{k-1})_{\ell j}\psi_j \otimes e_i, \qquad q = 1,\dots,p. \tag{3.21}$$

We now consider the term $\Psi^*(x_j v)$ in (3.14). By (3.16) we have for $v \in \mathbb{C}^p$

$$x_j \otimes v = \sum_{\ell=1}^{k}(M_k^T)_{j\ell}\psi_\ell \otimes v.$$

Substituting this in (3.19) gives

$$\begin{aligned}\Psi^*(x_j \otimes e_q) &= \sum_{\ell=1}^{k}(M_k^T)_{j\ell}\Psi^*(\psi_\ell \otimes e_q) = \sum_{i=1}^{n}\sum_{\ell=1}^{k}(M_k^T)_{j\ell}(W_k)_{i\ell q}e_i \\ &= \sum_{i=1}^{n}\sum_{\ell=1}^{k}(W_k)_{i\ell q}(M_k)_{\ell j}e_i. \end{aligned} \tag{3.22}$$

By (3.15) we have

$$z_\ell = \sum_{\beta=1}^{k} (T_k)_{\ell\beta}\psi_\beta.$$

It follows that

$$\sum_{\ell=1}^{k} (\widetilde{L}_{k-1})_{j\ell} z_\ell = \sum_{\ell=1}^{k}\sum_{\beta=1}^{k} (\widetilde{L}_{k-1})_{j\ell}(T_k)_{\ell\beta}\psi_\beta = \sum_{\beta=1}^{k} (\widetilde{L}_{k-1}T_k)_{j\beta}\psi_\beta. \tag{3.23}$$

From (3.22) and (3.23) we obtain that

$$\begin{aligned}\sum_{j=1}^{k} \Psi^*(x_j \otimes e_q) \sum_{\ell=1}^{k} (\widetilde{L}_{k-1})_{j\ell} z_\ell &= \sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\ell=1}^{k} (W_k)_{i\ell q}(M_k)_{\ell j} e_i \sum_{\beta=1}^{k} (\widetilde{L}_{k-1}T_k)_{j\beta}\psi_\beta \\ &= \sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\ell=1}^{k}\sum_{\beta=1}^{k} (W_k)_{i\ell q}(M_k)_{\ell j}(\widetilde{L}_{k-1}T_k)_{j\beta}\psi_\beta \otimes e_i \\ &= \sum_{i=1}^{n}\sum_{\ell=1}^{k}\sum_{\beta=1}^{k} (W_k)_{i\ell q}(M_k\widetilde{L}_{k-1}T_k)_{\ell\beta}\psi_\beta \otimes e_i. \end{aligned} \tag{3.24}$$

Substituting (3.18), (3.20), (3.21) and (3.24) in (3.14) gives

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\ell=1}^{k} (W_k)_{i\ell q}(L_k)_{\ell j}\psi_j \otimes e_i &= \gamma_k \sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{\ell=1}^{k} (W_k)_{i\ell q}(\widehat{L}_{k-1})_{\ell j}\psi_j \otimes e_i \\ &\quad - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \sum_{i=1}^{n}\sum_{\ell=1}^{k}\sum_{j=1}^{k} (W_k)_{i\ell q}(M_k\widetilde{L}_{k-1}T_k)_{\ell j}\psi_j \otimes e_i. \end{aligned} \tag{3.25}$$

We conclude that (3.25) is satisfied if L_k satisfies

$$L_k = \gamma_k(\widehat{L}_{k-1}) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) M_k \widetilde{L}_{k-1} T_k. \tag{3.26}$$

This recursively defines L_k and therefore, with this choice of L_k, the proof by induction of (3.10) (or equivalently (3.20)) is complete. Note that (3.26) is implemented in Algorithm 1. □

COROLLARY 3.3. *The matrix N_k, from* (3.9) *is given by Algorithm 1.*

Proof. We use the notation of the proof of Lemma 3.2. To complete the description of the algorithm, it only remains to re-formulate the tensors R_k and W_k and their relation (3.20) in matrix terms. Define the matrices

$$\widetilde{R}_k := \begin{bmatrix} R_{\cdot 1\cdot} \\ R_{\cdot 2\cdot} \\ \vdots \\ R_{\cdot k\cdot} \end{bmatrix} \in \mathbb{C}^{kn\times p}, \qquad \widetilde{W}_k := \begin{bmatrix} W_{\cdot 1\cdot} \\ W_{\cdot 2\cdot} \\ \vdots \\ W_{\cdot k\cdot} \end{bmatrix} \in \mathbb{C}^{kn\times p},$$

where $(R_{\cdot j\cdot})_{iq} = R_{ijq}$, $(W_{\cdot j\cdot})_{iq} = W_{ijq}$ for $i = 1,\dots,n$ and $q = 1,\dots,p$. Then (3.20) is equivalent to

$$(L_k^T \otimes I_n)\widetilde{W}_k = \widetilde{R}_k. \tag{3.27}$$

We have that N_k, from (3.9), satisfies

$$N_k^* = \begin{bmatrix} B^* R_{\cdot 1 \cdot} \\ B^* R_{\cdot 2 \cdot} \\ \vdots \\ B^* R_{\cdot k \cdot} \end{bmatrix}.$$

Algorithm 1 doesn't store $\widetilde{W}_k$ but instead the matrix $Q_k \in \mathbb{C}^{p \times km}$ defined through

$$Q_k^* = \begin{bmatrix} B^* W_{\cdot 1 \cdot} \\ B^* W_{\cdot 2 \cdot} \\ \vdots \\ B^* W_{\cdot k \cdot} \end{bmatrix}.$$

The relation (3.27) gives rise to $(L_k^T \otimes I_m) Q_k^* = N_k^*$, or equivalently (as it appears in Algorithm 1)

$$N_k = Q_k(\overline{L_k} \otimes I_m),$$

where $\overline{L_k}$ is the complex conjugate matrix of L_k.

From Corollary 2.11 a) we have

$$\Psi^*(\psi_j \otimes e_q) = \sqrt{2\mathrm{Re}(\alpha_j)} \sum_{i=1}^{n} (V_j)_{iq} e_i,$$

where V_j $(j = 1, \ldots, k)$ is as in (3.8). When compared with (3.19) this shows that

$$(W_k)_{ijq} = \sqrt{2\mathrm{Re}(\alpha_j)} (V_j)_{iq},$$

i.e. $W_{\cdot j \cdot} = \sqrt{2\mathrm{Re}(\alpha_j)} V_j$. Combining all of the above results gives Algorithm 1. □

4. The optimal control problem. In this section we consider the optimal control problem (1.2) & (1.3) and the optimal control problem (1.10) & (1.3).

LEMMA 4.1. *Let $A \in \mathbb{C}^{n \times n}$ be stable, $B \in \mathbb{C}^{n \times m}$, $C \in \mathbb{C}^{p \times n}$. Define Ψ, $\mathbb{F}$ and $\mathcal{R}_c$ by* (1.4), (1.5), (1.6). *The optimization problem* (1.2) *&* (1.3) *has a unique solution given by*

$$u^{\mathrm{opt}} = -\mathbb{F}^* \mathcal{R}_c^{-1} \Psi x_0.$$

The optimal cost is given by

$$\langle X x_0, x_0 \rangle, \quad X = \Psi^* \mathcal{R}_c^{-1} \Psi.$$

Proof. It is proven in [14, Proposition 7.2] that the optimal control is unique and is given by

$$u^{\mathrm{opt}} = -(I + \mathbb{F}^* \mathbb{F})^{-1} \mathbb{F}^* \Psi x_0,$$

and that the operator given the optimal cost is given by

$$X = \Psi^* \Psi - \Psi^* \mathbb{F} (I + \mathbb{F}^* \mathbb{F})^{-1} \mathbb{F}^* \Psi.$$

Using that $(I+\mathbb{F}^*\mathbb{F})^{-1}\mathbb{F}^* = \mathbb{F}^*\mathcal{R}_c^{-1}$, the given formulas follow. ∎

LEMMA 4.2. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$. Define Ψ_k, $\mathbb{F}_k$ and $\mathcal{R}_{c_k}$ by* (1.12), (1.13), (1.14), *where $P_k : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ is the orthogonal projection onto $\mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$ and $\mathscr{K}_k(\alpha)$ is the canonical rational Krylov subspace from Definition 2.3.*

The optimization problem (1.10) *&* (1.3) *has a unique solution given by*

$$u_k^{\mathrm{opt}} = -\mathbb{F}_k^*\mathcal{R}_{c,k}^{-1}\Psi_k x_0.$$

The optimal cost is given by

$$\langle X_k x_0, x_0\rangle, \quad X_k = \Psi_k^*\mathcal{R}_{c,k}^{-1}\Psi_k.$$

Proof. Noting that $P_k y = \Psi_k x_0 + \mathbb{F}_k u$, we use a "completing the square" argument similar to [14, Proposition 7.2]. That is, we make use of

$$\mathbb{F}_k^*\mathcal{R}_{c,k}^{-1} = \mathbb{F}_k^*(I+\mathbb{F}_k\mathbb{F}_k^*)^{-1} = (I+\mathbb{F}_k^*\mathbb{F}_k)^{-1}\mathbb{F}_k^*,$$

to see that

$$\begin{aligned}\|u\|_{L^2}^2 + \|P_k y\|_{L^2}^2 &= \|u\|_{L^2}^2 + \langle \Psi_k x_0 + \mathbb{F}_k u, \Psi_k x_0 + \mathbb{F}_k u\rangle_{L^2} \\ &= \langle \Psi_k^*\mathcal{R}_{c,k}^{-1}\Psi_k x_0, x_0\rangle + \langle (I+\mathbb{F}_k^*\mathbb{F}_k)(u+\mathbb{F}_k^*\mathcal{R}_{c,k}^{-1}\Psi_k x_0), (u+\mathbb{F}_k^*\mathcal{R}_{c,k}^{-1}\Psi_k x_0)\rangle_{L^2}.\end{aligned}$$

In particular, we have for $X_k = \Psi_k^*\mathcal{R}_{c,k}^{-1}\Psi_k$ that $\|u\|_{L^2}^2 + \|P_k y\|_{L^2}^2 \geq \langle X_k x_0, x_0\rangle$. In the case where the input reads $u = -\mathbb{F}_k^*\mathcal{R}_{c,k}^{-1}\Psi_k x_0$, the second summand vanishes. Thus, we have equality between $\|u\|_{L^2}^2 + \|P_k y\|_{L^2}^2$ and the quadratic form $\langle X_k x_0, x_0\rangle$ in this case. ∎

COROLLARY 4.3. *Under the assumptions and with the notation of Lemma 4.2, we have*

$$u_k^{\mathrm{opt}} \in \mathscr{K}_k(\alpha)\otimes\mathbb{C}^m.$$

Proof. By Lemma 4.2 we have $u_k^{\mathrm{opt}} = \mathbb{F}^* z$ for $z := -P_k\mathcal{R}_{c,k}^{-1}\Psi_k x_0 \in \mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$. From Proposition 2.14 d) we see that $\mathbb{F}^*$ maps $\mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$ into $\mathscr{K}_k(\alpha)\otimes\mathbb{C}^m$. Therefore $u_k^{\mathrm{opt}} \in \mathscr{K}_k(\alpha)\otimes\mathbb{C}^m$, as desired. ∎

THEOREM 4.4. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$. Define Ψ, $\mathbb{F}$ and $\mathcal{R}_c$ by* (1.4), (1.5), (1.6) *and $X = \Psi^*\mathcal{R}_c^{-1}\Psi$. Define Ψ_k, $\mathbb{F}_k$ and $\mathcal{R}_{c,k}$ by* (1.12), (1.13), (1.14), *where $P_k : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ is the orthogonal projection onto $\mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$ and $\mathscr{K}_k(\alpha)$ is the canonical rational Krylov subspace from Definition 2.3. Define $X_k = \Psi_k^*\mathcal{R}_{c,k}^{-1}\Psi_k$. Then*

$$X_k \leq X_{k+1}, \quad X_k \leq X.$$

Proof. For $x_0 \in \mathbb{C}^n$ and $u \in L^2(0,\infty;\mathbb{C}^m)$ with corresponding output y defined through (1.3) we have

$$\|P_k y\|_{L^2(0,\infty;\mathbb{C}^p)}^2 \leq \|P_{k+1} y\|_{L^2(0,\infty;\mathbb{C}^p)}^2,$$

since $\mathscr{K}_k(\alpha) \subset \mathscr{K}_{k+1}(\alpha)$. It follows that

$$\langle X_k x_0, x_0\rangle = \inf_{u\in L^2(0,\infty;\mathbb{C}^m)} \|u\|^2 + \|P_k y\|^2 \\ \leq \inf_{u\in L^2(0,\infty;\mathbb{C}^m)} \|u\|^2 + \|P_{k+1} y\|^2 = \langle X_{k+1} x_0, x_0\rangle.$$

Similarly, using that

$$\|P_k y\|^2_{L^2(0,\infty;\mathbb{C}^p)} \leq \|y\|^2_{L^2(0,\infty;\mathbb{C}^p)},$$

we obtain

$$\langle X_k x_0, x_0\rangle \leq \langle X x_0, x_0\rangle.$$

□

5. Convergence of Riccati-ADI. In this section we prove convergence of the Riccati-ADI method.

THEOREM 5.1. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$. Let $(\alpha_j)_{j=1}^\infty$ be such that $\mathrm{Re}(\alpha_j) > 0$ for all j. For $k \in \mathbb{N}$, let X_k be the operator obtain by Riccati-ADI. Then X_k converges as $k \to \infty$. If $(\alpha_j)_{j=1}^\infty$ satisfies the non-Blaschke condition* (1.17)*, then X_k converges to X, the nonnegative definite solution of the algebraic Riccati equation* (1.1)*.*

Proof. This is a special case of Theorem 5.2 where by finite-dimensionality the topology in which convergence occurs is irrelevant. □

We formulate the following theorem in the infinite-dimensional context. In the finite-dimensional case it simply reduces to Theorem 5.1. We refer to [12] for the terminology used in the statement of the following theorem (readers not familiar with this may simply consider the proof of the following theorem as a proof of Theorem 5.1).

THEOREM 5.2. *Consider a well-posed linear system on Hilbert spaces $\mathscr{U}$, $\mathscr{Y}$ and $\mathscr{X}$ that is output stable and input-output stable and whose semigroup is uniformly bounded. Denote its output map by Ψ and its input-output map by $\mathbb{F}$. Let $(\alpha_j)_{j=1}^\infty$ be such that $\mathrm{Re}(\alpha_j) > 0$ for all j and for $k \in \mathbb{N}$ let $P_k : L^2(0,\infty;\mathscr{Y}) \to L^2(0,\infty;\mathscr{Y})$ be the orthogonal projection onto $\mathscr{K}_k(\alpha) \otimes \mathscr{Y}$, where $\mathscr{K}_k(\alpha)$ is the canonical rational Krylov subspace from Definition 2.3. Define X_k by* (1.11)*. Then X_k converges in the strong operator topology as $k \to \infty$. Let X be given by* (1.7) *and assume that $(\alpha_j)_{j=1}^\infty$ satisfies the non-Blaschke condition* (1.17)*. Then X_k converges to X in the strong operator topology as $k \to \infty$. If moreover X is compact, then X_k converges to X in the uniform operator topology and if X is in the Schatten class $S_p(\mathscr{X})$ for $p \in [1,\infty]$, then X_k converges to X in the topology of $S_p(\mathscr{X})$.*

Proof. We first note that the results proven in the earlier parts of this article hold in the setting of this theorem (with essentially the same proofs).

Since, by Theorem 4.4, X_k is a non-decreasing sequence which is bounded from above, we obtain convergence in the strong operator topology.

Since $\mathscr{K}_k(\alpha) \subset \mathscr{K}_{k+1}(\alpha)$ we have $P_k \leq P_{k+1}$. Since P_k is an orthogonal projection, we have $P_k \leq I$. It follows that P_k converges in the strong operator topology to some orthogonal projection P. It was shown in [9, Lemma 4.4] that $P = I$ if and only if the non-Blaschke condition is satisfied (this result is shown there actually only for the case $\mathscr{Y} = \mathbb{C}$, but the behavior of tensor products under the strong operator topology [5, Theorem 1 part b] gives the general case).

From now on we assume the non-Blaschke condition, so that $P = I$. Then $\mathcal{R}_{c,k} = I + \mathbb{F}P_k\mathbb{F}^* \to I + \mathbb{F}\mathbb{F}^* = \mathcal{R}_c$ in the strong operator topology. It follows from [3, Theorem 7.6.1] that $\mathcal{R}_{c,k}^{-1} \to \mathcal{R}_c^{-1}$ in the strong operator topology. By sequential continuity of the strong operator topology we then have

$$X_k = \Psi^* P_k \mathcal{R}_{c,k}^{-1} P_k \Psi \to \Psi^* \mathcal{R}_c^{-1} \Psi = X,$$

in the strong operator topology.

If X is compact, then (since $\mathcal{R}_c$ is self-adjoint and invertible), Ψ is compact. As above we have that $P_k\mathcal{R}_{c,k}^{-1}P_k$ converges in the strong operator topology to $\mathcal{R}_c^{-1}$. From [9, Theorem A.2 part a] (which is a slight modification of [4, Thm. III.6.3]) we then obtain that $P_k\mathcal{R}_{c,k}^{-1}P_k\Psi_k$ converges to $\mathcal{R}_c^{-1}\Psi$ in the uniform operator topology. It follows that $X_k \to X$ in the uniform operator topology. The argument for Schatten class convergence is similar (see e.g. [9, Appendix A] for the needed relation between Schatten class membership of X and of Ψ). □

6. Comparison with the rational Krylov–Galerkin method. More generally than in the introduction (to allow for non-distinct α_j), we choose the Galerkin space

$$\mathscr{X}_k := \Psi^*(\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p), \tag{6.1}$$

which we characterized in terms of rational Krylov subspaces for A^* in Lemma 2.12 and Remark 2.13. We recall that in this section we make a dissipativity assumption on A (rather than just a stability assumption as done previously) so as to ensure stability of the Galerkin approximation A_q.

LEMMA 6.1. *Let $A \in \mathbb{C}^{n\times n}$ be dissipative (i.e. $A + A^* < 0$), $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$. Let $\mathscr{X}_k$ be the rational Krylov–Galerkin trial space from* (6.1). *For $x_0 \in \mathscr{X}_k$ and $u \in L^2(0,\infty;\mathbb{C}^m)$, let y_q be the output of* (1.18) *and let y be the output of* (1.3). *Let $P_k : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ be the orthogonal projection onto $\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$. Then $P_k y = P_k y_q$.*

Proof. We first note that $P_k y = P_k y_q$ holds if and only if for all $v \in \mathbb{C}^p$ and $j = 1, \dots, k$

$$\langle y, \psi_j v\rangle = \langle y_q, \psi_j v\rangle,$$

where $(\psi_j)_{j=1}^k$ is the Takenaka–Malmquist system defined in Definition 2.6. This in turn is equivalent to

$$\langle y, \varphi_j v\rangle = \langle y_q, \varphi_j v\rangle,$$

where $(\varphi_j)_{j=1}^k$ is the convolution system defined in Definition 2.4.

Let Ψ be the output map and $\mathbb{F}$ be the input-output map of (1.3). For Ψ_q the output map and $\mathbb{F}_q$ the input-output map of (1.18), we have $y = \Psi x_0 + \mathbb{F}u$ and $y_q = \Psi_q W_q^* x_0 + \mathbb{F}_q u$ so that the above equalities are implied by

$$\Psi^*(\varphi_j v) = W_q \Psi_q^*(\varphi_j v), \quad \mathbb{F}^*(\varphi_j v) = \mathbb{F}_q^*(\varphi_j v). \tag{6.2}$$

Both of these equalities are implied by

$$\Phi^t(\varphi_j v) = W_q \Phi_q^t(\varphi_j v), \tag{6.3}$$

for all $t \geq 0$, where Φ^t is defined in (2.5) and Φ^t_q is defined similarly. The first equation in (6.2) is obtained by putting $t = 0$ in (6.3) and the second equality in (6.2) is obtained from

$$\mathbb{F}^*_q(\varphi_j v) = t \mapsto B^*_q \Phi^t_q(\varphi_j v) = t \mapsto B^* W_q \Phi^t_q(\varphi_j v) = t \mapsto B^* \Phi^t(\varphi_j v) = \mathbb{F}^*(\varphi_j v).$$

We now show that (6.3) is implied by

$$W^*_q \Phi^t(\varphi_j v) = \Phi^t_q(\varphi_j v). \tag{6.4}$$

This can be seen by applying W_q to (6.4) to obtain

$$W_q W^*_q \Phi^t(\varphi_j v) = W_q \Phi^t_q(\varphi_j v),$$

using that $W_q W^*_q$ is the orthogonal projection onto $\mathscr{X}_k$ and that $\Phi^t(\varphi_j v) \in \mathscr{X}_k$ by Proposition 2.14 c), so that (6.2) is obtained as desired.

We now prove (6.4) by induction. By Proposition 2.14 a) we have $(\alpha_1 - A^*)^{-1} C^* v \in \mathscr{X}_k$. Using that $W_q W^*_q$ is the orthogonal projection onto $\mathscr{X}_k$ we then obtain

$$W^*_q(\alpha_1 - A^*) W_q W^*_q (\alpha_1 - A^*)^{-1} C^* v = W^*_q C^* v.$$

Using that $W^*_q W_q = I$ and the definitions of A_q and C_q this gives

$$(\alpha_1 - A_q) W^*_q (\alpha_1 - A^*)^{-1} C^* v = C^*_q v.$$

It follows that

$$W^*_q (\alpha_1 - A^*)^{-1} C^* v = (\alpha_1 - A_q)^{-1} C^*_q v.$$

We have by Corollary 2.10 a)

$$\Phi^t(\varphi_1 v) = (\alpha_1 - A^*)^{-1} C^* v \varphi_1(t), \quad \Phi^t_q(\varphi_1 v) = (\alpha_1 - A^*_q)^{-1} C^*_q v \varphi_1(t),$$

so that by the earlier computations we have $W^*_q \Phi^t(\varphi_1 v) = \Phi^t_q(\varphi_1 v)$, as desired.

Utilizing φ_j rather than φ_1 we similarly obtain

$$W^*_q (\alpha_j - A^*)^{-1} C^* v = (\alpha_j - A^*_q)^{-1} C^*_q v.$$

By Corollary 2.10 a), to show $W^*_q \Phi^t(\varphi_j v) = \Phi^t_q(\varphi_j v)$ it therefore only remains to show that

$$W^*_q (\alpha_j - A^*)^{-1} \Phi^t(\varphi_{j-1} v) = (\alpha_j - A^*_q)^{-1} \Phi^t_q(\varphi_{j-1} v). \tag{6.5}$$

Define

$$z := (\alpha_j - A^*)^{-1} \Phi^t(\varphi_{j-1} v).$$

Since $\Phi^t(\varphi_{j-1} v) \in \mathscr{X}_{j-1}$ by Proposition 2.14 c) and using Proposition 2.14 b), we have $z \in \mathscr{X}_j \subset \mathscr{X}_k$. Since $W_q W^*_q$ is the orthogonal projection onto $\mathscr{X}_k$ we then have $z = W_q W^*_q z$ and $W_q W^*_q \Phi^t(\varphi_{j-1} v) = \Phi^t(\varphi_{j-1} v)$. We therefore have

$$W_q W^*_q (\alpha_j - A^*) W_q W^*_q z = \Phi^t(\varphi_{j-1} v),$$

which by definition of A_q and the fact that $W^*_q W_q = I$ is equivalent to

$$W_q (\alpha_j - A^*_q) W^*_q z = \Phi^t(\varphi_{j-1} v).$$

Applying W_q^* and using that $W_q^* W_q = I$ we obtain

$$(\alpha_j - A_q^*) W_q^* z = W_q^* \Phi^t(\varphi_{j-1} v),$$

which gives

$$W_q^* z = (\alpha_j - A_q^*)^{-1} W_q^* \Phi^t(\varphi_{j-1} v).$$

By the induction hypothesis we have $W_q^* \Phi^t(\varphi_{j-1} v) = \Phi_q^t(\varphi_{j-1} v)$; using this and the definition of z we obtain from the above

$$W_q^* (\alpha_j - A^*)^{-1} \Phi^t(\varphi_{j-1} v) = (\alpha_j - A_q^*)^{-1} \Phi_q^t(\varphi_{j-1} v),$$

which is (6.5). □

REMARK 6.2. *Lemma 6.1 in particular implies that applying Riccati-ADI with parameters $(\alpha_j)_{j=1}^k$ to the original system and applying it to the rational Krylov–Galerkin approximation (with Galerkin space obtained using the same parameters $(\alpha_j)_{j=1}^k$) gives the same answer.*

THEOREM 6.3. *Let $A \in \mathbb{C}^{n\times n}$ be dissipative (i.e. $A + A^* < 0$), $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$. Let $k \in \mathbb{N}$, $(\alpha_j)_{j=1}^k$ be such that $\operatorname{Re}(\alpha_j) > 0$ for all j, let X_k be the operator obtained by Riccati-ADI and let $\widetilde{X}_k \in \mathbb{C}^{n\times n}$ be the operator obtained from solving the Riccati equation for the rational Krylov–Galerkin approximation with parameters $(\alpha_j)_{j=1}^k$. Then*

$$X_k \le \widetilde{X}_k.$$

Proof. For $x_0 \in \mathscr{X}_k$, $u \in L^2(0,\infty;\mathbb{C}^m)$, y the output of the original system (1.3), y_q the output of the rational Krylov–Galerkin system (1.18), and $P_k : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ the orthogonal projection onto $\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$, we have from Lemma 6.1

$$\begin{aligned}\|u\|^2_{L^2(0,\infty;\mathbb{C}^m)} + \|P_k y\|^2_{L^2(0,\infty;\mathbb{C}^p)} &= \|u\|^2_{L^2(0,\infty;\mathbb{C}^m)} + \|P_k y_q\|^2_{L^2(0,\infty;\mathbb{C}^p)} \\ &\le \|u\|^2_{L^2(0,\infty;\mathbb{C}^m)} + \|y_q\|^2_{L^2(0,\infty;\mathbb{C}^p)}.\end{aligned}$$

By infimizing over $u \in L^2(0,\infty;\mathbb{C}^m)$ we conclude that for $x_0 \in \mathscr{X}_k$

$$\langle X_k x_0, x_0\rangle \le \langle \widetilde{X}_k x_0, x_0\rangle. \tag{6.6}$$

We note that $\mathcal{R}_{c,k}$ maps $\mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$ into itself and that hence $\mathcal{R}_{c,k}^{-1}$ maps $\mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$ into itself as well. Since $\Psi_k\mathbb{C}^n \subset \mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$, it follows that $\mathcal{R}_{c,k}^{-1}\Psi_k\mathbb{C}^n \subset \mathscr{K}_k(\alpha)\otimes\mathbb{C}^p$ and therefore $\Psi_k^*\mathcal{R}_{c,k}^{-1}\Psi_k\mathbb{C}^n \subset \mathscr{X}_k$. So X_k maps $\mathbb{C}^n$ into $\mathscr{X}_k$. We conclude that

$$\langle X_k x_0, x_0\rangle = 0 \text{ for } x_0 \in \mathscr{X}_k^{\perp}.$$

By definition of $\widetilde{X}_k$ we also have $\langle \widetilde{X}_k x_0, x_0\rangle = 0$ for $x_0 \in \mathscr{X}_k^{\perp}$. Combined with (6.6) this gives the desired conclusion. □

THEOREM 6.4. *Let $A \in \mathbb{C}^{n\times n}$ be dissipative (i.e. $A + A^* < 0$), $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$. Let $k \in \mathbb{N}$, $(\alpha_j)_{j=1}^k$ be such that $\operatorname{Re}(\alpha_j) > 0$ for all j, let X_k be the operator obtained by Riccati-ADI and let $\widetilde{X}_k \in \mathbb{C}^{n\times n}$ be the operator obtained from solving the Riccati equation for the rational Krylov–Galerkin approximation with parameters*

$(\alpha_j)_{j=1}^k$. Let $X_q^G \in \mathbb{C}^{q\times q}$ be the nonnegative definite solution of the Riccati equation for the rational Krylov–Galerkin approximation. We have

$$X_k = \widetilde{X}_k,$$

if $\alpha = -\sigma(A_q^{\mathrm{opt}})$, where $A_q^{\mathrm{opt}} := A_q - B_q B_q^* X_q^G$ and eigenvalues are repeated according to algebraic multiplicity. The converse holds provided that X_q^G is positive definite.

Proof. By standard optimal control theory, the optimal output of (1.2) & (1.18) for $x_0 \in \mathscr{X}_k$ is given by:

$$y_q^{\mathrm{opt}}(t) = C_q e^{A_q^{\mathrm{opt}} t} x_0. \tag{6.7}$$

So if $\alpha = -\sigma(A_q^{\mathrm{opt}})$, we have $y_q^{\mathrm{opt}} \in \mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$. It then follows that $P_k y_q^{\mathrm{opt}} = y_q^{\mathrm{opt}}$ and from the proof of Theorem 6.3 we then obtain that $\langle X_k x_0, x_0\rangle = \langle \widetilde{X}_k x_0, x_0\rangle$. In that same proof it was already remarked that equality holds for $x_0 \in \mathscr{X}_k^{\perp}$ (both sides being equal to zero), so that we have $X_k = \widetilde{X}_k$.

For the converse, we see from the proof of Theorem 6.3 that $X_k = \widetilde{X}_k$ implies that $y_q^{\mathrm{opt}} \in \mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$ for all initial conditions $x_0 \in \mathscr{X}_k$. From that proof we also see, by infimizing over $u \in \mathscr{K}_k(\alpha) \otimes \mathbb{C}^m$ instead and using Corollary 4.3, that $u_q^{\mathrm{opt}} \in \mathscr{K}_k(\alpha) \otimes \mathbb{C}^m$. Let x_0 be an eigenvector of A_q^{opt} with corresponding eigenvalue λ. Then, using (6.7) and the similar formula $u_q^{\mathrm{opt}}(t) = -B_q^* X_q^G e^{A_q^{\mathrm{opt}} t} x_0$ for the optimal control, we have

$$y_q^{\mathrm{opt}}(t) = C_q x_0 \mathrm{e}^{\lambda t}, \quad u_q^{\mathrm{opt}}(t) = -B_q^* X_q^G x_0 \mathrm{e}^{\lambda t}.$$

Since X_q^G is positive definite it follows that at least one of $C_q x_0 \neq 0$ and $B_q^* X_q^G x_0 \neq 0$ holds true (since otherwise the optimal cost for initial condition x_0 would be zero). We conclude that $t \mapsto \mathrm{e}^{\lambda t}$ is in $\mathscr{K}_k(\alpha)$, which implies that $-\lambda \in \alpha$. If A_q^{opt} is not diagonalizable, then a similar argument using generalized eigenvectors shows that an eigenvalue must occur in the sequence α according to its algebraic multiplicity. ∎

7. Algorithm.

Algorithm 1 ADI iteration for Riccati equations.

Input: $A \in \mathbb{C}^{n\times n}$ a stable matrix, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$ and shift parameters $\alpha_1, \ldots, \alpha_k \in \mathbb{C}$ with $\mathrm{Re}(\alpha_i) > 0$.

Output: $S_k \in \mathbb{C}^{kp\times n}$, $F_k \in \mathbb{C}^{kp\times km}$ such that $S_k^*(I_{kp} + F_k F_k^*)^{-1} S_k \approx X$, where X is the unique nonnegative definite solution of the algebraic Riccati equation

$$A^*X + XA + C^*C - XBB^*X = 0.$$

$V_1 = (\alpha_1 - A^*)^{-1}C^*$

$S_1 = \sqrt{2\mathrm{Re}(\alpha_1)} \cdot V_1^*$

$Q_1 = \sqrt{2\mathrm{Re}(\alpha_1)} \cdot V_1^* B$

$L_1 = \frac{1}{\sqrt{2\mathrm{Re}(\alpha_1)}}$

$F_1 = Q_1 L_1$

for $i = 2, 3, \ldots, k$ **do**

$V_i = V_{i-1} - (\alpha_i + \overline{\alpha_{i-1}}) \cdot (\alpha_i - A^*)^{-1} V_{i-1}$

$S_i = [\, S_{i-1}^*\,,\ \sqrt{2\mathrm{Re}(\alpha_i)} \cdot V_i \,]^*$

$Q_i = [\, Q_{i-1}\,,\ \sqrt{2\mathrm{Re}(\alpha_i)} \cdot V_i^* B \,]$

$\gamma_i := \sqrt{\frac{\mathrm{Re}(\alpha_j)}{\mathrm{Re}(\alpha_{j-1})}}$

$$M_{i,1} := \begin{bmatrix} \frac{1}{\sqrt{2\mathrm{Re}(\alpha_1)}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{2\mathrm{Re}(\alpha_i)}} \end{bmatrix}, \quad M_{i,2} = \begin{bmatrix} \frac{\overline{\alpha_1}+\alpha_i}{\alpha_1-\alpha_i} & & \\ & \frac{\overline{\alpha_2}+\alpha_i}{} & \\ & \ddots & \\ & \alpha_{i-1}-\alpha_i & \frac{\overline{\alpha_i}+\alpha_i}{} \end{bmatrix},$$

$$M_{i,3} = \begin{bmatrix} 1 & \ldots & 1 \\ & \ddots & \vdots \\ & & 1 \end{bmatrix}, \quad M_{i,4} = \begin{bmatrix} 0 & I \\ 1 & 0 \end{bmatrix}, \quad M_{i,5} = \begin{bmatrix} -\sqrt{2\mathrm{Re}(\alpha_1)} & & \\ & \ddots & \\ & & -\sqrt{2\mathrm{Re}(\alpha_{i-1})} \\ & & 1 \end{bmatrix}$$

$M_i = M_{i,1}^{-1} M_{i,2}^{-1} M_{i,3}^{-1} M_{i,4}^{-1} M_{i,5}^{-1}$

$$L_i = \begin{bmatrix} \gamma_i L_{i-1} & 0 \\ 0 & 0 \end{bmatrix} - M_i \begin{bmatrix} L_{i-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_i(\alpha_i + \overline{\alpha_{i-1}})I & 0 \\ [0, \gamma_i] & -1 \end{bmatrix}$$

$$F_i = \begin{bmatrix} [F_{i-1}, 0] \\ Q_i \left(\overline{L_i} \otimes I_m\right) \end{bmatrix}$$

end for

REMARK 7.1. *For the optimal control problem* (1.2) *subject to*

$$E\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad y(t) = Cx(t), \tag{7.1}$$

with invertible $E \in \mathbb{C}^{n\times n}$, the above procedure can be done without explicit inversion of E: We have to make the replacements

1: $V_1 = (\alpha_1 E - A^*)^{-1} C^*$

7: $V_i = V_{i-1} - (\alpha_i + \overline{\alpha_{i-1}}) \cdot (\alpha_i E - A^*)^{-1} E V_{i-1}$

in the above algorithm (cf. [7, Remark 3.3]).

REMARK 7.2. *The choice of shift parameters is essential for the speed of convergence. In [7, Section 3.2] it is stated that a choice based on the stable eigenvalues of the Hamiltonian*

$$\mathcal{H} = \begin{bmatrix} A & -BB^* \\ -C^*C & -A^* \end{bmatrix}$$

is effective. Our approach gives an alternative interpretation of this fact as follows. Since the stable eigenvalues of the Hamiltonian $\mathcal{H}$ are the eigenvalues of $A - BB^*X$ [15, Chap. 13], we have, in case where the first n shifts are (counted by multiplicity) the stable eigenvalues of $\mathcal{H}$, that the output corresponding to the optimal control for the optimal control problem (1.2) *&* (1.3) *fulfills*

$$y^{\mathrm{opt}} \in \mathscr{K}_n(\alpha) \otimes \mathbb{C}^p.$$

As a consequence, this particular choice gives rise to the fact that our projected optimal control problem (1.10) *coincides with the original optimal control problem, which gives $X = X_n$ (for this particular choice of shift parameters).*
In [7, Section 5] the following reasonable approach to shift parameter selection is proposed. Choose $N \in \mathbb{N}$. Then perform N iterations with N shift parameters chosen by using the method of Wachspress *[13] on the basis of the eigenvalues of A. Thereafter, determine N Wachspress parameters on the basis of the eigenvalues of $A - BB^*X_N$, and perform the next N iterations with these shift parameters. After that, compute N Wachspress parameters on the basis of the eigenvalues of $A - BB^*X_{2N}$, and perform the next N iterations with these shift parameters; repeat this approach any N steps. By convergence of (X_k) to X (established in Section 5), these parameters converge to the eigenvalues of $A - BB^*X$.*
The efficient numerical computation of dominant stable eigenvalues of a Hamiltonian matrix seems not to have been explored so far. The ADI method for Riccati equations would be an application for this research area.

8. Numerical Results. We present two numerical examples to show the applicability of our algorithm and to demonstrate the expected performance of the Riccati-ADI iteration in terms of monotonicity and convergence behavior. All the calculations are done using MATLAB 7.10.0 (R2010a).

8.1. Two Dimensional Convection-Diffusion Equation. Let $\Omega := [0,1] \times [0,1]$ be the unit square with boundary $\partial\Omega := \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$, where $\Gamma_1 := \{0\} \times [0,1]$, $\Gamma_2 := [0,1] \times \{0\}$, $\Gamma_3 := [0,1] \times \{1\}$, and $\Gamma_4 := \{1\} \times [0,1]$.

We consider the two-dimensional convection-diffusion equation

$$\frac{\partial x}{\partial t}(\xi,t) = \Delta x(\xi,t) + b^\top \nabla x(\xi,t), \quad (\xi,t) \in \Omega \times \mathbb{R}_{\geq 0}, \tag{8.1}$$

with Robin boundary conditions

$$\begin{aligned} u(t) &= \nu(\xi)^\top \nabla x(\xi,t) + \alpha x(\xi,t), && (\xi,t) \in (\Gamma_1 \cup \Gamma_2) \times \mathbb{R}_{\geq 0}, \\ 0 &= \nu(\xi)^\top \nabla x(\xi,t) + \alpha x(\xi,t), && (\xi,t) \in (\Gamma_3 \cup \Gamma_4) \times \mathbb{R}_{\geq 0}. \end{aligned}$$

and two-dimensional output

$$y(t) = \begin{bmatrix} \int_{\Gamma_1} x(\xi,t) d\sigma_\xi \\ \int_{\Gamma_3} x(\xi,t) d\sigma_\xi \end{bmatrix},$$

where σ_ξ denotes the surface measure and $\nu(\xi)$ denotes the outward normal.

We consider $b = \left[\begin{smallmatrix} 10 \\ 10 \end{smallmatrix}\right]$ and set $\alpha = 1$. To discretize the PDE (8.1), we apply a finite element discretization with uniform triangular elements of fixed size $h = \frac{1}{N-1}$, where $N \in \mathbb{N}$ is the number of points in each coordinate direction. In addition, we define the subspace $V_h \subset H^1(\Omega)$ using piecewise-linear basis functions. As a result, we obtain a finite dimensional dynamical system (7.1) with state-space dimension $n = N^2$. The

matrix $E \in \mathbb{R}^{n\times n}$ is a symmetric positive definite mass matrix, $A \in \mathbb{R}^{n\times n}$ is a non-symmetric stiffness matrix, $B \in \mathbb{R}^{n\times 1}$ is the input matrix, and $C \in \mathbb{R}^{2\times n}$ the output matrix.

In order to find an approximate solution of the corresponding algebraic Riccati equation, we apply Algorithm 1 with the modifications in Remark 7.1.

The choice of the shift parameters has a major effect on the convergence speed of the Riccati-ADI algorithm. In this example, we show that if the shift parameters do not satisfy the non-Blaschke condition (1.17), then the matrix X_k obtained by Algorithm 1 may converge to a nonnegative matrix which is not the solution of algebraic Riccati equation corresponding to the system (7.1) (cf. Theorem 5.1). To this end, we choose the following two different sets of shift parameters to use in our example.

1. The first set of shift parameters is chosen using Penzl's heuristic procedure [10] on the matrix pencil $\lambda E - A$. By this choice we generate a set of 10 shift parameters, which we re-use every 10 iterations. We sort these 10 shift parameters in an increasing order with respect to the values of their real parts in order to obtain a smooth convergence in Algorithm 1. This cyclic choice of shift parameters satisfies the non-Blaschke condition (1.17).
2. As a second set of shift parameters, we choose the infinite sequence $p_i = i^3$, $i = 1, 2, \ldots$, for which the non-Blaschke condition is not satisfied.

We have performed the simulation using several values of the state space di-

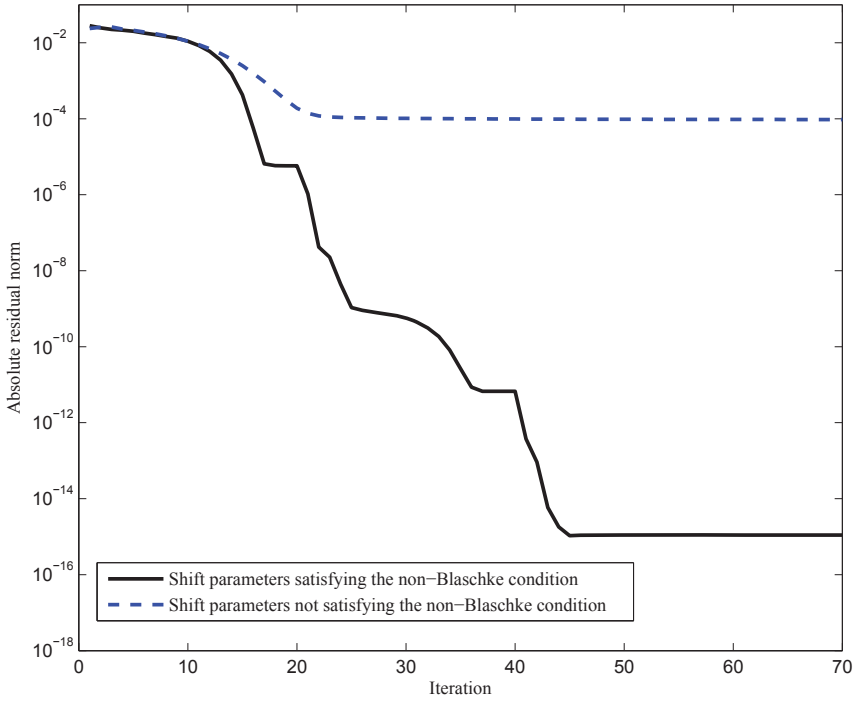

FIGURE 1. *Comparison of two sets of shift parameters for Riccati-ADI: convection-diffusion equation with the state space dimension* $n = 3600$

mension and with the two sets of shift parameters which we introduced above. At each iteration k, we observe the residual norm using the approach proposed in [7, Sec. 3.3]. That is, we exploit the low-rank form of the approximate solution $X_k = S_k(I+F_kF_k^*)^{-1}S_k^*$ to calculate the residual norm. Figure 1 shows the absolute residual norm with respect to the iteration for problem dimension $n = 3600$.

Considering Figure 1, we observe that by choosing the second set of shift parameters, $p_i = i^3$, our sequence converges to a matrix which is not the solution of the corresponding algebraic Riccati equation. In addition, with a tolerance of 10^{-14} on the residual norm, the first choice of shift parameters provides convergence to the desired solution in less than 45 iterations for state space dimensions satisfying $n \leq 3600$. We use the first set of shift parameters to continue with further analyses in our example.

In order to illustrate Theorem 6.3, we have implemented the rational Krylov subspace method (RKSM) based on [16, 17] and compared the iteration history of this method with that of Riccati-ADI by using the same set of shift parameters for both algorithms. Specifically, we use the first set of shift parameters which we have already computed in the previous analysis. At each iteration k, we compute the traces of X_k (the solution obtained using Riccati-ADI) and $\widetilde{X_k}$ (the solution obtained using RKSM) denoted by $tr\,(X_k)$ and $tr\left(\widetilde{X_k}\right)$, respectively. The trace of X_k (and

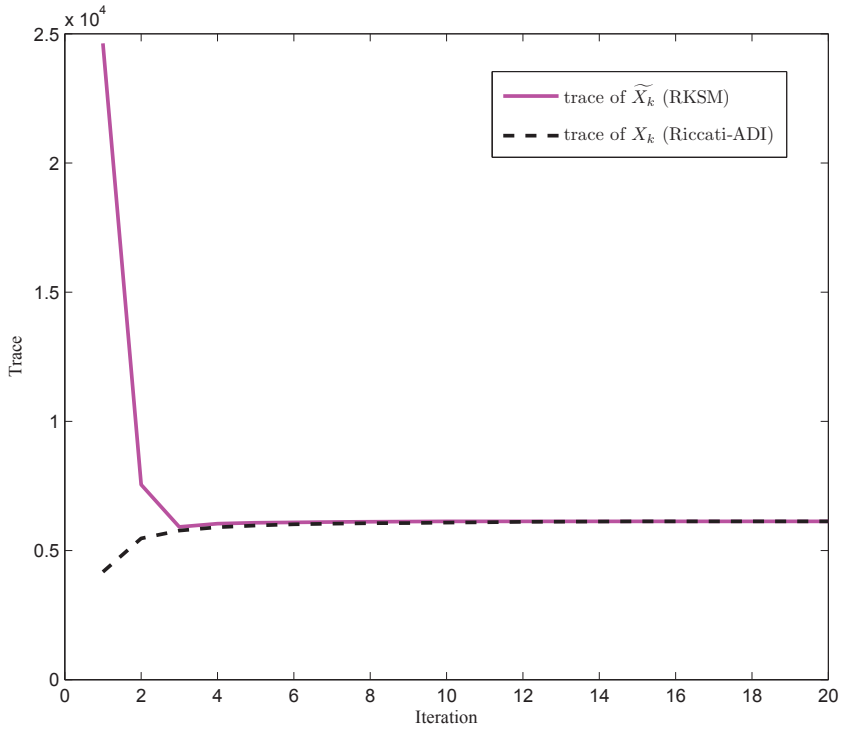

FIGURE 2. *Trace of the solution using RKSM and Riccati-ADI: convection-diffusion equation with the state space dimension* $n = 3600$

similarly that of $\widetilde{X_k}$) can be computed efficiently as follows. We compute the Cholesky factorization of $I_{kp} + F_k F_k^* = U_k^* U_k$ and therefore we obtain

$$tr\,(X_k) = tr\left(S_k (U_k^* U_k)^{-1} S_k^*\right) = tr\left(S_k U_k^{-1} U_k^{-*} S_k^*\right) = \|S_k U_k^{-1}\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Figure 2 illustrates the traces of X_k and $\widetilde{X_k}$ at each iteration. We observe that for all $k \in \mathbb{N}$ there holds $tr\,(X_k) \leq tr\left(\widetilde{X_k}\right)$ (consistent with Theorem 6.3). Note that in our example, by setting a tolerance of 10^{-9}, RKSM terminates in less than 23 iterations for state space dimensions satisfying $n \leq 3600$. In addition, from Figure 3, we observe that $tr\,(X_k)$ is a non-decreasing function of the iteration k. This illustrates the monotonicity of Riccati-ADI which we have proven in Theorem 4.4.

8.2. Euler-Bernoulli beam. As a second example, we consider the Euler-Bernoulli beam problem taken from [18]. Discretization of this PDE gives small enough matrices for a direct method for the solution of Riccati equations to be utilized; this allows for comparison of Riccati ADI with this "true solution". For ease of reference, we briefly present the problem in the following. The beam is clamped at one end ($r = 0$) and is free to vibrate at the other end ($r = R$); the control acts at the free end. The deflection of the beam from its rigid body motion at time t and position r is denoted by $w(r, t)$. As a result, the corresponding partial differential

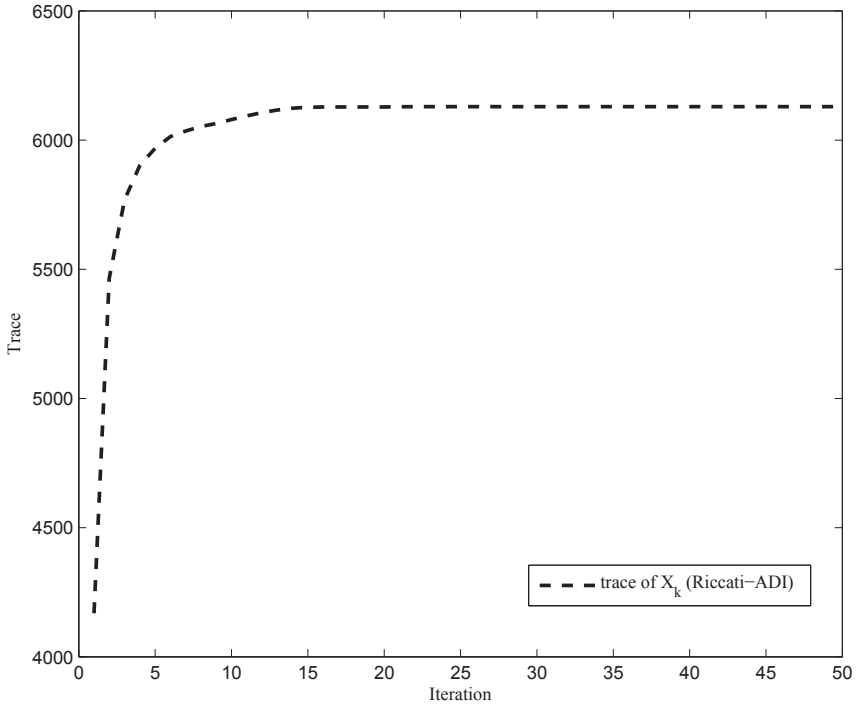

FIGURE 3. *Monotonicity of Riccati-ADI: convection-diffusion equation with the state space dimension* $n = 3600$

equation model with Kelvin-Voigt and viscous damping is

$$\rho w_{tt}(r,t) + C_v w_t(r,t) + \frac{\partial^2}{\partial r^2}\left[C_d I_b w_{rrt}(r,t) + EI_b w_{rr}(r,t)\right] = 0, \tag{8.2}$$

with boundary conditions and controls

$$\begin{aligned} w(0,t) &= 0, \\ w_r(0,t) &= 0, \\ \left[C_d I_b w_{rrt}(r,t) + EI_b w_{rr}(r,t)\right]_{r=R} &= u_1(t), \\ \left[C_d I_b w_{rrrt}(r,t) + EI_b w_{rrr}(r,t)\right]_{r=R} &= u_2(t). \end{aligned} \tag{8.3}$$

where we have used the notations $w_t := \frac{\partial}{\partial t} w$ and $w_r := \frac{\partial}{\partial r} w$. In addition, we consider a two dimensional boundary observation at the free end ($r = R$) described by

$$y(t) = \begin{bmatrix} w_t(R,t) \\ w_{rt}(R,t) \end{bmatrix}.$$

The values of the physical parameters are chosen as in Table 1. These values coincide with [18, Table 11], with modifications in the values of C_v and C_d (instead of the values $C_v = 2$ and $C_d = 5 \times 10^8$ from [18, Table 11], we use $C_v = 20$ and $C_d = 10^8$). We partition the interval $[0, R]$ into N uniform subintervals and consider a finite

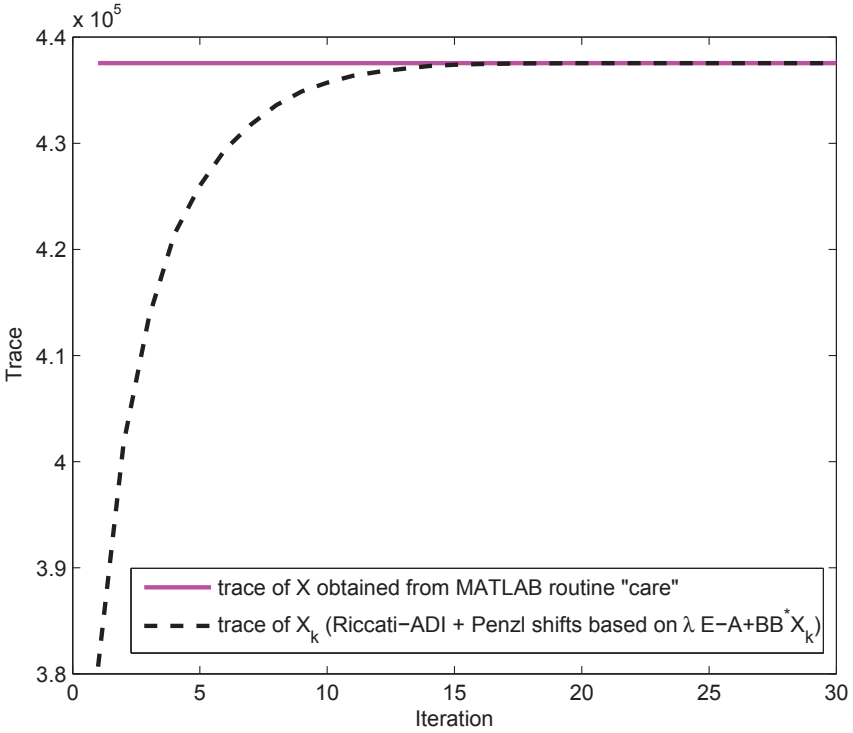

FIGURE 4. *Trace of the solution using Riccati-ADI and the "care" routine in MATLAB: Euler-Bernoulli beam with the state space dimension* $n = 96$

Parameter	Value
E	$2.68 \times 10^{10}\ N/m^2$
I_b	$1.64 \times 10^{-9}\ m^4$
ρ	1.02087
C_v	$20\ Ns/m^2$
C_d	$10^8\ Ns/m^2$
L	$1\ m$

TABLE 1
Physical parameters of the Euler-Bernoulli beam

element discretization of (8.2) using standard cubic B-splines. As a result, we obtain the following finite dimensional dynamical system in $H^{2N} \times H^{2N}$:

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t), \end{aligned} \tag{8.4}$$

with the state-space dimension $n = 4N$. The matrix $E \in \mathbb{R}^{n\times n}$ is a symmetric positive definite mass matrix, $A \in \mathbb{R}^{n\times n}$ is a non-symmetric stiffness matrix, $B \in \mathbb{R}^{n\times 2}$ is the input matrix, and $C \in \mathbb{R}^{2\times n}$ the output matrix.

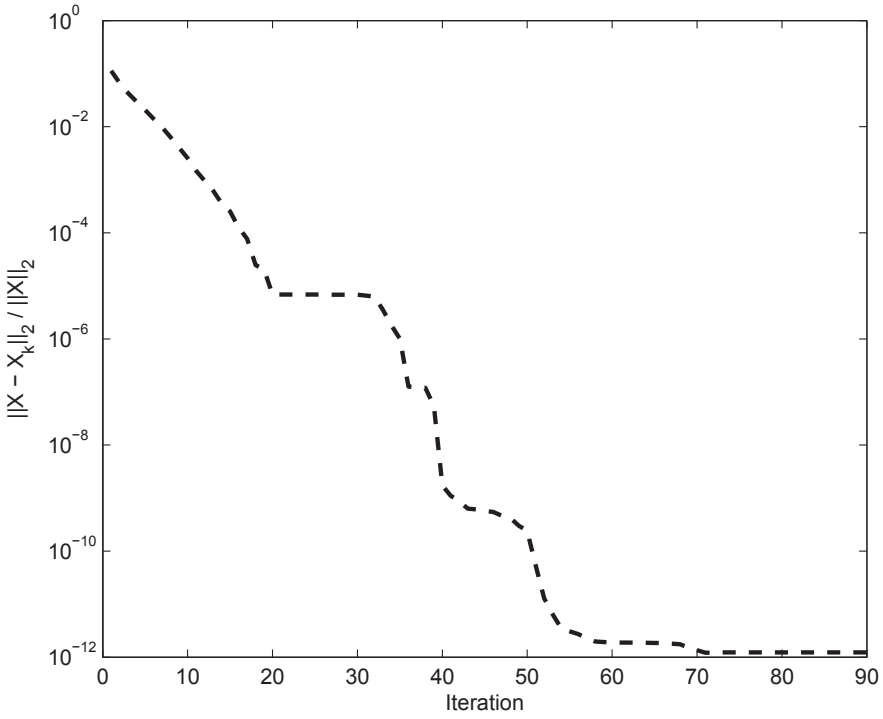

FIGURE 5. *The relative error of the solution obtained by Riccati-ADI with respect to the solution obtained by the "care" routine in MATLAB: Euler-Bernoulli beam with the state space dimension* $n = 96$

We consider $N = 24$ and solve the algebraic Riccati equation corresponding to the system (8.4) once using the "care" routine of MATLAB and once using Algorithm 1 with the modifications in Remark 7.1. The shift parameters are chosen similarly to the first set of parameters in the previous example. That is, we apply Penzl's heuristic method to the matrix pencil $\lambda E - A$. By this procedure we generate a set of 20 shift parameters, which we re-use every 20 iterations. For this example, we sort these 20 shift parameters in a decreasing order with respect to the values of their real parts in order to obtain a smooth convergence in Algorithm 1. Note that this cyclic set of shift parameters satisfies the non-Blaschke condition (1.17).

We denote by X the solution obtained from the "care" routine and we use it as a reference for the comparisons with the solution obtained by Algorithm 1 (denoted by X_k). Also note that the modifications in the values of C_v and C_d that we mentioned earlier, ensure that the associated Hamiltonian pencil has eigenvalues far from the imaginary axis and therefore we obtain a more accurate solution using the "care" routine in MATLAB.

In order to illustrate Theorem 4.4, we consider the traces of X and X_k denoted respectively by $tr\,(X)$ and $tr\,(X_k)$. Note that $tr\,(X_k)$ can be computed efficiently by considering the Cholesky decomposition of $I_{kp} + F_k F_k^*$ as we have already shown in the previous example. Figure 4 shows that $tr\,(X_k) \leq tr\,(X)$ for all $k \in \mathbb{N}$. In addition, we observe the relative error $\dfrac{\|X_k - X\|_2}{\|X\|_2}$ at every iteration to show the convergence behavior of the Riccati-ADI algorithm. Figure 5 shows the relative error of the solutions obtained by Algorithm 1 with respect to the solution obtained by the "care" routine in MATLAB.

REFERENCES

[1] Peter Benner and Jens Saak. Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey. *GAMM-Mitt.*, 36(1):32–52, 2013.

[2] Klaus-Jochen Engel and Rainer Nagel. *One-parameter semigroups for linear evolution equations*. Springer-Verlag, New York, 2000.

[3] Arthur E. Frazho and Wisuwat Bhosri. *An operator perspective on signals and systems*, volume 204 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, 2010.

[4] Israel C. Gohberg and Mark G. Kreĭn. *Introduction to the theory of linear nonselfadjoint operators*. Translated from the Russian by A. Feinstein. Translations of Mathematical Monographs, Vol. 18. American Mathematical Society, Providence, R.I., 1969.

[5] Carlos S. Kubrusly and Paulo C. M. Vieira. Convergence and decomposition for tensor products of Hilbert space operators. *Oper. Matrices*, 2(3):407–416, 2008.

[6] Jing-Rebecca Li and Jacob White. Low rank solution of Lyapunov equations. *SIAM J. Matrix Anal. Appl.*, 24(1):260–280, 2002.

[7] Yiding Lin and Valeria Simoncini. A new subspace iteration method for the algebraic Riccati equation. *Numer. Linear Algebra Appl.*, 2014. Article first published online, available at `http://onlinelibrary.wiley.com/doi/10.1002/nla.1936/abstract`.

[8] An Lu and Eugene L. Wachspress, Solution of Lyapunov equations by alternating direction implicit iteration, Comput. Math. Appl., 21(9):43–58, 1991.

[9] Mark R. Opmeer, Timo Reis, and Winnifried Wollner. Finite-rank ADI iteration for operator Lyapunov equations. *SIAM J. Control Optim.*, 51(5):4084–4117, 2013.

[10] T. Penzl. *A cyclic low-rank Smith method for large sparse Lyapunov equations. Siam. J. Sci. Comput.*, 21(4):1401–1418, 1999/00.

[11] Valeria Simoncini. Computational methods for linear matrix equations. preprint, 2013.

[12] Olof J. Staffans. *Well-posed linear systems*. Cambridge University Press, Cambridge, 2005.

[13] Eugene L. Wachspress. *Iterative solution of the Lyapunov matrix equation. Appl. Math. Lett.*, 1:87–90, 1988.

[14] Martin Weiss and George Weiss. Optimal control of stable weakly regular linear systems. *Math.*

Control Signals Systems, 10(4):287–330, 1997.

[15] Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, NJ, 1996.

[16] Valeria Simoncini. *A new iterative method for solving large-scale Lyapunov matrix equations.*. SIAM J. Sci. Comput., 29(3):1268–1288, 2007.

[17] Valeria Simoncini, Daniel B. Szyld, and Marlliny Monsalve *On two numerical methods for the solution of large-scale algebraic Riccati equations*. IMA Journal of Numerical Analysis, 1–17, 2013.

[18] Kirsten A. Morris and Carmeliza Navasca. *Approximation of low rank solutions for linear quadratic control of partial differential equations. Comput Optim Appl*, 46:93–111, 2010.