SIAM J. Matrix Anal. Appl.
Vol. 37, No. 2, pp. 624–648

ANALYSIS OF AN ITERATION METHOD FOR THE ALGEBRAIC RICCATI EQUATION*

ARASH MASSOUDI†, MARK R. OPMEER‡, and TIMO REIS†

Abstract. We consider a recently published method for solving algebraic Riccati equations. We present a new perspective on this method in terms of the underlying linear-quadratic optimal control problem: we prove that the matrix obtained by this method expresses the optimal cost for a projected optimal control problem. The projection is determined by the so-called shift parameters of the method. Our representation in terms of the optimal control problem gives rise to a simple and very general convergence analysis.

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

DOI. 10.1137/140985792

1. Introduction.

We consider an algorithm developed in [6] for obtaining an approximation of the unique positive semidefinite 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 of its eigenvalues are in the open left half-plane), $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$, and the superscript $*$ denotes the complex conjugate transpose.

The algorithm is iterative in nature and at step k produces an approximate solution of the form $X_k = S_k T_k^{-1} S_k^*$ for $S_k \in \mathbb{C}^{n\times kp}$ and positive definite $T_k \in \mathbb{C}^{kp\times kp}$. The main computational cost in the algorithm consists of, at each iteration step, solving a linear system of the form $(\alpha_i I - A)x = v$, where $v \in \mathbb{C}^{n\times p}$ and the "shift parameter" $\alpha_i \in \mathbb{C}$ satisfies $\mathrm{Re}(\alpha_i) > 0$. 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 what follows we present an algorithm mathematically equivalent to [6, Algorithm 2] which, when compared with [6, Algorithm 2], has a redistribution of terms which is important for our interpretation of the algorithm.

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 [5].

We note that our analysis of Algorithm 1 is completely based on the underlying optimal control problem and not on the algebraic Riccati equation (1.1) itself. Namely, we use the well-known fact (see, e.g., [15, section III.1.4] or [4, Chapter 16]) that the unique positive semidefinite solution X of the algebraic Riccati equation (1.1) fulfills

*Received by the editors September 8, 2014; accepted for publication (in revised form) by T. Stykel February 23, 2016; published electronically May 17, 2016.
http://www.siam.org/journals/simax/37-2/98579.html

†Fachbereich Mathematik, Universität Hamburg, 20146 Hamburg, Germany (arash.massoudi@uni-hamburg.de, timo.reis@uni-hamburg.de).

‡Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (m.opmeer@maths.bath.ac.uk).

for all $x_0 \in \mathbb{C}^n$,

$$x_0^* X x_0 = \min_{u \in L^2(0,\infty;\mathbb{C}^m)} \int_0^\infty \|u(t)\|^2 + \|y(t)\|^2 \, dt, \tag{1.2}$$

where

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

Algorithm 1. An iteration method for the algebraic Riccati equation.

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 $\alpha_1, \dots, \alpha_k \in \mathbb{C}$ 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 positive semidefinite solution of the algebraic Riccati equation (1.1)

```
$V_1 = (\alpha_1 I - A^*)^{-1} C^*$
$S_1 = \sqrt{2\mathrm{Re}(\alpha_1)} \cdot V_1^*$
$Q_1 = \sqrt{2\mathrm{Re}(\alpha_1)} \cdot V_1^* B$
$L_1 = \frac{1}{\sqrt{2\mathrm{Re}(\alpha_1)}}$
$F_1 = Q_1 L_1$
for $i = 2, 3, \dots, k$ do
    $V_i = V_{i-1} - (\alpha_i + \overline{\alpha_{i-1}}) \cdot (\alpha_i I - A^*)^{-1} V_{i-1}$
    $S_i = [\, S_{i-1}^* \,,\, \sqrt{2\mathrm{Re}(\alpha_i)} \cdot V_i \,]^*$
    $Q_i = [\, Q_{i-1} \,,\, \sqrt{2\mathrm{Re}(\alpha_i)} \cdot V_i^* B \,]$
    $\gamma_i := \sqrt{\frac{\mathrm{Re}(\alpha_j)}{\mathrm{Re}(\alpha_{j-1})}}$
    $M_{i,1} := \begin{bmatrix} \frac{1}{\sqrt{2\mathrm{Re}(\alpha_1)}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{2\mathrm{Re}(\alpha_i)}} \end{bmatrix}, \quad M_{i,2} = \begin{bmatrix} \overline{\alpha_1}+\alpha_i & & & \\ \alpha_1-\alpha_i & \overline{\alpha_2}+\alpha_i & & \\ & & \ddots & \\ & & \alpha_{i-1}-\alpha_i & \overline{\alpha_i}+\alpha_i \end{bmatrix},$
        $M_{i,3} = \begin{bmatrix} 1 & \dots & 1 \\ & \ddots & \vdots \\ & & 1 \end{bmatrix}, \quad M_{i,4} = \begin{bmatrix} 0 & I \\ 1 & 0 \end{bmatrix}, \quad M_{i,5} = \begin{bmatrix} -\sqrt{2\mathrm{Re}(\alpha_1)} & & & \\ & \ddots & & \\ & & -\sqrt{2\mathrm{Re}(\alpha_{i-1})} & \\ & & & 1 \end{bmatrix}$
    $M_i = M_{i,1}^{-1} M_{i,2}^{-1} M_{i,3}^{-1} M_{i,4}^{-1} M_{i,5}^{-1}$
    $L_i = \begin{bmatrix} \gamma_i L_{i-1} & 0 \\ 0 & 0 \end{bmatrix} - M_i \begin{bmatrix} L_{i-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_i(\alpha_i + \overline{\alpha_{i-1}})I & 0 \\ [0, \gamma_i] & -1 \end{bmatrix}$
    $F_i = \begin{bmatrix} [F_{i-1}, 0] \\ Q_i \left(\overline{L_i} \otimes I_m\right) \end{bmatrix}$
end for
```

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

$$\mathscr{V}_k := \mathrm{span}\{t \mapsto \mathrm{e}^{-\alpha_1 t}, \dots, t \mapsto \mathrm{e}^{-\alpha_k t}\}. \tag{1.4}$$

In this introduction, we assume for notational simplicity that the shift parameters α_j are distinct (in the sections following this introduction, we drop this assumption; the definition of $\mathscr{V}_k$ has to be modified in the case of nondistinct parameters). Let

$P_{k,p} : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ denote the orthogonal projection onto $\mathscr{V}_k \otimes \mathbb{C}^p \subset L^2(0,\infty;\mathbb{C}^p)$. The matrix X_k computed by Algorithm 1 gives the optimal cost for an optimal control problem (see Theorems 5.1 and 5.3): for all $x_0 \in \mathbb{C}^n$ there holds

$$x_0^* X_k x_0 = \min_{u \in L^2(0,\infty;\mathbb{C}^m)} \int_0^\infty \|u(t)\|^2 + \|(P_{k,p}y)(t)\|^2 \, dt \tag{1.5}$$

subject to (1.3). In Corollary 5.2 we show that the minimizer in (1.5) fulfills $u_k^{\text{opt}} \in \mathscr{V}_k \otimes \mathbb{C}^m$. This implies that in (1.5) we can equivalently minimize over $\mathscr{V}_k \otimes \mathbb{C}^m$ rather than over all of $L^2(0,\infty;\mathbb{C}^m)$.

It follows immediately from $\mathscr{V}_k \subset \mathscr{V}_{k+1}$ and (1.5) that (Theorem 5.4)

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

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

We obtain (Theorem 6.1) that the approximation computed by Algorithm 1 converges to X, i.e.,

$$\lim_{k\to\infty} X_k = X, \tag{1.6}$$

provided that

$$\overline{\bigcup_{k\in\mathbb{N}} \mathscr{V}_k} = L^2(0,\infty). \tag{1.7}$$

The property (1.7) is proven in [8] to be equivalent to the *non-Blaschke condition*

$$\sum_{j=1}^\infty \frac{\operatorname{Re}(\alpha_j)}{1+|\alpha_j|^2} = \infty. \tag{1.8}$$

We note that this non-Blaschke condition is, for example, satisfied if all of the parameters 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 [8].

As noted above, the approximate solution X_k obtained using Algorithm 1 is identical to that obtained in [6]. That the sequence X_k is nondecreasing is also obtained in [6, Theorem 4.2] by using very different arguments. Convergence of X_k to X is not obtained in [6]. An upper bound for the distance between X_k and X in the gap metric was considered in [6]. However, it was left open there whether or not this upper bound converges to zero.

The following example shows that, even for the case where the Riccati equation reduces to a Lyapunov equation, X_k may converge to something different from X if the non-Blaschke condition (1.8) is not satisfied.

Example 1.1. Let $n = p = m = 1$, $A = -1$, $B = 0$, $C = \sqrt{2}$. Then the unique solution of the algebraic Riccati equation is $X = 1$. From [11, Lemma 4.5] we obtain $X - X_n = X|T_n|^2$, where

$$T_n := \prod_{k=1}^n \frac{A+\alpha_k}{A-\overline{\alpha_k}}.$$

Choosing $\alpha_k := \frac{1}{8k^2-1}$ results in $T_n = \prod_{k=1}^n \left(1 - \frac{1}{4k^2}\right)$. By Wallis' formula (or from Euler's product formula for the sine) we then have $T_n \to \frac{2}{\pi}$ as $n \to \infty$. It follows that

$X_n \to 1 - \frac{4}{\pi^2}$. In particular, X_n does not converge to X. We note that the above sequence of shift parameters does not satisfy the non-Blaschke condition (1.8).

The proof of convergence of the algorithm under the very nonrestrictive non-Blaschke condition (1.8) is the main contribution of this paper and shows the usefulness of the new optimal control interpretation of the algorithm.

In Theorem 7.1, for completeness we briefly consider the extension of the obtained results to the case where $\mathbb{C}^n$, $\mathbb{C}^p$, and $\mathbb{C}^m$ are replaced by (possibly infinite-dimensional) Hilbert spaces. The analysis in fact goes through almost unchanged; the only real difference is that in the infinite-dimensional case we have to specify in what topology the convergence (1.6) takes place.

The remainder of this paper is organized as follows. Section 2 provides the connection between the algebraic Riccati equation and the optimal control problem following the setting of [14]. In section 3 we take a closer look at the space $\mathscr{V}_k$ from (1.4). In particular, we consider an orthonormal basis for this space (the Takenaka–Malmquist system). This is used in section 4 to determine matrix representations of the (projected) input-output and output maps of (1.3). We show that these are exactly the matrices F_k and S_k in Algorithm 1. These results are used in section 5 to derive (1.5). Using this connection, convergence of Algorithm 1 is shown in section 6. Section 7 briefly considers the extension to the infinite-dimensional case. Finally, section 8 illustrates the obtained results using numerical examples arising from a convection-diffusion equation.

2. The linear-quadratic optimal control problem.

In this section, we present the connection between the algebraic Riccati equation (1.1) and the optimal control problem (1.2), (1.3). Specifically, in Proposition 2.2 we give an explicit formula for the solution X based on the operator framework from [14]. This formula is crucial in the derivation and convergence analysis of Algorithm 1. The formula is in terms of certain maps associated to the dynamical system (1.3) which we introduce in the following definition.

Definition 2.1 (output map, input-output map, complementary Popov operator). *Assume that $A \in \mathbb{C}^{n\times n}$ is stable, $B \in \mathbb{C}^{n\times m}$, and $C \in \mathbb{C}^{p\times n}$. Consider the following maps associated to the 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 $u = 0$),*

$$\Psi x_0 = t \mapsto C\mathrm{e}^{At}x_0, \tag{2.1}$$

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

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

- *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 $x_0 = 0$),*

$$\mathbb{F}u = t \mapsto \int_0^t C\mathrm{e}^{A(t-\tau)}Bu(\tau)\,d\tau, \tag{2.2}$$

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

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

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

$$\mathcal{R}_c := I + \mathbb{F}\mathbb{F}^* = \begin{bmatrix} I & \mathbb{F} \end{bmatrix} \begin{bmatrix} I \\ \mathbb{F}^* \end{bmatrix}. \tag{2.3}$$

With the above-introduced mappings, the minimized expression in (1.2) becomes $\|\Psi x_0 + \mathbb{F}u\|_{L^2}^2 + \|u\|_{L^2}^2$. In addition, we note that the complementary Popov operator is bounded, self-adjoint, and positive definite, and has a bounded inverse.

The following proposition characterizes the unique positive semidefinite solution of the algebraic Riccati equation (1.1) in terms of the output map and the complementary Popov operator corresponding to the dynamical system (1.3).

Proposition 2.2. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, and $C \in \mathbb{C}^{p\times n}$. Define Ψ, $\mathbb{F}$, and $\mathcal{R}_c$ by* (2.1), (2.2), (2.3). *The unique minimizer of the optimal control problem* (1.2), (1.3) *is*

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

The optimal cost is given by $x_0^* X x_0$ with

$$X = \Psi^*\mathcal{R}_c^{-1}\Psi. \tag{2.4}$$

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 the optimal cost is induced 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. □

3. Convolution and Takenaka–Malmquist systems.

In this section we consider special subspaces of $L^2(0,\infty)$ which are relevant for our considerations on the optimal control interpretation (1.5) of the iteratively determined matrices X_k in Algorithm 1.

Definition 3.1. *Let $(\alpha_j)_{j=1}^\infty$ be such that* $\mathrm{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$. We define the corresponding* convolution system $(\varphi_j)_{j=1}^\infty$, $\varphi_j \in L^2(0,\infty)$, *by*

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

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

$$\mathscr{K}_k(\alpha) := \mathrm{span}\{\varphi_1, \ldots, \varphi_k\}. \tag{3.1}$$

Remark 3.2. Let $(\alpha_j)_{j=1}^k$ be a tuple of numbers in the open right complex half-plane, and 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

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

and therefore,

$$\widehat{\varphi}_j(s) = \prod_{\ell=1}^{j} \frac{1}{s+\alpha_\ell}. \tag{3.2}$$

(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 $k = \ell_1 + \cdots + \ell_J$). 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

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

(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

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

Next we consider a special basis of $\mathscr{K}_k(\alpha)$.

Definition 3.3. *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*

$$\begin{aligned} \phi_1 &= t \mapsto \mathrm{e}^{-\alpha_1 t}, & \psi_1 &= \sqrt{2\operatorname{Re}(\alpha_1)} \cdot \phi_1, \\ \phi_j &= \phi_{j-1} - (\alpha_j + \overline{\alpha_{j-1}}) \cdot (\mathrm{e}^{-\alpha_j \cdot} * \phi_{j-1}), & \psi_j &= \sqrt{2\operatorname{Re}(\alpha_j)} \cdot \phi_j. \end{aligned} \tag{3.3}$$

Remark 3.4.

(a) The Takenaka–Malmquist system is orthonormal (see, e.g., [8, Appendix B] for a proof).

(b) Laplace transformation of (3.3) yields that for all $s \in \mathbb{C}$ with $\operatorname{Re}(s) > 0$ there holds

$$\begin{aligned} \widehat{\phi}_1(s) &= \frac{1}{s+\alpha_1}, & \widehat{\psi}_1(s) &= \sqrt{2\operatorname{Re}(\alpha_1)} \cdot \widehat{\phi}_1(s), \\ \widehat{\phi}_j(s) &= \widehat{\phi}_{j-1}(s) - (\alpha_j + \overline{\alpha_{j-1}}) \cdot \frac{1}{s+\alpha_j} \cdot \widehat{\phi}_{j-1}(s), & \widehat{\psi}_j(s) &= \sqrt{2\operatorname{Re}(\alpha_j)} \cdot \widehat{\phi}_j(s). \end{aligned} \tag{3.4}$$

Therefore, we obtain by induction that

$$\widehat{\psi}_j(s) = \frac{\sqrt{2\operatorname{Re}(\alpha_j)}}{(s+\alpha_j)} \cdot \prod_{\ell=1}^{j-1} \frac{s - \overline{\alpha_\ell}}{s+\alpha_\ell}. \tag{3.5}$$

(c) By using partial fraction expansions of their Laplace transforms (see (3.2) and (3.5)), we obtain

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

Consequently, $\{\psi_1, \dots, \psi_k\}$ is a basis of $\mathscr{K}_k(\alpha)$.

Now we determine how the operators Ψ^* and $\mathbb{F}^*$ act on the considered bases of $\mathscr{K}_k(\alpha)$. We first define the following two operators (for $t \geq 0$):

$$\Phi_t : L^2(0,\infty;\mathbb{C}^p) \to \mathbb{C}^n, \quad \Phi_t z := \int_t^\infty \mathrm{e}^{A^*(\tau-t)} C^* z(\tau)\, d\tau, \tag{3.6}$$

$$\Lambda : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^n), \quad \Lambda z := t \mapsto \int_t^\infty \mathrm{e}^{A^*(\tau-t)} C^* z(\tau)\, d\tau. \tag{3.7}$$

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 used to show how the operators Ψ^* and $\mathbb{F}^*$ act on the convolution and Takenaka–Malmquist systems.

Lemma 3.5. *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 (3.6). Then for $\mu \in \mathbb{C}$ with* $\mathrm{Re}(\mu) > 0$*, $v \in \mathbb{C}^p$, and $z \in L^2(0,\infty;\mathbb{C}^p)$, there holds*

$$\Phi_t(\mathrm{e}^{-\mu\cdot} v) = (\mu I - A^*)^{-1} C^* v \mathrm{e}^{-\mu t} \tag{3.8}$$

and

$$\Phi_t(\mathrm{e}^{-\mu\cdot} * z) = (\mu I - A^*)^{-1} C^* (\mathrm{e}^{-\mu\cdot} * z)(t) + (\mu I - A^*)^{-1} \Phi_t(z). \tag{3.9}$$

Proof. We first consider (3.8). We have, by the change of variables $\theta := \tau - t$,

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

and elementary integration then gives the result.

We now consider (3.9). We have

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

Interchanging the order of integration gives that the above equals

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

as claimed. □

A consequence of the above lemma is a result regarding the action of Φ_t, Ψ^*, and Λ on the convolution system.

PROPOSITION 3.6. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, $(\alpha_j)_{j=1}^{\infty}$ such that* $\operatorname{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$, $(\varphi_j)_{j=1}^{\infty}$ as in Definition* 3.1, *and $v \in \mathbb{C}^p$.*

(a) *Let $t \geq 0$. With Φ_t as in* (3.6) *there holds*

$$\Phi_t(\varphi_1 v) = (\alpha_1 I - A^*)^{-1} C^* v \varphi_1(t),$$
$$\Phi_t(\varphi_j v) = (\alpha_j I - A^*)^{-1} C^* v \varphi_j(t) + (\alpha_j I - A^*)^{-1} \Phi_t(\varphi_{j-1} v).$$

(b) *With Ψ as in* (2.1) *there holds*

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

(c) *With Λ as in* (3.7) *there holds*

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

Proof. We first prove part (a). The first formula follows directly from (3.8) with $\mu := \alpha_1$. The second formula follows from multiplying the iterative definition of $(\varphi_j)_{j=1}^{\infty}$ from Definition 3.1 by v, applying Φ_t to the result, and using that by Lemma 3.5,

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

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 $\Lambda z = t \mapsto \Phi_t z$. ◻

We can immediately conclude from the previous result and $\mathbb{F}^* = B^*\Lambda$ that $\mathscr{K}_k(\alpha)$ is, in a certain sense, an invariant subspace of the adjoint input-output map.

COROLLARY 3.7. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, and $(\alpha_j)_{j=1}^{\infty}$ such that* $\operatorname{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$. Then for $\mathscr{K}_k(\alpha)$ as in Definition* 3.1 *there holds*

$$\mathbb{F}^*(\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p) \subset (\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p). \tag{3.10}$$

Now we describe the action of Φ_t, Ψ^*, and Λ on the Takenaka–Malmquist system.

PROPOSITION 3.8. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, $(\alpha_j)_{j=1}^{\infty}$ such that* $\operatorname{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$, $(\phi_j)_{j=1}^{\infty}$ and $(\psi_j)_{j=1}^{\infty}$ as in Definition* 3.3 *and $v \in \mathbb{C}^p$.*

(a) *With Ψ as in* (2.1) *there holds*

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

(b) *For $j > 1$ and with Λ as in* (3.7) *and $\gamma_j := \frac{\operatorname{Re}(\alpha_j)}{\operatorname{Re}(\alpha_{j-1})}$ there holds*

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

Proof. We first prove part (a). The first equation follows from (3.8) with $\mu := \alpha_1$ using that $\Psi^* = \Phi_0$. The second equation is obtained by multiplying (3.3) by v, applying Ψ^* to the result, and using that by Lemma 3.5 (using that $\Psi^* = \Phi_0$),

$$\Psi^*(\mathrm{e}^{-\alpha_j \cdot} * \phi_{j-1} v) = (\alpha_j I - A^*)^{-1} \Psi^*(\phi_{j-1} v). \tag{3.11}$$

We now prove part (b). From (3.3) 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 3.5 we obtain that

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

and the desired result follows. □

4. Matrix representations.

In this section we use the results from section 3 to show that Algorithm 1 computes matrix representations of (projection times) $\mathbb{F}$ and Ψ with respect to the Takenaka–Malmquist system.

We first introduce the canonical embeddings associated to the Takenaka–Malmquist system.

DEFINITION 4.1. *Let* $(\alpha_j)_{j=1}^{\infty}$ *be such that* $\mathrm{Re}(\alpha_j) > 0$ *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* (3.3). *For* $k \in \mathbb{N}$, *the mapping* ι_k *is defined by*

$$\begin{aligned} \iota_k : \quad & \mathbb{C}^k \to L^2(0,\infty), \\ & x \mapsto \sum_{j=1}^{k} x_j \cdot \psi_j. \end{aligned} \tag{4.1}$$

Further, for $\ell \in \mathbb{N}$ *and denoting the identity matrix* $I_\ell \in \mathbb{C}^{\ell \times \ell}$, *we set*

$$\iota_{k,\ell} := \iota_k \otimes I_\ell : \ \mathbb{C}^{k\ell} \to L^2(0,\infty;\mathbb{C}^\ell).$$

It follows immediately from the orthonormality of the Takenaka–Malmquist system that $\iota_{k,\ell}$ defines an isometric embedding. In particular, the operator

$$P_{k,\ell} := \iota_{k,\ell} \iota_{k,\ell}^* : L^2(0,\infty;\mathbb{C}^\ell) \to L^2(0,\infty;\mathbb{C}^\ell)$$

is the orthogonal projector onto $\mathscr{V}_k \otimes \mathbb{C}^\ell$. With operators Ψ and $\mathbb{F}$ as in (2.1) and (2.2), we define the operators

$$\Psi_k : \ \mathbb{C}^n \to L^2(0,\infty;\mathbb{C}^p), \qquad \Psi_k = P_{k,p}\Psi, \tag{4.2}$$

$$\mathbb{F}_k : \ L^2(0,\infty;\mathbb{C}^m) \to L^2(0,\infty;\mathbb{C}^p), \qquad \mathbb{F}_k = P_{k,p}\mathbb{F}. \tag{4.3}$$

We further introduce the matrices

$$S_k = \iota_{k,p}^* \Psi \in \mathbb{C}^{kp \times n}, \tag{4.4}$$

$$F_k = \iota_{k,p}^* \mathbb{F} \iota_{k,m} \in \mathbb{C}^{kp \times km}. \tag{4.5}$$

It follows from (4.2) that

$$\Psi_k = P_{k,p}\Psi = \iota_{k,p}\iota_{k,p}^*\Psi = \iota_{k,p} S_k.$$

We conclude that the matrix S_k as in (4.4) 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 $\mathbb{C}^p$. Further, with the matrix F_k as in (4.5) and $\mathbb{F}_k$ as in (4.3) we have

$$\iota_{k,p} F_k = P_{k,p} \mathbb{F} \iota_{k,m} = \mathbb{F}_k \iota_{k,m},$$

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, $\mathbb{C}^p$).

We now proceed to develop a recursive determination of S_k and F_k. These results will imply that S_k and F_k in (4.4) and (4.5) are indeed the matrices computed in Algorithm 1.

THEOREM 4.2. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, and $(\alpha_j)_{j=1}^\infty$ such that* $\mathrm{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$. Then for all $k \in \mathbb{N}$, the matrix S_k determined by Algorithm* 1 *fulfills* (4.4)*.*

Proof. By Algorithm 1, we have

$$S_k = \left[\sqrt{2\mathrm{Re}(\alpha_1)} \cdot V_1 \quad \dots \quad \sqrt{2\mathrm{Re}(\alpha_k)} \cdot V_k\right]^*,$$

where the sequence (V_k) is recursively defined by

$$V_1 = (\alpha_1 I - A^*)^{-1} C^*, \qquad V_k = V_{k-1} - (\alpha_k + \overline{\alpha_{k-1}}) \cdot (\alpha_k I - A^*)^{-1} V_{k-1}. \tag{4.6}$$

The result then follows from Proposition 3.8(a) together with the definition of the Takenaka–Malmquist system in (3.3).

We note that Theorem 4.2 was already established in [8], where the case $B = 0$ (for which the Riccati equation becomes a Lyapunov equation) was considered. □

In the following we prove that the matrix F_k in (4.5) is as determined in Algorithm 1.

THEOREM 4.3. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$, and $(\alpha_j)_{j=1}^\infty$ such that* $\mathrm{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$. Then for all $k \in \mathbb{N}$, the matrix F_k determined by Algorithm* 1 *fulfills* (4.5)*.*

Proof. The proof is given in the appendix. □

5. The projected optimal control problem.

In this section we consider the optimal control problem (1.5), (1.3). By using that, by the results of the previous section, the matrices F_k and S_k in Algorithm 1 are matrix representations of the (projected) input-output and output mappings of the system (1.3), we show that Algorithm 1 indeed provides the solution of (1.5), (1.3). These results will be essential for our convergence analysis in section 6.

We first present a projected version of Proposition 2.2.

THEOREM 5.1. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$, and $(\alpha_j)_{j=1}^\infty$ such that* $\mathrm{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$. Define Ψ_k and $\mathbb{F}_k$ by* (4.2) *and* (4.3)*, where $P_{k,p} : L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p)$ is the orthogonal projection onto $\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$ with $\mathscr{K}_k(\alpha)$ as in Definition* 3.1*. In addition, define the* projected complementary Popov operator *by*

$$\mathcal{R}_{c,k} : \ L^2(0,\infty;\mathbb{C}^p) \to L^2(0,\infty;\mathbb{C}^p), \quad \mathcal{R}_{c,k} = I + \mathbb{F}_k \mathbb{F}_k^* = [I \ \ \mathbb{F}_k] \begin{bmatrix} I \\ \mathbb{F}_k^* \end{bmatrix}. \tag{5.1}$$

The unique minimizer of the optimal control problem (1.5), (1.3) *is given by*

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

The optimal cost is given by $x_0^* X_k x_0$ *with*

$$X_k = \Psi_k^* \mathcal{R}_{c,k}^{-1} \Psi_k. \tag{5.2}$$

Proof. Noting that $P_{k,p} y = \Psi_k x_0 + \mathbb{F}_k u$, 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

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

In particular, we have for $X_k = \Psi_k^* \mathcal{R}_{c,k}^{-1} \Psi_k$ that $\|u\|_{L^2}^2 + \|P_{k,p} y\|_{L^2}^2 \geq x_0^* X_k x_0$. 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 $\|u\|_{L^2}^2 + \|P_{k,p} y\|_{L^2}^2$ and the quadratic form $x_0^* X_k x_0$ in this case. □

COROLLARY 5.2. *Under the assumptions and with the notation of Theorem* 5.1, *we have*

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

Proof. By Theorem 5.1 we have $u_k^{\text{opt}} = \mathbb{F}^* z$ for $z := -P_{k,p} \mathcal{R}_{c,k}^{-1} \Psi_k x_0 \in \mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$. From Corollary 3.7 we see that $\mathbb{F}^*$ maps $\mathscr{K}_k(\alpha) \otimes \mathbb{C}^p$ into $\mathscr{K}_k(\alpha) \otimes \mathbb{C}^m$. Therefore $u_k^{\text{opt}} \in \mathscr{K}_k(\alpha) \otimes \mathbb{C}^m$, as desired. □

Next we show that the matrix X_k in Theorem 5.1 is indeed the one determined in Algorithm 1.

THEOREM 5.3. *Let* $A \in \mathbb{C}^{n \times n}$ *be stable,* $B \in \mathbb{C}^{n \times m}$, $C \in \mathbb{C}^{p \times n}$, *and* $(\alpha_j)_{j=1}^{\infty}$ *such that* $\operatorname{Re}(\alpha_j) > 0$ *for all* $j \in \mathbb{N}$. *Then for all* $k \in \mathbb{N}$, *the matrix* X_k *determined by Algorithm* 1 *fulfills* (5.2).

Proof. By Corollary 3.7 we have

$$P_{k,m} \mathbb{F}^* \iota_{k,p} = \mathbb{F}^* \iota_{k,p}.$$

It follows that

$$\begin{aligned} I_{kp} + (\mathbb{F}^* \iota_{k,p})^* \cdot (\mathbb{F}^* \iota_{k,p}) &= I_{kp} + (\mathbb{F}^* \iota_{k,p})^* P_{k,m} (\mathbb{F}^* \iota_{k,p}) \\ &= I_{kp} + (\mathbb{F}^* \iota_{k,p})^* \iota_{k,m} (\iota_{k,m}^* \mathbb{F}^* \iota_{k,p}) \\ &= I_{kp} + (\iota_{k,m}^* \mathbb{F}^* \iota_{k,p})^* \cdot (\iota_{k,m}^* \mathbb{F}^* \iota_{k,p}) \\ &= I_{kp} + F_k F_k^*, \end{aligned} \tag{5.3}$$

and thus

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

Using the above relation, together with the definition of S_k and F_k as in (4.4) and (4.5), we obtain

$$\begin{aligned}\Psi_k^*\mathcal{R}_{c,k}^{-1}\Psi_k &= S_k\iota_{k,p}P_k\mathcal{R}_{c,k}^{-1}P_k\iota_{k,p}^*S_k^*\\ &= S_k\iota_{k,p}\iota_{k,p}(I+F_kF_k^*)^{-1}\iota_{k,p}^*\iota_{k,p}^*S_k^*\\ &= S_k(I+F_kF_k^*)^{-1}S_k^*.\end{aligned}$$

The desired statement can now be concluded using Theorems 4.2 and 4.3. □

Theorems 5.3 and 5.1 imply that the matrix X_k computed by Algorithm 1 indeed expresses the optimal cost (1.5) of the projected optimal control problem (1.5), (1.3). Since the ranges of the projectors $P_{k,p}$ are nested, we can easily deduce that the sequence (X_k) is monotone and bounded from above by X with respect to semidefiniteness.

THEOREM 5.4. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, $C \in \mathbb{C}^{p\times n}$, and $(\alpha_j)_{j=1}^\infty$ such that* $\operatorname{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$. Let $X \in \mathbb{C}^{n\times n}$ be the unique positive semidefinite solution of the algebraic Riccati equation* (1.1)*. Let the sequence (X_k) be determined by Algorithm* 1*. Then for all $k \in \mathbb{N}$ there holds*

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

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

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

since $\mathscr{K}_k(\alpha) \subset \mathscr{K}_{k+1}(\alpha)$. It follows from Theorems 5.1 and 5.3 that

$$\begin{aligned}x_0^*X_kx_0 = &\min_{u\in L^2(0,\infty;\mathbb{C}^m)} \|u\|^2 + \|P_{k,p}y\|^2\\ &\le \inf_{u\in L^2(0,\infty;\mathbb{C}^m)} \|u\|^2 + \|P_{k+1,p}y\|^2 = x_0^*X_{k+1}x_0.\end{aligned}$$

Similarly, using that

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

we obtain from (1.2) (cf. Proposition 2.2) that

$$x_0^*X_kx_0 \le x_0^*Xx_0.$$ □

6. Convergence of the algorithm.

The following theorem gives convergence of Algorithm 1.

THEOREM 6.1. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $B \in \mathbb{C}^{n\times m}$, and $C \in \mathbb{C}^{p\times n}$. Let $(\alpha_j)_{j=1}^\infty$ be such that* $\operatorname{Re}(\alpha_j) > 0$ *for all $j \in \mathbb{N}$. For $k \in \mathbb{N}$, let (X_k) be the sequence obtained by Algorithm* 1*. Then (X_k) converges as $k \to \infty$. If $(\alpha_j)_{j=1}^\infty$ satisfies the non-Blaschke condition* (1.8)*, then (X_k) converges to X, the unique positive semidefinite solution of the algebraic Riccati equation* (1.1)*.*

Proof. Since, by Theorem 5.4, (X_k) is a nondecreasing sequence which is bounded from above by X, we obtain convergence to some matrix $Q = Q^* \in \mathbb{C}^{n\times n}$ with $Q \le X$.

Since $\mathscr{K}_k(\alpha) \subset \mathscr{K}_{k+1}(\alpha)$ we have $P_{k,p} \le P_{k+1,p}$. Since $P_{k,p}$ is an orthogonal projection, we have $P_{k,p} \le I$. It follows from [12, p. 263] that $P_{k,p}$ converges strongly to some orthogonal projection P with $P \leq I$. It was shown in [8, 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 $L^2(0, \infty)$, but using the tensor product of $\{\psi_1, \ldots, \psi_k\}$ and the canonical basis of $\mathbb{C}^p$, as in Definition 4.1, one can prove the general case on $L^2(0, \infty; \mathbb{C}^p)$).

From now on we assume the non-Blaschke condition, so that $P = I$. Then $(\mathcal{R}_{c,k}) = (I + \mathbb{F}P_{k,p}\mathbb{F}^*)$ converges strongly to $I + \mathbb{F}\mathbb{F}^* = \mathcal{R}_c$. As a result, we have strong convergence of $(\mathcal{R}_{c,k}^{-1})$ to $\mathcal{R}_c^{-1}$ (e.g., by [2, Theorem 7.6.1]). By sequential continuity, we then have strong convergence of $(\Psi^* P_{k,p}\mathcal{R}_{c,k}^{-1}P_{k,p}\Psi)$ to $\Psi^*\mathcal{R}_c^{-1}\Psi$. We have $X_k = \Psi^* P_{k,p}\mathcal{R}_{c,k}^{-1}P_{k,p}\Psi$ by Theorem 5.1 and $X = \Psi^*\mathcal{R}_c^{-1}\Psi$ by Proposition 2.2. Hence, the sequence (X_k) converges strongly to X. The finite-dimensionality of $\mathbb{C}^{n\times n}$ then implies that the sequence (X_k) converges to X. □

Remark 6.2. The choice of shift parameters is essential for the speed of convergence. In [6, 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 $A - BB^*X$ [16, Chap. 13], we have, in the 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

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

As a consequence, for this particular choice of shift parameters the projected optimal control problem (1.5) coincides with the original optimal control problem, so that $X = X_n$.

In [6, 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 Lu and Wachspress [7] 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 for any N steps. By convergence of (X_k) to X (established in Theorem 6.1), these parameters converge to the eigenvalues of $A - BB^*X$ (which coincide with the stable eigenvalues of the Hamiltonian).

The efficient numerical computation of dominant stable eigenvalues of a Hamiltonian matrix seems not to have been explored so far. The considered iteration method for Riccati equations would be an application for this research area.

7. Extension to infinite-dimensional spaces.

In this section, we formulate the extension of Theorem 6.1 to the infinite-dimensional spaces. We refer the reader to [13] for the terminology used in the statement of the following theorem. Note that Theorem 6.1 is a special case of Theorem 7.1 where by finite-dimensionality the topology in which convergence occurs is irrelevant.

THEOREM 7.1. *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* $j \in \mathbb{N}$, *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* $\mathscr{K}_k(\alpha) \otimes \mathscr{Y}$ *with* $\mathscr{K}_k(\alpha)$ *as in* (3.1). *Define* X_k *by* (5.2). *Then* X_k *converges in the strong operator topology as* $k \to \infty$. *Let* X *be given by* (2.4) *and assume that* $(\alpha_j)_{j=1}^{\infty}$ *satisfies the non-Blaschke condition* (1.8). *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 paper hold in the setting of this theorem (with essentially the same proofs) if we interpret the superscript $*$ as the adjoint with respect to the given inner-products and if expressions such as $x_0^* Y x_0$ are interpreted as $\langle Y x_0, x_0\rangle$, where the latter denotes the given inner-product. Also, all of the claims in Theorem 7.1, except those in the last sentence, follow as in the proof of Theorem 6.1. Therefore, it only remains to show the claims in the last sentence.

If X is compact, then (since $\mathcal{R}_c$ is self-adjoint and invertible) Ψ is compact. As in the proof of Theorem 6.1 we have that $P_{k,p}\mathcal{R}_{c,k}^{-1}P_{k,p}$ converges in the strong operator topology to $\mathcal{R}_c^{-1}$. From [8, Theorem A.2, part a] (which is a slight modification of [3, Theorem III.6.3]) we then obtain that $P_{k,p}\mathcal{R}_{c,k}^{-1}P_{k,p}\Psi_k$ 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., [8, Appendix A] for the needed relation between Schatten class membership of X and of Ψ). ∎

8. Numerical results.

We present a numerical example to show the applicability of the algorithm and to demonstrate the expected behavior of Algorithm 1 in terms of monotonicity and convergence. All the calculations were done using MATLAB 8.5 (R2015a) on a 64-bit server with 24 CPU cores (using hyperthreading) of type Intel Xeon X5650 at 2.67GHz and 48 GB main memory available.

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

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

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 $a = 1$. 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. An example of the grid

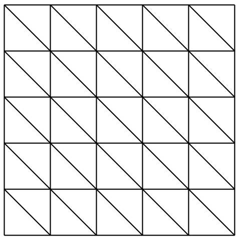

FIG. 1. *An example of the chosen triangular element for $N = 6$.*

(for $N = 6$) that we used in our computations is shown in Figure 1. 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

$$E\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad y(t) = Cx(t), \tag{8.2}$$

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 nonsymmetric matrix, $B \in \mathbb{R}^{n\times 1}$ is the input matrix, and $C \in \mathbb{R}^{2\times n}$ the output matrix.

Remark 8.1. For invertible $E \in \mathbb{C}^{n\times n}$, the unique positive semidefinite solution of the algebraic Riccati equation

$$A^*XE + E^*XA + C^*C - E^*XBB^*XE = 0 \tag{8.3}$$

satisfies

$$x_0^* E^* X E x_0 = \min_{u\in L^2(0,\infty;\mathbb{C}^m)} \int_0^\infty \|u(t)\|^2 + \|y(t)\|^2 \, dt$$

subject to (8.2). If in Algorithm 1 we make the replacements

```
1:    $V_1 = (\alpha_1 E^* - A^*)^{-1} C^*$,
7:    $V_i = V_{i-1} - (\alpha_i + \overline{\alpha_{i-1}}) \cdot (\alpha_i E^* - A^*)^{-1} E^* V_{i-1}$,
```

then for X_k as computed by this adapted version of Algorithm 1 we have

$$x_0^* E^* X_k E x_0 = \min_{u\in L^2(0,\infty;\mathbb{C}^m)} \int_0^\infty \|u(t)\|^2 + \|(P_{k,p}y)(t)\|^2 \, dt$$

subject to (8.2). Therefore the convergence results in this paper remain valid for this modification if we replace (1.1) by (8.3).

We note that if E is the positive definite mass matrix of a finite element discretization, then the expression $x_0^* E^* X E x_0$ equals $\langle XEx_0, x_0\rangle$, where the inner-product in this latter expression is the one induced by the underlying function space.

We consider $N = 60$ (so that $n = 3600$) and solve the algebraic Riccati equation corresponding to the system (8.2) once using the "care" routine of MATLAB and once using Algorithm 1 with the modifications in Remark 8.1. We note that although "care" does work for the example considered, the computation takes two hours. For comparison, Algorithm 1 requires just 20 seconds. We denote by X the solution obtained from the "care" routine and use it as a reference for the comparisons with the solution obtained by Algorithm 1 (denoted by X_k).

The choice of the shift parameters has a major effect on the convergence speed of Algorithm 1. We first illustrate that if the shift parameters do not satisfy the non-Blaschke condition (1.8), then the matrix X_k obtained by Algorithm 1 may converge to a positive semidefinite matrix, which is not a solution of the algebraic Riccati equation corresponding to the system (8.2) (cf. Theorem 6.1). Toward 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 [9, 10] on the matrix pencil $\lambda E - A$. The underlying Arnoldi process is initialized with a random vector in $\mathbb{R}^n$. We compute 32 Ritz values by the Arnoldi process to approximate the eigenvalues of the matrix pencil $\lambda E - A$. Out of these 32 Ritz values, 11 values are calculated using the inverse Arnoldi method (to increase the accuracy of approximation). By this choice we generate a set of 10 shift parameters, which we reuse 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 (1.8).
2. As a second set of shift parameters, we choose the infinite sequence $\alpha_i = i^3$, $i = 1, 2, \dots,$ for which the non-Blaschke condition is not satisfied.

We perform the simulation using the two sets of shift parameters which we introduced above, and at each iteration k we observe the absolute residual norm using the approach proposed in [6, sect. 3.3]. That is, we exploit the low-rank form of the approximate solution $X_k = S_k(I + F_k F_k^*)^{-1} S_k^*$ to calculate the residual norm. Figure 2 shows the absolute residual norm with respect to the iteration for problem dimension $n = 3600$.

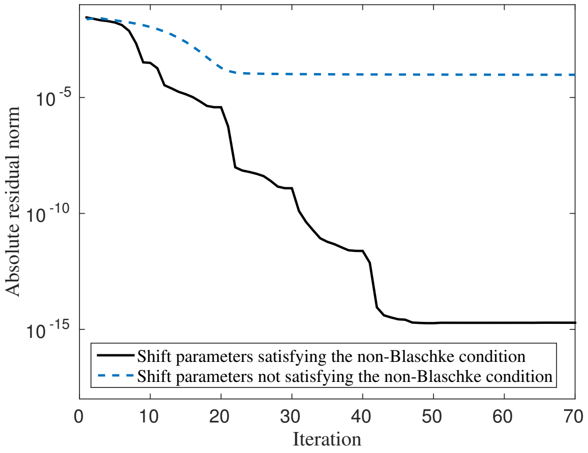

FIG. 2. *Comparison of two sets of shift parameters for Algorithm* 1: *convection-diffusion equation with the state-space dimension* $n = 3600$.

Considering Figure 2, we observe that by choosing the second set of shift parameters, $\alpha_i = i^3$, our sequence converges to a matrix which is not a solution of the corresponding algebraic Riccati equation. In addition, with a tolerance of 10^{-14} on the absolute residual norm, the first choice of shift parameters provides convergence to the desired solution in fewer than 50 iterations for state-space dimension $n = 3600$. We use the first set of shift parameters to continue with further analyses in our example.

In order to illustrate the monotonicity of Algorithm 1, which we have proven in Theorem 5.4, we compute the traces of X and X_k. The trace of X_k can be computed efficiently using the low-rank factors. Specifically, we compute the Cholesky factorization of $I + F_k F_k^* = U_k^* U_k$, and therefore we obtain

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

where $\|\cdot\|_F$ denotes the Frobenius norm. From Figure 3, we observe that trace (X_k) is a nondecreasing function of the iteration k. In addition, we have that trace $(X_k) \leq$ trace (X) for all $k \in \mathbb{N}$.

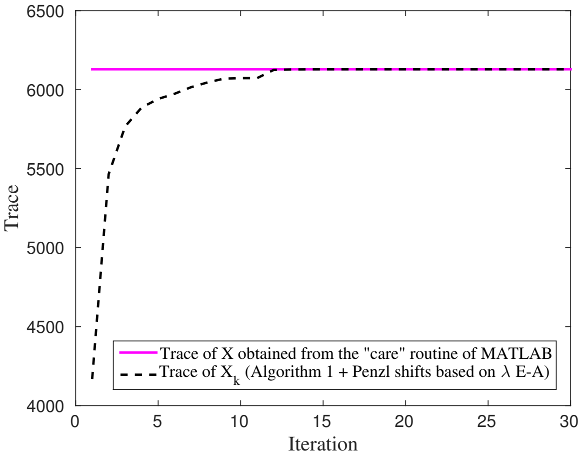

FIG. 3. *Monotonicity of Algorithm* 1*: convection-diffusion equation with the state space dimension* $n = 3600$.

We finish our analyses by observing the relative 2-norm difference $\frac{\|X_k - X\|_2}{\|X\|_2}$ at every iteration to show the convergence behavior of Algorithm 1. Figure 4 shows the relative 2-norm difference of the solutions obtained by Algorithm 1 with respect to the solution obtained by the "care" routine in MATLAB. Note that since the matrices X_k and X are self-adjoint, their 2-norm difference equals the absolute value of the largest eigenvalue of $(X_k - X)$. This eigenvalue can be approximated efficiently using a power iteration without forming the product $X_k = S_k(I + F_k F_k^*)^{-1} S_k^*$ (see, e.g., [1]).

9. Conclusions. The purpose of this paper was to show the convergence of Algorithm 1 which is missing in [6]. Toward this end, we established the connection between this algorithm and the underlying linear-quadratic optimal control problem. We considered the main operator A to be stable, so that the output map Ψ and the input-output map $\mathbb{F}$ are bounded. This allows for the use of an explicit formula for the solution of the algebraic Riccati equation in terms of Ψ and $\mathbb{F}$. The link to the optimal control problem was established by considering a sequence of subspaces of $L^2(0, \infty)$. We chose the Takenaka–Malmquist basis for these subspaces which allowed us to construct matrix representations for the (projected) solution maps associated to the dynamical system (1.3). The sequence of subspaces is determined by the choice of the shift parameters. We showed that if these shift parameters satisfy the non-Blaschke condition, then the matrix calculated by Algorithm 1 converges to the unique positive semidefinite solution of the algebraic Riccati equation. Furthermore, the sequence of approximate solutions is monotonically nondecreasing with respect to

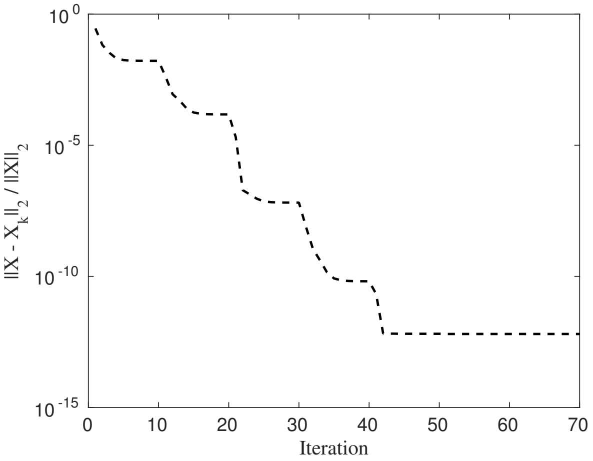

FIG. 4. *The relative* 2*-norm difference of the solution obtained by Algorithm* 1 *with respect to the solution obtained by the "care" routine in MATLAB: convection-diffusion equation with the state-space dimension* $n = 3600$.

definiteness. Finally, we noted that our analyses can be extended to the case where the input, state, and output spaces are infinite-dimensional Hilbert spaces.

Appendix A. Proof of Theorem 4.3. Using that, by Corollary 3.7, the invariance $\mathbb{F}^*\left(\mathscr{K}_{k-1}(\alpha)\otimes\mathbb{C}^p\right)\subset\mathscr{K}_{k-1}(\alpha)\otimes\mathbb{C}^m$ holds true, we see that

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

Thus, we obtain that F_k has the form

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

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.

We need the following additional bases for $\mathscr{K}_k(\alpha)$.

DEFINITION A.1. *Let $k \in \mathbb{N}$ with $k > 1$. Let $(\alpha_j)_{j=1}^k$ be such that* $\mathrm{Re}(\alpha_j) > 0$ *for all $j \in \{1,\ldots,k\}$. Let $(\psi_j)_{j=1}^{k-1}$ be as in Definition* 3.3. *Define the functions $z_j \in L^2(0,\infty)$ and $x_j \in L^2(0,\infty)$ for $j \in \{1,\ldots,k\}$ by*

$$z_j = \begin{cases} \psi_j, & j \in \{1,\ldots,k-1\}, \\ \mathrm{e}^{-\alpha_k\cdot} * \psi_{k-1}, & j = k, \end{cases}$$

$$x_j = \begin{cases} \mathrm{e}^{-\alpha_k\cdot} * \psi_j, & j \in \{1,\ldots,k-1\}, \\ \mathrm{e}^{-\alpha_k\cdot}, & j = k. \end{cases}$$

The following lemma shows how $(z_j)_{j=1}^k$ and $(x_j)_{j=1}^k$ can be obtained as linear combinations of $(\psi_j)_{j=1}^k$.

LEMMA A.2. *Let $k \in \mathbb{N}$ with $k > 1$. Let $(\alpha_j)_{j=1}^k$ be such that* $\mathrm{Re}(\alpha_j) > 0$ *for all* $j \in \{1, \dots, k\}$. *Let $(\psi_j)_{j=1}^k$ as in Definition* 3.3, *and let $(z_j)_{j=1}^k$ and $(x_j)_{j=1}^k$ as in Definition* A.1. *Define $\gamma_k := \sqrt{\frac{\mathrm{Re}(\alpha_k)}{\mathrm{Re}(\alpha_{k-1})}}$ and*

$$T_k = \begin{bmatrix} I_{k-1} & 0 \\ \begin{bmatrix} 0 & \frac{1}{\alpha_k + \overline{\alpha_{k-1}}} \end{bmatrix} & \frac{-1}{\gamma_k(\alpha_k + \overline{\alpha_{k-1}})} \end{bmatrix}, \tag{A.2}$$

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

with

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

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

Then for all $j \in \{1, \dots, k\}$,

$$z_j = \sum_{\ell=1}^{k} (T_k)_{j\ell} \psi_\ell, \qquad x_j = \sum_{\ell=1}^{k} (M_k)_{j\ell}^T \psi_\ell. \tag{A.3}$$

Proof. The equality involving z_j follows by first noting that only the equality involving z_k is nontrivial and then using the recursive definition of the Takenaka–Malmquist basis:

$$\begin{aligned} z_k &= \mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1} = \sqrt{2\mathrm{Re}(\alpha_{k-1})} \mathrm{e}^{-\alpha_k \cdot} * \phi_{k-1} = \sqrt{2\mathrm{Re}(\alpha_{k-1})} \frac{\phi_{k-1} - \phi_k}{\alpha_k + \overline{\alpha_{k-1}}} \\ &= \frac{1}{\alpha_k + \overline{\alpha_{k-1}}} \left(\psi_{k-1} - \frac{\sqrt{2\mathrm{Re}(\alpha_{k-1})}}{\sqrt{2\mathrm{Re}(\alpha_k)}} \psi_k \right) = \frac{1}{\alpha_k + \overline{\alpha_{k-1}}} \psi_{k-1} + \frac{1}{\gamma_k(\alpha_k + \overline{\alpha_{k-1}})} \psi_k. \end{aligned}$$

Applying the Laplace transform, we see that the second equality in (A.3) is equivalent to

$$\widehat{x}_j = \sum_{\ell=1}^{k} (M_k)_{j\ell}^T \widehat{\psi}_\ell.$$

From Definition A.1 we obtain, using the Laplace transform,

$$\widehat{x}_j(s) = \begin{cases} \frac{1}{s+\alpha_k} \widehat{\psi}_j(s), & j \in \{1, \dots, k-1\}, \\ \frac{1}{s+\alpha_k}, & j = k. \end{cases}$$

Therefore, the second equality in (A.3) (for all $j \in \{1, \dots, k\}$) is equivalent to (for all s with $\mathrm{Re}(s) > 0$)

$$\begin{bmatrix} \frac{1}{s+\alpha_k} \widehat{\psi}_1(s) \\ \vdots \\ \frac{1}{s+\alpha_k} \widehat{\psi}_{k-1}(s) \\ \frac{1}{s+\alpha_k} \end{bmatrix} = M_k^T \begin{bmatrix} \widehat{\psi}_1(s) \\ \vdots \\ \widehat{\psi}_{k-1}(s) \\ \widehat{\psi}_k(s) \end{bmatrix}. \tag{A.4}$$

We have, by (3.5),

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

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

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

Taking transposes, this establishes (A.4). We note that this argumentation is similar to the proof of [6, Proposition 3.2]. □

LEMMA A.3. *Let $A \in \mathbb{C}^{n\times n}$ be stable, $C \in \mathbb{C}^{p\times n}$, and $k \in \mathbb{N}$ with $k > 1$, and let $(\alpha_j)_{j=1}^k$ be such that $\mathrm{Re}(\alpha_j) > 0$ for all $j \in \{1, \dots, k\}$. Let $(\psi_j)_{j=1}^k$ as in Definition 3.3, and let $(z_j)_{j=1}^k$ and $(x_j)_{j=1}^k$ as in Definition A.1. Let Ψ as in (2.1) and Λ as in (3.7). Assume that there exists an $L_{k-1} \in \mathbb{C}^{(k-1)\times(k-1)}$ such that for all $v \in \mathbb{C}^p$,*

$$\Lambda(\psi_{k-1}v) = \sum_{j=1}^{k-1} \Psi^*(\psi_j v) \sum_{\ell=1}^{k-1} (L_{k-1})_{j\ell}\psi_\ell. \tag{A.5}$$

Then for all $v \in \mathbb{C}^p$,

$$\Lambda(\psi_k v) = \gamma_k \Lambda(\psi_{k-1}v) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \sum_{j=1}^{k} \Psi^*(x_j v) \sum_{\ell=1}^{k} (\widetilde{L}_{k-1})_{j\ell} z_\ell, \tag{A.6}$$

where

$$\widetilde{L}_{k-1} := \begin{bmatrix} L_{k-1} & 0 \\ 0 & 1 \end{bmatrix}, \qquad \gamma_k := \sqrt{\frac{\mathrm{Re}(\alpha_k)}{\mathrm{Re}(\alpha_{k-1})}}.$$

Proof. We obtain from Proposition 3.8(b) that

$$\begin{aligned} \Lambda(\psi_k v) &= \gamma_k \Lambda(\psi_{k-1}v) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \\ &\quad \cdot \left[(\alpha_k I - A^*)^{-1} C^* v \cdot (\mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}) + (\alpha_k I - A^*)^{-1} \Lambda(\psi_{k-1}v) \right]. \end{aligned} \tag{A.7}$$

Substituting (A.5) in (A.7) gives

$$\begin{aligned}\text{(A.8)}\quad \Lambda(\psi_k v) &= \gamma_k \Lambda(\psi_{k-1} v) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \\ &\cdot \left[(\alpha_k I - A^*)^{-1} C^* v \cdot (\mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}) + \sum_{j=1}^{k-1} (\alpha_k I - A^*)^{-1} \Psi^*(\psi_j v) \sum_{\ell=1}^{k-1} (L_{k-1})_{j\ell} \psi_\ell \right].\end{aligned}$$

From (3.8) with $\mu := \alpha_k$ and $t = 0$ (noting that $\Psi^* = \Phi_0$) and (3.11), we have

$$\text{(A.9)}\qquad \begin{aligned}(\alpha_k I - A^*)^{-1} C^* v &= \Psi^*(\mathrm{e}^{-\alpha_k \cdot} v), \\ (\alpha_k I - A^*)^{-1} \Psi^*(\psi_j v) &= \Psi^*(\mathrm{e}^{-\alpha_k \cdot} * \psi_j v), \qquad j = 1, \dots, k-1.\end{aligned}$$

Inserting this in (A.8) gives

$$\begin{aligned}\text{(A.10)}\quad \Lambda(\psi_k v) &= \gamma_k \Lambda(\psi_{k-1} v) - \gamma_k(\alpha_k + \overline{\alpha_{k-1}}) \\ &\cdot \left[\Psi^*(\mathrm{e}^{-\alpha_k \cdot} v) \cdot (\mathrm{e}^{-\alpha_k \cdot} * \psi_{k-1}) + \sum_{j=1}^{k-1} \Psi^*(\mathrm{e}^{-\alpha_k \cdot} * \psi_j v) \sum_{\ell=1}^{k-1} (L_{k-1})_{j\ell} \psi_\ell \right].\end{aligned}$$

By definition of $(x_j)_{j=1}^k$, $(z_j)_{j=1}^k$, and $\widetilde{L}_{k-1}$, this equals (A.6). □

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

$$\text{(A.11)}\qquad \Lambda(\psi_k e_q) = \sum_{i=1}^{n} \sum_{j=1}^{k} (R_k)_{ijq} \psi_j e_i, \qquad q = 1, \dots, p,$$

$$\text{(A.12)}\qquad \Psi^*(\psi_j e_q) = \sum_{i=1}^{n} (W_k)_{ijq} e_i, \quad j = 1, \dots, k, \quad q = 1, \dots, p.$$

Proposition A.4. *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 $j \in \mathbb{N}$, $(\psi_j)_{j=1}^\infty$ as in Definition* 3.3, *Ψ as in* (2.1), *and Λ as in* (3.7). *Then, for each $k \in \mathbb{N}$, there exists an $L_k \in \mathbb{C}^{k \times k}$ such that for all $v \in \mathbb{C}^p$,*

$$\text{(A.13)}\qquad \Lambda(\psi_k v) = \sum_{j=1}^{k} \Psi^*(\psi_j v) \sum_{\ell=1}^{k} (L_k)_{j\ell} \psi_\ell.$$

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

We note that in terms of the tensors defined through (A.11) and (A.12), the equality (A.13) can be written as

$$\text{(A.14)}\qquad (R_k)_{ijq} = \sum_{\ell=1}^{k} (W_k)_{i\ell q} (L_k)_{\ell j}.$$

Proof of Proposition A.4. We prove this by induction. For $k = 1$ we have by Proposition 3.6(c) that $\Lambda(\psi_1 v) = \psi_1 (\alpha_1 I - A^*)^{-1} C^* v$, and by Proposition 3.6(b) that $\Psi^*(\psi_1 v) = \sqrt{2\mathrm{Re}(\alpha_1)} (\alpha_1 I - A^*)^{-1} C^* v$. Hence for $k = 1$, (A.13) is satisfied with $L_1 = \frac{1}{\sqrt{2\mathrm{Re}(\alpha_1)}}$.

By the induction hypothesis, (A.13) holds with k replaced by $k-1$, i.e., (A.5) holds. From Lemma A.3 we then obtain that (A.6) holds. We now write all the terms in (A.6) (with $v := e_q$) with respect to the tensors defined through (A.11) and (A.12).

We consider the term $\Lambda(\psi_{k-1}e_q)$ in (A.6). We have by (A.11)

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

which by the induction hypothesis in tensor form (i.e., (A.14) with k replaced by $k-1$) can be written as

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

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

$$\Lambda(\psi_{k-1}e_q) = \sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{\ell=1}^{k}(W_k)_{i\ell q}(\widehat{L}_{k-1})_{\ell j}\psi_j e_i, \qquad q = 1, \dots, p. \tag{A.15}$$

We now consider the term $\Psi^*(x_j e_q)$ in (A.6). By Lemma A.2 we have

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

Using (A.12), this gives

$$\begin{aligned}\Psi^*(x_j e_q) &= \sum_{\ell=1}^{k}(M_k^T)_{j\ell}\Psi^*(\psi_\ell e_q) = \sum_{i=1}^{n}\sum_{\ell=1}^{k}(M_k^T)_{j\ell}(W_k)_{i\ell q}e_i \\ &= \sum_{i=1}^{n}\sum_{\ell=1}^{k}(W_k)_{i\ell q}(M_k)_{\ell j}e_i. \qquad \text{(A.16)}\end{aligned}$$

By Lemma A.2 we have

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

It follows that

$$\sum_{\ell=1}^{k}(\widetilde{L}_{k-1})_{j\ell}z_\ell = \sum_{\ell=1}^{k}\sum_{\beta=1}^{k}(\widetilde{L}_{k-1})_{j\ell}(T_k)_{\ell\beta}\psi_\beta = \sum_{\beta=1}^{k}(\widetilde{L}_{k-1}T_k)_{j\beta}\psi_\beta. \tag{A.17}$$

From (A.16) and (A.17) we obtain that

$$\begin{aligned}\sum_{j=1}^{k}\Psi^*(x_je_q)\sum_{\ell=1}^{k}(\widetilde{L}_{k-1})_{j\ell}z_\ell &= \sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\ell=1}^{k}(W_k)_{i\ell q}(M_k)_{\ell j}e_i\sum_{\beta=1}^{k}(\widetilde{L}_{k-1}T_k)_{j\beta}\psi_\beta\\ &= \sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{\ell=1}^{k}\sum_{\beta=1}^{k}(W_k)_{i\ell q}(M_k)_{\ell j}(\widetilde{L}_{k-1}T_k)_{j\beta}\psi_\beta e_i\\ &= \sum_{i=1}^{n}\sum_{\ell=1}^{k}\sum_{\beta=1}^{k}(W_k)_{i\ell q}(M_k\widetilde{L}_{k-1}T_k)_{\ell\beta}\psi_\beta e_i. \qquad \text{(A.18)}\end{aligned}$$

Substituting (A.11), (A.15), and (A.18) in (A.6) gives

$$\text{(A.19)}\quad \begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{k}(R_k)_{ijq}\psi_je_i &= \gamma_k\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{\ell=1}^{k}(W_k)_{i\ell q}(\widehat{L}_{k-1})_{\ell j}\psi_je_i\\ &\quad - \gamma_k(\alpha_k+\overline{\alpha_{k-1}})\sum_{i=1}^{n}\sum_{\ell=1}^{k}\sum_{j=1}^{k}(W_k)_{i\ell q}(M_k\widetilde{L}_{k-1}T_k)_{\ell j}\psi_je_i.\end{aligned}$$

We define L_k by (note that this is equivalent to how it is defined in Algorithm 1)

$$\text{(A.20)}\qquad L_k := \gamma_k(\widehat{L}_{k-1}) - \gamma_k(\alpha_k+\overline{\alpha_{k-1}})M_k\widetilde{L}_{k-1}T_k.$$

The right-hand side of (A.19) can then be written as

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{\ell=1}^{k}(W_k)_{i\ell q}(L_k)_{\ell j}\psi_je_i.$$

Hence,

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

Since (ψ_je_i) with $j=1,\dots,k$ and $i=1,\dots,n$ are linearly independent, it follows that $(R_k)_{ijq} = \sum_{\ell=1}^{k}(W_k)_{i\ell q}(L_k)_{\ell j}$, i.e., that (A.14) holds (or, equivalently, that (A.13) holds). □

Proof of Theorem 4.3. We prove this by induction. The relation $\iota_{1,m}^*\mathbb{F}\iota_{1,p} = F_1 = B^*(\alpha_1 I - A^*)^{-1}C^*$ follows from Proposition 3.6(c) together with $\mathbb{F}^* = B^*\Lambda$ and $\psi_1 = \sqrt{2\mathrm{Re}(\alpha_1)}\cdot\varphi_1$.

We first reformulate the tensors R_k and W_k defined through (A.11) and (A.12) and their relation (A.14) 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 (A.14) is equivalent to

$$\text{(A.21)}\qquad (L_k^T\otimes I_n)\widetilde{W_k} = \widetilde{R}_k.$$

Since $\mathbb{F}^* = B^*\Lambda$ we have by (A.11) that N_k from (A.1) 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 does not 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 (A.21) 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 Proposition 3.8(a) we have

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

where V_j $(j = 1, \ldots, k)$ is as in (4.6). When compared with (A.12) this shows that

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

i.e., $W_{\cdot j\cdot} = \sqrt{2\mathrm{Re}(\alpha_j)}V_j$. This shows that N_k from (A.1) is indeed as computed in Algorithm 1. It follows that F_k determined by Algorithm 1 fulfills (4.5). □

REFERENCES

[1] P. Benner and J. Saak, *A Galerkin-Newton-ADI Method for Solving Large-Scale Algebraic Riccati Equations*, Technical report SPP1253-090, Deutsche Forschungsgemeinschaft–Priority Program 1253, Bonn, Germany, 2010.

[2] A. E. Frazho and W. Bhosri, *An Operator Perspective on Signals and Systems*, Oper. Theory Adv. Appl. 204, Birkhäuser Verlag, Basel, 2010.

[3] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, translated from the Russian by A. Feinstein, Transl. Math. Monogr. 18, American Mathematical Society, Providence, RI, 1969.

[4] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, 1995.

[5] J.-R. Li and J. White, *Low rank solution of Lyapunov equations*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 260–280. http://epubs.siam.org/doi/abs/10.1137/S0895479801384937.

[6] Y. Lin and V. Simoncini, *A new subspace iteration method for the algebraic Riccati equation*, Numer. Linear Algebra Appl., 22 (2015), pp. 26–47. http://onlinelibrary.wiley.com/doi/10.1002/nla.1936/abstract.

[7] A. Lu and E. L. Wachspress, *Solution of Lyapunov equations by alternating direction implicit iteration*, Comput. Math. Appl., 21 (1991), pp. 43–58. http://dx.doi.org/10.1016/0898-1221(91)90124-M.

[8] M. R. Opmeer, T. Reis, and W. Wollner, *Finite-rank ADI iteration for operator Lyapunov equations*, SIAM J. Control Optim., 51 (2013), pp. 4084–4117. http://epubs.siam.org/doi/abs/10.1137/120885310.

[9] T. Penzl, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (2000), pp. 1401–1418. http://epubs.siam.org/doi/abs/10.1137/S1064827598347666.

[10] T. Penzl, *LYAPACK: A Matlab Toolbox for Large Lyapunov and Riccati Equations, Model Reduction Problems, and Linear-Quadratic Optimal Control Problems. Users' Guide (Version* 1.0*)*, http://www.netlib.org/lyapack/guide.pdf (2000).

[11] T. Reis and W. Wollner, *Finite-Rank ADI Iteration for Operator Lyapunov Equations*, Hamburger Beiträge zur Angewandten Mathematik, 2012-09.

[12] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Blackie & Son Limited, London, 1956.

[13] O. J. Staffans, *Well-Posed Linear Systems*, Encyclopedia Math. Appl. 103, Cambridge University Press, Cambridge, UK, 2005.

[14] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 10 (1997), pp. 287–330. http://link.springer.com/article/10.1007/BF01211550.

[15] J. Zabczyk, *Mathematical Control Theory: An Introduction*, Birkhäuser, Boston, MA, 2008.

[16] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*, Prentice Hall, Upper Saddle River, NJ, 1996.