The intrinsic Toeplitz structure and its applications in algebraic Riccati equations

Zhen-Chen Guo* Xin Liang†

Abstract

In this paper we derive a Toeplitz-structured closed form of the unique positive semi-definite stabilizing solution for the discrete-time algebraic Riccati equations, especially for the case that the state matrix is not stable. Based on the found form and fast Fourier transform, we propose a new algorithm for solving both discrete-time and continuous-time large-scale algebraic Riccati equations with low-rank structure. It works without unnecessary assumptions, shift selection strategies, or matrix calculations of the cubic order with respect to the problem scale. Numerical examples are given to illustrate its features. Besides, we show that it is theoretically equivalent to several algorithms existing in the literature in the sense that they all produce the same sequence under the same parameter setting.

Key words. Toeplitz matrix, FFT, algebraic Riccati equations, large-scale, low-rank

AMS subject classifications. 15A24, 15B05, 65F45, 93B52

1 Introduction

Consider a continuous-time algebraic/limiting Riccati equation (CARE)

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

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{l\times n}$. The CAREs arise in various models related to control theory, such as linear-quadratic optimal regulator design, and H^2 and H^∞ controller design for linear systems, see, e.g., [35, 3]. They also arise in nonlinear systems, like nonlinear controller design by state-dependent Riccati equations [13], or solving differential Riccati equations by implicit integration schemes [17, 8]. Usually (1.1) has infinite many solutions, but in many applications including those mentioned above only the so-called c-stabilizing solution is hoped to be chased. Here a solution X is called c-stabilizing if $A - BB^{\mathrm{T}}X$ is stable, namely all the eigenvalues of $A - BB^{\mathrm{T}}X$ lie in the open left half complex plane $\mathbb{C}_-$. Its existence and uniqueness are guaranteed by the assumption that the pairs (A, BB^{T}) and $(A^{\mathrm{T}}, C^{\mathrm{T}}C)$ are c-stabilizable, or equivalently, $\operatorname{rank}(\begin{bmatrix} A-\lambda I & BB^{\mathrm{T}}\end{bmatrix}) = \operatorname{rank}(\begin{bmatrix} A^{\mathrm{T}}-\lambda I & C^{\mathrm{T}}C\end{bmatrix}) = n$ for any $\lambda \in \mathbb{C}\setminus\mathbb{C}_-$.

During many years, people have developed many numerical methods to find out the c-stabilizing solution of (1.1). Reader are referred to [12] to obtain an overview. In this paper, we are focusing on a special case that A is large-scale and sparse, and B, C are low-rank, namely $m, l \ll n$. The existing methods are categorized into four classes:

1. projection methods, including extended Krylov subspace method [23], rational Krylov subspace method [18], tangential rational Krylov subspace method [19], global extended Krylov subspace method [27], etc.;
2. non-projective iterations, including quadratic ADI [44], Cayley transformed Hamiltonian subspace iteration [38], RADI [5], etc.;
3. Newton-type methods, including the Galerkin projected variant of Newton-Kleinman ADI [9] and its inexact line-search variant [6], etc.;
4. methods adopted from those suited for small-scale problems, including structure-preserving doubling algorithm (SDA) [14, 36], and Hamiltonian stable subspace methods [1, 4], etc..

*Department of Mathematics, Nanjing University, Nanjing 210093, China; e-mail: guozhenchen@nju.edu.cn. Supported in part by NSFC-11901290 and Fundamental Research Funds for the Central Universities.

†Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China, and Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, China; e-mail: liangxinslm@tsinghua.edu.cn. Supported in part by NSFC-11901340.

Many more methods and references can be listed if we bring in more details. Interested readers are encouraged to look through a comparison paper [11] and the references therein.

The methods in the former three classes use a lot of shifts in the calculation process, so a shift selection strategy rather than several pre-chosen shifts is needed. Different shifts or strategies usually affect the convergence speed significantly. Moreover, the convergence of those methods usually relies on more assumptions, for example, A is stable. To deal with the problems without the guarantee, the preprocessing is necessary and costs not little calculations. On the opposite, the methods in the latter class, like SDA, only use a few shifts, which helps decrease the calculation that is not directly related to the solution.

On the other hand, the methods in the former three classes only use matrix-vector multiplication and inverse-vector multiplication (that is actually done by linear system solvers), while SDA uses matrix-matrix and inverse-matrix multiplication (also done by linear system solvers), which implies that SDA consumes much more time than those in the former classes.

In this paper, first we contribute a Toeplitz-structured closed form of the d-stabilizing solution of discrete-time algebraic Riccati equations (DAREs) by theoretical analysis, which naturally induces a new algorithm named FFT-based Toeplitz-structured approximation (FTA) to solve DAREs. The proposed FTA method exploits the fast Fourier transform (FFT) to reduce the time complexity. Then using a Cayley transformation that transforms CAREs to DAREs, the FTA is successfully adopted to solve CAREs, where the incorporation technique (a.k.a. defect correction) is applied to deal with the case that the truncated approximation does not provide enough accuracy. The FTA solves DAREs and CAREs without more assumptions, shift selection strategies, or matrix-matrix/inverse-matrix multiplications. As a by-product, we show that FTA, SDA, and many other methods like RADI are equivalent under the same parameter setting including the same initial guess 0 and the same consistent shift, in the sense that they all produce the same sequence (or subsequence).

The rest of the paper is organized as follows. First, some notations are used. In Section 2 we present a detailed form of the inverse of special matrices of the form $I + TT^{\mathrm{T}}$ where T is block-Toeplitz, whose proof, not an easy consequence of the theory on Toeplitz matrices, is put in Appendix A for readability. Section 3 generalizes the idea on the Toeplitz operator in the associated discrete-time dynamic systems under good conditions to those without good conditions, and then naturally induces a closed form of the d-stabilizing solution of DAREs, where the special-structured matrices are involved, which suggests us to develop the FTA method to solve DAREs. As is shown in Section 4, an variant of FTA for CAREs is obtained with the help of Cayley transformation that transforms CAREs to DAREs. Numerical tests and discussions are given in Section 5. Some concluding remarks are provided in Section 6.

Notation. Throughout this paper, I_n (or simply I if its dimension is clear from the context) is the $n \times n$ identity matrix. Given a vector or matrix X, X^{T}, X^{H}, $\|X\|$, $\|X\|_{\mathrm{F}}$, $\rho(X)$ are its transpose, conjugate transpose, spectral norm, Frobenius norm, and spectral radius respectively. By $X \otimes Y$ denote the Kronecker product of X and Y. By $\Re\alpha$ denote the real part of a complex number α.

We use $X \succ 0$ ($X \succeq 0$) to indicate that X is symmetric positive (semi-)definite, and $X \prec 0$ ($X \preceq 0$) if $-X \succ 0$ ($-X \succeq 0$). Some easy identities are given:

$$U(I + V^{\mathrm{T}}U) = (I + UV^{\mathrm{T}})U, \qquad U(I + V^{\mathrm{T}}U)^{-1} = (I + UV^{\mathrm{T}})^{-1}U. \tag{1.2}$$

Here is the Sherman-Morrison-Woodbury formula:

$$(M + UDV^{\mathrm{T}})^{-1} = M^{-1} - M^{-1}U(D^{-1} + V^{\mathrm{T}}M^{-1}U)^{-1}V^{\mathrm{T}}M^{-1}. \tag{1.3}$$

The inverse sign in (1.2) and (1.3) indicates invertibility. Both will be applied occasionally.

In addition, all the discussions below are based on the field $\mathbb{R}$. They are also valid on the field $\mathbb{C}$, with all $(\cdot)^{\mathrm{T}}$ replaced by $(\cdot)^{\mathrm{H}}$.

2 Preliminary

The block-Toeplitz matrices appear in the subsequent sections and play an important role in the proposed algorithms. Since 1970s, people have known that fast and superfast algorithms are valid for Toeplitz matrices, due to its low displacement rank, see, e.g., [32, 30, 31, 33, 20]. However, to keep algebraic Riccati equations in mind, here we only introduce the notations related to block-Toeplitz matrices, and give a lemma that is used in the discussions on algebraic Riccati equations, while its proof is placed in Appendix A.

Given $A_0, A_1, \dots, A_{m-1} \in \mathbb{R}^{p_1 \times p_2}$, we will use

$$\mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{m-1} \end{bmatrix}\right) = \begin{bmatrix} A_0 & & & & & \\ A_1 & A_0 & & & & \\ A_2 & A_1 & \ddots & & & \\ \vdots & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & A_1 & A_0 & \\ A_{m-1} & \cdots & \cdots & A_2 & A_1 & A_0 \end{bmatrix} \in \mathbb{R}^{p_1 m \times p_2 m}.$$

For ease, $\mathcal{L}_{p_1\times p_2}(A) = \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{m-1} \end{bmatrix}\right)$ if $A = \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{m-1} \end{bmatrix}$, and this notation makes no confusion for the subscript $\cdot_{p_1\times p_2}$ demonstrates how the matrix is composed. Similarly,

$$\mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{m-1} \end{bmatrix}\right) = \begin{bmatrix} A_{m-1} & \cdots & \cdots & A_2 & A_1 & A_0 \\ & A_{m-1} & \ddots & & A_2 & A_1 \\ & & \ddots & \ddots & & A_2 \\ & & & \ddots & \ddots & \vdots \\ & & & & A_{m-1} & \vdots \\ & & & & & A_{m-1} \end{bmatrix} \in \mathbb{R}^{p_1 m \times p_2 m}.$$

Besides,

$$\mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} A_0 & A_1 & \cdots & A_{m-1} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} A_{m-1} \\ \vdots \\ A_1 \\ A_0 \end{bmatrix}\right),$$

$$\mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} A_0 & A_1 & \cdots & A_{m-1} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} A_{m-1} \\ \vdots \\ A_1 \\ A_0 \end{bmatrix}\right).$$

The following lemma will be used several times later.

Lemma 2.1. *Given $Y \in \mathbb{R}^{p_1\times p_2}$, $Y^L \in \mathbb{R}^{p_2\times p_1}$, $D_{t-1} \in \mathbb{R}^{p_1(t-1)\times p_2}$, $D_{t-1}^L \in \mathbb{R}^{p_2 \times p_1(t-1)}$, let*

$$T_t = \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y \\ D_{t-1} \end{bmatrix}\right) = \begin{bmatrix} Y & 0 \\ D_{t-1} & T_{t-1} \end{bmatrix} \in \mathbb{R}^{p_1 t\times p_2 t}, \quad T_t^L = \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y^L & D_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} = \begin{bmatrix} Y^L & D_{t-1}^L \\ 0 & T_{t-1}^L \end{bmatrix} \in \mathbb{R}^{p_2 t\times p_1 t}.$$

If $I - T_t T_t^L$ is nonsingular, then

$$(I - T_t T_t^L)^{-1} = \mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix}\right)(I \otimes Q_{2,b})^{-1}\mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} Q_2^L & Q_{2,b} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \tag{2.1a}$$
$$+ \mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I \otimes [W + WY^L YW])^{-1}\mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}},$$

where the following equations are solvable and $Q_{2,b}, W + WY^L YW$ are nonsingular:

$$Q_3^L\left[I - D_{t-1}D_{t-1}^L - T_{t-1}T_{t-1}^L\right] = D_{t-1}^L, \tag{2.1b}$$
$$\left[I - D_{t-1}D_{t-1}^L - T_{t-1}T_{t-1}^L\right] Q_3 = D_{t-1}, \qquad W = -(I + Q_3^L D_{t-1}), \tag{2.1c}$$
$$\begin{bmatrix} Q_2^L & Q_{2,b} \end{bmatrix}(I - T_t T_t^L) = \begin{bmatrix} 0 & I_{p_1} \end{bmatrix}, \tag{2.1d}$$
$$(I - T_t T_t^L)\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix} = \begin{bmatrix} 0 \\ I_{p_1} \end{bmatrix}, \qquad Q_{2,b} \in \mathbb{R}^{p_1\times p_1}. \tag{2.1e}$$

3 DARE

Given a linear time-invariant control system in discrete-time:

$$\begin{aligned} x_0 &\text{ is given,} \\ x_{k+1} &= Ax_k + Bu_k, \qquad k = 0, 1, 2, \dots, \\ y_k &= Cx_k, \end{aligned} \tag{3.1}$$

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{l\times n}$. Suppose the following condition holds through out this section:

$$\boxed{(A, B) \text{ is d-stabilizable and } (C, A) \text{ is d-detectable,}}$$

or equivalently, $\operatorname{rank}(\begin{bmatrix} A - \lambda I & BB^{\mathrm{T}} \end{bmatrix}) = \operatorname{rank}(\begin{bmatrix} A^{\mathrm{T}} - \lambda I & C^{\mathrm{T}}C \end{bmatrix}) = n$ for any $\lambda \in \mathbb{C} \setminus \mathbb{D}$, where $\mathbb{D}$ is the open unit disk.

Its linear-quadratic optimal control can be expressed as

$$\arg\min_{\{u_k\}} \frac{1}{2} \sum_{k=0}^{\infty} (y_k^{\mathrm{T}} y_k + u_k^{\mathrm{T}} u_k) = \{u_k = -(I + B^{\mathrm{T}} X_\star B)^{-1} B^{\mathrm{T}} X_\star A x_k\}, \tag{3.2}$$

where $X_\star$ is the unique symmetric positive semi-definite d-stabilizing solution of the DARE [12, 15, 34, 39]:

$$-X + A^{\mathrm{T}} X (I + BB^{\mathrm{T}} X)^{-1} A + C^{\mathrm{T}} C = 0. \tag{3.3}$$

Here a solution X is called d-stabilizing, if the closed loop matrix $A_X = (I + BB^{\mathrm{T}}X)^{-1}A$ is d-stable, namely all of its eigenvalues lie in the open unit disk $\mathbb{D}$, or equivalently, $\rho(A_X) < 1$.

3.1 In the operator view

In this section, we briefly state the existence and uniqueness of $X_\star$ shown by the operator theory, which is based on the monograph [26].

In order to make things simple, first we assume that A is d-stable. Write $\boldsymbol{x} = \{x_k\}_{k\in\mathbb{N}}, \boldsymbol{u} = \{u_k\}_{k\in\mathbb{N}}, \boldsymbol{y} = \{y_k\}_{k\in\mathbb{N}}$. Let $\ell_+^{2,n}$ denote the Hilbert space of norm-square summable $\mathbb{R}^n$-valued series.

Suppose $\boldsymbol{u} \in \ell_+^{2,m}$ and consider the cost functional (a.k.a. restricted quadratic index)

$$J(\boldsymbol{u}) = \sum_{k=0}^{+\infty} \begin{bmatrix} x_k \\ u_k \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} Q & L \\ L^{\mathrm{T}} & R \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} = \left\langle \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} Q & L \\ L^{\mathrm{T}} & R \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{u} \end{bmatrix} \right\rangle_{\ell_+^{2,n} \times \ell_+^{2,m}},$$

where $\boldsymbol{x}$ is the solution to (3.1). Here for any matrix U and any series $\boldsymbol{z} = \{z_k\}_{k\in\mathbb{N}}$, $U\boldsymbol{z}$ is understood as $U\boldsymbol{z} := \{Uz_k\}_{k\in\mathbb{N}}$.

In fact $\boldsymbol{x} = \mathscr{F}x_0 + \mathscr{L}\boldsymbol{u} \in \ell_+^{2,n}$, where $\mathscr{F} \colon \mathbb{R}^n \to \ell_+^{2,n}, (\mathscr{F}x_0)_k = A^k x_0, k \ge 0$, and $\mathscr{L} \colon \ell_+^{2,m} \to \ell_+^{2,n}, (\mathscr{L}\boldsymbol{u})_0 = 0, (\mathscr{L}\boldsymbol{u})_k = \sum_{i=0}^{k-1} A^{k-i-1} B u_i, k \ge 1$. Clearly $\mathscr{F}$ and $\mathscr{L}$ are bounded linear operators. Also, it is not difficult to find $\mathscr{L}$ is a Toeplitz operator. Hence

$$\begin{aligned} J(\boldsymbol{u}) &= \left\langle \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} Q & L \\ L^{\mathrm{T}} & R \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{u} \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} \mathscr{F} & \mathscr{L} \\ & I \end{bmatrix} \begin{bmatrix} x_0 \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} Q & L \\ L^{\mathrm{T}} & R \end{bmatrix} \begin{bmatrix} \mathscr{F} & \mathscr{L} \\ & I \end{bmatrix} \begin{bmatrix} x_0 \\ \boldsymbol{u} \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} x_0 \\ \boldsymbol{u} \end{bmatrix}, \begin{bmatrix} \mathscr{P}_o & \mathscr{P} \\ \mathscr{P}^* & \mathscr{R} \end{bmatrix} \begin{bmatrix} x_0 \\ \boldsymbol{u} \end{bmatrix} \right\rangle \qquad (\mathscr{A}^* \text{ is the adjoint of operator } \mathscr{A}) \end{aligned}$$

where $\mathscr{P}_o = \mathscr{F}^* Q \mathscr{F}, \mathscr{P} = \mathscr{F}^*(Q\mathscr{L} + L), \mathscr{R} = R + L^{\mathrm{T}}\mathscr{L} + \mathscr{L}^* L + \mathscr{L}^* Q \mathscr{L}$. Then the unique symmetric d-stabilizing solution $X_\star$ is given by

$$X_\star = \mathscr{P}_o - \mathscr{P}\mathscr{R}^{-1}\mathscr{P}^*. \tag{3.4}$$

Clearly, $\mathscr{P}_o, \mathscr{P}, \mathscr{R}$ are bounded linear operators. [26, Theorem 4.7.1] tells that the DARE (3.3) has a unique d-stabilizing solution, if and only if the Toeplitz-like operator $\mathscr{R}$ has a bounded inverse.

Then for the case that A is not d-stable, a similar result still holds. Since (A, B) is d-stabilizable, there exists $F \in \mathbb{R}^{m\times n}$ such that $\widetilde{A} = A + BF$ is d-stable. Note the system defined by $A, B, Q = C^{\mathrm{T}}C, L = 0, R = I$ is equivalent to the system defined by $\widetilde{A}, \widetilde{B} = B, \widetilde{Q} = C^{\mathrm{T}}C + F^{\mathrm{T}}F, \widetilde{L} = F^{\mathrm{T}}, \widetilde{R} = I$ in the sense that they share the same d-stabilizing solution. Hence

$$X_\star = \widetilde{\mathscr{P}}_o - \widetilde{\mathscr{P}}\widetilde{\mathscr{R}}^{-1}\widetilde{\mathscr{P}}^*. \tag{3.5}$$

In the following, we derive the (infinite) matrix representation of (3.4) and (3.5), which is not provided in [26]. First consider the case that A is d-stable. Obviously, the matrix representations of $\mathscr{F},\mathscr{L}$, still denoted by $\mathscr{F},\mathscr{L}$, are

$$\mathscr{F}=\begin{bmatrix} I\\ A\\ A^2\\ A^3\\ \vdots\end{bmatrix},\qquad \mathscr{L}=\begin{bmatrix} 0 & & & & \\ B & 0 & & & \\ AB & B & 0 & & \\ A^2B & AB & B & 0 & \\ \vdots & \ddots & \ddots & \ddots & \ddots\end{bmatrix}.$$

Hence

$$\begin{aligned}X_\star &= \mathscr{F}^*Q\mathscr{F}-\mathscr{F}^*(Q\mathscr{L}+L)(R+L^{\mathrm T}\mathscr{L}+\mathscr{L}^*L+\mathscr{L}^*Q\mathscr{L})^{-1}(\mathscr{L}^*Q^{\mathrm T}+L^{\mathrm T})\mathscr{F}\\ &= \mathscr{F}^*C^{\mathrm T}C\mathscr{F}-\mathscr{F}^*C^{\mathrm T}C\mathscr{L}(I+\mathscr{L}^*C^{\mathrm T}C\mathscr{L})^{-1}\mathscr{L}^*C^{\mathrm T}C\mathscr{F}\\ &= \mathscr{F}^*C^{\mathrm T}\left[I-C\mathscr{L}(I+\mathscr{L}^*C^{\mathrm T}C\mathscr{L})^{-1}\mathscr{L}^*C^{\mathrm T}\right]C\mathscr{F}\\ &\overset{(1.3)}{=} \mathscr{F}^*C^{\mathrm T}(I+C\mathscr{L}\mathscr{L}^*C^{\mathrm T})^{-1}C\mathscr{F}.\end{aligned}$$

Write

$$\mathscr{V}=C\mathscr{F}=\begin{bmatrix} C\\ CA\\ CA^2\\ CA^3\\ \vdots\end{bmatrix},\quad \mathscr{T}=C\mathscr{L}=\begin{bmatrix} 0 & & & & \\ CB & 0 & & & \\ CAB & CB & 0 & & \\ CA^2B & CAB & CB & 0 & \\ \vdots & \ddots & \ddots & \ddots & \ddots\end{bmatrix}, \tag{3.6}$$

and then $\mathscr{V}\colon\mathbb{R}^n\to\ell_+^{2,l},\mathscr{T}\colon\ell_+^{2,m}\to\ell_+^{2,l},(I+\mathscr{T}\mathscr{T}^*)^{-1}\colon\ell_+^{2,l}\to\ell_+^{2,l}$ are bounded linear operators, and $\mathscr{T}$ is also a Toeplitz operator. Thus

$$X_\star=\mathscr{V}^*(I+\mathscr{T}\mathscr{T}^*)^{-1}\mathscr{V}, \tag{3.7}$$

which is a closed form of the d-stabilizing solution.

Then consider the case A is not d-stable. It follows from (3.5) that

$$\begin{aligned}X_\star &= \widetilde{\mathscr{F}}^*\widetilde{Q}\widetilde{\mathscr{F}}-\widetilde{\mathscr{F}}^*(\widetilde{Q}\widetilde{\mathscr{L}}+\widetilde{L})(\widetilde{R}+\widetilde{L}^{\mathrm T}\widetilde{\mathscr{L}}+\widetilde{\mathscr{L}}^*\widetilde{L}+\widetilde{\mathscr{L}}^*\widetilde{Q}\widetilde{\mathscr{L}})^{-1}(\widetilde{\mathscr{L}}^*\widetilde{Q}^{\mathrm T}+\widetilde{L}^{\mathrm T})\widetilde{\mathscr{F}}\\ &= \widetilde{\mathscr{F}}^*[C^{\mathrm T}C+F^{\mathrm T}F]\widetilde{\mathscr{F}}\\ &\quad -\widetilde{\mathscr{F}}^*([C^{\mathrm T}C+F^{\mathrm T}F]\widetilde{\mathscr{L}}+F^{\mathrm T})(I+F\widetilde{\mathscr{L}}+\widetilde{\mathscr{L}}^*F^{\mathrm T}+\widetilde{\mathscr{L}}^*[C^{\mathrm T}C+F^{\mathrm T}F]\widetilde{\mathscr{L}})^{-1}(\widetilde{\mathscr{L}}^*[C^{\mathrm T}C+F^{\mathrm T}F]+F)\widetilde{\mathscr{F}}\\ &= \widetilde{\mathscr{F}}^*\begin{bmatrix}F\\ C\end{bmatrix}^{\mathrm T}\begin{bmatrix}F\\ C\end{bmatrix}\widetilde{\mathscr{F}}-\widetilde{\mathscr{F}}^*\begin{bmatrix}F\\ C\end{bmatrix}^{\mathrm T}\begin{bmatrix}I+F\widetilde{\mathscr{L}}\\ C\widetilde{\mathscr{L}}\end{bmatrix}\left(\begin{bmatrix}I+F\widetilde{\mathscr{L}}\\ C\widetilde{\mathscr{L}}\end{bmatrix}^*\begin{bmatrix}I+F\widetilde{\mathscr{L}}\\ C\widetilde{\mathscr{L}}\end{bmatrix}\right)^{-1}\begin{bmatrix}I+F\widetilde{\mathscr{L}}\\ C\widetilde{\mathscr{L}}\end{bmatrix}^*\begin{bmatrix}F\\ C\end{bmatrix}\widetilde{\mathscr{F}}\\ &\overset{(1.3)}{=} \widetilde{\mathscr{F}}^*\begin{bmatrix}F\\ C\end{bmatrix}^{\mathrm T}\left(I-\mathscr{Q}(\mathscr{Q}^*\mathscr{Q})^{-1}\mathscr{Q}^*\right)\begin{bmatrix}F\\ C\end{bmatrix}\widetilde{\mathscr{F}}\qquad\left(\text{write}\quad \mathscr{Q}=\begin{bmatrix}I+F\widetilde{\mathscr{L}}\\ C\widetilde{\mathscr{L}}\end{bmatrix}\right).\end{aligned}$$

Clearly $I-\mathscr{Q}(\mathscr{Q}^*\mathscr{Q})^{-1}\mathscr{Q}^*$ is an orthogonal projection operator onto $\mathcal{R}(\mathscr{Q})^\perp$, the orthogonal complement of the image set of $\mathscr{Q}$. We temporarily forget the convergence of series and the boundedness of operators. Then $C\widetilde{\mathscr{L}}=C\mathscr{L}(I+F\widetilde{\mathscr{L}})$, because the three linear operators $C\widetilde{\mathscr{L}},C\mathscr{L},I+F\widetilde{\mathscr{L}}$ are all Toeplitz operators, and the generating polynomials satisfy

$$\begin{aligned}&(CB\lambda+CAB\lambda^2+CA^2B\lambda^3+\cdots)(I+FB\lambda+F(A+BF)B\lambda^2+F(A+BF)^2B\lambda^3+\cdots)\\ &\quad= CB\lambda+(CBFB+CAB)\lambda^2+(CBF(A+BF)B+CABFB+CA^2B)\lambda^3+\cdots\\ &\quad= CB\lambda+C(A+BF)B\lambda^2+C(A+BF)^2B\lambda^3+\cdots.\end{aligned}$$

Hence $\mathscr{Q}=\begin{bmatrix}I+F\widetilde{\mathscr{L}}\\ C\mathscr{L}(I+F\widetilde{\mathscr{L}})\end{bmatrix}=\begin{bmatrix}I\\ C\mathscr{L}\end{bmatrix}(I+F\widetilde{\mathscr{L}})$ where $I+F\widetilde{\mathscr{L}}$ is invertible. Since

$$\begin{bmatrix}I & -\mathscr{L}^*C^{\mathrm T}\\ C\mathscr{L} & I\end{bmatrix}=\begin{bmatrix}I & \\ C\mathscr{L} & I\end{bmatrix}\begin{bmatrix}I & \\ & I+C\mathscr{L}\mathscr{L}^*C^{\mathrm T}\end{bmatrix}\begin{bmatrix}I & -\mathscr{L}^*C^{\mathrm T}\\ & I\end{bmatrix}$$

is invertible, and $\begin{bmatrix}I\\ C\mathscr{L}\end{bmatrix}^*\begin{bmatrix}-\mathscr{L}^*C^{\mathrm T}\\ I\end{bmatrix}=0$, we know that $\mathscr{Q}_\perp:=\begin{bmatrix}-\mathscr{L}^*C^{\mathrm T}\\ I\end{bmatrix}$ satisfies $\mathscr{Q}_\perp(\mathscr{Q}_\perp^*\mathscr{Q}_\perp)^{-1}\mathscr{Q}_\perp^*=I-\mathscr{Q}(\mathscr{Q}^*\mathscr{Q})^{-1}\mathscr{Q}^*$. Similarly we have $C\widetilde{\mathscr{F}}=C\mathscr{F}+C\mathscr{L}F\widetilde{\mathscr{F}}$. Thus

$$X_\star=\begin{bmatrix}F\widetilde{\mathscr{F}}\\ C\mathscr{F}+C\mathscr{L}F\widetilde{\mathscr{F}}\end{bmatrix}^*\begin{bmatrix}-\mathscr{L}^*C^{\mathrm T}\\ I\end{bmatrix}(I+C\mathscr{L}\mathscr{L}^*C^{\mathrm T})^{-1}\begin{bmatrix}-\mathscr{L}^*C^{\mathrm T}\\ I\end{bmatrix}^*\begin{bmatrix}F\widetilde{\mathscr{F}}\\ C\mathscr{F}+C\mathscr{L}F\widetilde{\mathscr{F}}\end{bmatrix}$$

$$= \mathscr{F}^* C^{\mathrm{T}}(I + C\mathscr{L}\mathscr{L}^* C^{\mathrm{T}})^{-1} C\mathscr{F}$$
$$= \mathscr{V}^*(I + \mathscr{T}\mathscr{T}^*)^{-1}\mathscr{V},$$

where $\mathscr{V}, \mathscr{T}$ are as in (3.6), which implies that the solution in the unstable case would have the same form as that in the stable case. However, is this true, considering the convergence of series? In the next subsection, we will show the validity of (3.7).

3.2 In the matrix view

It is well known that $X_\star = \lim_{t\to\infty} X_t$, where X_t is generated by the difference Riccati equation (DRE):

$$X_0 = 0, \qquad X_{t+1} = \mathscr{D}(X_t) := C^{\mathrm{T}}C + A^{\mathrm{T}}X_t(I + BB^{\mathrm{T}}X_t)^{-1}A, \tag{3.8}$$

which can be recognized as a variant of fixed point iteration for (3.3).

Based on the fixed point iteration, in 1970s, people have developed the doubling algorithm to solve DAREs (3.3) and CAREs. Anderson [2] proposed a variant, which is recently usually called SDA and has three iterative recursions:

$$A_{k+1} = A_k(I_n + G_kH_k)^{-1}A_k, \tag{3.9a}$$
$$G_{k+1} = G_k + A_k(I_n + G_kH_k)^{-1}G_kA_k^{\mathrm{T}}, \tag{3.9b}$$
$$H_{k+1} = H_k + A_k^{\mathrm{T}}H_k(I_n + G_kH_k)^{-1}A_k, \tag{3.9c}$$

provided that all matrix inversions are feasible (i.e., $I_n + G_kH_k$ are nonsingular for $k = 0, 1, \cdots$). The initial terms are usually set by

$$A_0 = A, \qquad G_0 = BB^{\mathrm{T}}, \qquad H_0 = C^{\mathrm{T}}C.$$

It has been shown that for those initial terms, $I_n + G_kH_k$ are nonsingular for $k \ge 0$, and $A_k \to 0$, $G_k \to Y_\star$ (the solution to the dual DARE) and $H_k \to X_\star$, all quadratically [39] except for the critical case [25].

In [2], it is stated clearly that $H_k = X_{2^k}$, implying that the iteration for H_k can be treated as an acceleration of (3.8), because it only computes the terms $X_1, X_2, X_4, \dots, X_{2^k}, \dots$. Moreover, [2] also argued that (3.8) with any initial $X_0 \succeq 0$ leads $X_t \to X_\star$ in usual situation (but did not mention which situation satisfies).

Questions arise naturally, of which two are:

1. can we even only compute less terms in the sequence $\{X_t\}$, namely accelerate (3.8) even further?
2. how things go when arbitrary initial terms are set?

Before we begin the analysis, a simple property of the DRE (3.8) is given.

Lemma 3.1. *The operator $\mathscr{D}$ is monotonic on the set consisting of all positive semi-definite matrices with respect to the partial order "$\succeq$". In details, if $Z_1 \succeq 0, Z_2 \succeq 0$, then*

$$Z_1 \succeq Z_2 \Rightarrow \mathscr{D}(Z_1) \succeq \mathscr{D}(Z_2).$$

Proof. First suppose $Z_2 \succ 0$ and thus Z_2 is nonsingular. Then

$$Z_1 \succeq Z_2 \Leftrightarrow Z_1^{-1} \preceq Z_2^{-1} \Leftrightarrow (Z_1^{-1} + BB^{\mathrm{T}})^{-1} \succeq (Z_2^{-1} + BB^{\mathrm{T}})^{-1}$$
$$\Leftrightarrow Z_1(I + BB^{\mathrm{T}}Z_1)^{-1} \succeq Z_2(I + BB^{\mathrm{T}}Z_2)^{-1} \Rightarrow \mathscr{D}(Z_1) \succeq \mathscr{D}(Z_2).$$

If Z_2 is singular, then $Z_2 + \varepsilon I \succ 0$ for any $\varepsilon > 0$. Thus, taking limits yields

$$Z_1 \succeq Z_2 \Leftrightarrow Z_1 + \varepsilon I \succeq Z_2 + \varepsilon I \Rightarrow \mathscr{D}(Z_1 + \varepsilon I) \succeq \mathscr{D}(Z_2 + \varepsilon I) \Rightarrow \mathscr{D}(Z_1) \succeq \mathscr{D}(Z_2).$$ □

Let us observe the first several iterations of the DRE (3.8). Clearly $X_0 = 0, X_1 = C^{\mathrm{T}}C$. Since $X_1 \succeq X_0 = 0$, by Lemma 3.1, $X_2 = \mathscr{D}(X_1) \succeq \mathscr{D}(X_0) = X_1$. Similarly $0 = X_0 \preceq X_1 \preceq X_2 \preceq \cdots \preceq X_t \preceq \cdots$, namely the sequence $\{X_t\}$ generated by (3.8) is monotonic. On the other hand, the d-stabilizing solution $X_\star \succeq 0 = X_0$, and thus $X_\star = \mathscr{D}(X_\star) \succeq \mathscr{D}(X_0) = X_1$. Similarly $X_\star \succeq X_t$ for any t, namely the sequence $\{X_t\}$ is bounded. As a result, it holds that $X_t \to X_\star$.

More structure of X_t can be obtained. Using some calculations, we have

$$X_2 = C^{\mathrm{T}}C + A^{\mathrm{T}}C^{\mathrm{T}}C(I + BB^{\mathrm{T}}C^{\mathrm{T}}C)^{-1}A$$
$$= C^{\mathrm{T}}C + A^{\mathrm{T}}C^{\mathrm{T}}(I + CBB^{\mathrm{T}}C^{\mathrm{T}})^{-1}CA$$

$$= \begin{bmatrix} C \\ CA \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I & \\ & I + CBB^{\mathrm{T}}C^{\mathrm{T}} \end{bmatrix}^{-1} \begin{bmatrix} C \\ CA \end{bmatrix}$$
$$= \begin{bmatrix} C \\ CA \end{bmatrix}^{\mathrm{T}} \left(I + \begin{bmatrix} 0 & \\ CB & 0 \end{bmatrix} \begin{bmatrix} 0 & \\ CB & 0 \end{bmatrix}^{\mathrm{T}} \right)^{-1} \begin{bmatrix} C \\ CA \end{bmatrix}.$$

This encourages us to use an ansatz $X_t = V_t^{\mathrm{T}}(I + T_tT_t^{\mathrm{T}})^{-1}V_t$. Note that this ansatz is always valid because $X_t \succeq 0$. The following theorem, namely Theorem 3.1, gives one of the detailed forms of the ansatz.

Theorem 3.1. *Write*

$$V_t = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ \vdots \\ CA^{t-1} \end{bmatrix}, \quad T_t = \begin{bmatrix} 0 & & & & & \\ CB & 0 & & & & \\ CAB & CB & \ddots & & & \\ \vdots & & \ddots & \ddots & & \\ \vdots & & & CB & 0 & \\ CA^{t-2}B & \cdots & \cdots & CAB & CB & 0 \end{bmatrix}, \tag{3.10}$$

$T_1 = 0$. *Then the terms of the sequence* $\{X_t\}$ *generated by the DRE* (3.8) *are*

$$X_t = V_t^{\mathrm{T}}(I + T_tT_t^{\mathrm{T}})^{-1}V_t, \qquad t = 1, 2, \ldots. \tag{3.11}$$

Moreover, $\{X_t\}$ *is monotonically nondecreasing, and* $X_t \to X_\star$*, the solution of DARE* (3.3).

Proof. The "moreover" part has been illustrated above. Only (3.11) is proved here.

We have already shown (3.11) is correct for $t = 1, 2$. Assuming (3.11) is correct for t, we are going to prove it is also correct for $t + 1$. By the DRE (3.8),

$$\begin{aligned} X_{t+1} &= C^{\mathrm{T}}C + A^{\mathrm{T}}V_t^{\mathrm{T}}\left(I + T_tT_t^{\mathrm{T}}\right)^{-1}V_t\left(I + BB^{\mathrm{T}}V_t^{\mathrm{T}}\left(I + T_tT_t^{\mathrm{T}}\right)^{-1}V_t\right)^{-1}A \\ &\overset{(1.2)}{=} C^{\mathrm{T}}C + A^{\mathrm{T}}V_t^{\mathrm{T}}\left(I + T_tT_t^{\mathrm{T}}\right)^{-1}\left(I + V_tBB^{\mathrm{T}}V_t^{\mathrm{T}}\left(I + T_tT_t^{\mathrm{T}}\right)^{-1}\right)^{-1}V_tA \\ &= C^{\mathrm{T}}C + A^{\mathrm{T}}V_t^{\mathrm{T}}\left(I + T_tT_t^{\mathrm{T}} + V_tBB^{\mathrm{T}}V_t^{\mathrm{T}}\right)^{-1}V_tA \\ &= \begin{bmatrix} C \\ V_tA \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I & \\ & I + T_tT_t^{\mathrm{T}} + V_tBB^{\mathrm{T}}V_t^{\mathrm{T}} \end{bmatrix}^{-1} \begin{bmatrix} C \\ V_tA \end{bmatrix} \\ &= \begin{bmatrix} C \\ V_tA \end{bmatrix}^{\mathrm{T}} \left(I + \begin{bmatrix} 0 & \\ V_tB & T_t \end{bmatrix} \begin{bmatrix} 0 & \\ V_tB & T_t \end{bmatrix}^{\mathrm{T}} \right)^{-1} \begin{bmatrix} C \\ V_tA \end{bmatrix} \\ &= V_{t+1}^{\mathrm{T}}(I + T_{t+1}T_{t+1}^{\mathrm{T}})^{-1}V_{t+1}. \end{aligned}$$

□

It is not difficult to discover that Theorem 3.1 coincides with the decoupled formulae of the dSDA for DAREs introduced in [21] at $t = 2^k$, which is actually guaranteed by the fact that the sequence $\{H_k\}$ generated by SDA (3.9c) is a subsequence of $\{X_t\}$.

One can easily find (3.11) is the truncated form of (3.7), a Toeplitz-structured closed form of $X_\star$, whose validity for the d-stable case has been proved by the operator theory in Section 3.1. Note that under the assumption that A is d-stable, V_t and T_t, treated as the truncations of $\mathscr{V}$ and $\mathscr{T}$, converges to $\mathscr{V}$ and $\mathscr{T}$ respectively, by the fact that $\mathscr{V}, \mathscr{T}$ are bounded linear operators. With the help of operator theory, $X_t \to X_\star$. To the opposite, for the case that A is not d-stable, $\mathscr{V}$ and $\mathscr{T}$ are no longer bounded, and it would be difficult to show $X_t \to X_\star$ by the operator theory. However, the matrix analysis reveals that $X_t \to X_\star$, which implies $X_\star$ indeed has the closed form (3.7) in the unstable case.

3.3 Efficient method

Now we acquire the non-iterative form (3.11) of X_t, which allows us to compute the terms X_t directly for arbitrary t. In the following, we will work on an efficient method to compute X_t for any given t.

Using the notations for Toeplitz matrices in Section 2, we have

$$T_t = \mathcal{L}_{l\times m}\left(\begin{bmatrix} 0 \\ V_{t-1}B \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 \\ \mathcal{L}_{l\times m}(V_{t-1}B) & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ V_{t-1}B & T_{t-1} \end{bmatrix}.$$

Thus,

$$\mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}} = T_{t-1}T_{t-1}^{\mathrm{T}} + V_{t-1}BB^{\mathrm{T}}V_{t-1}^{\mathrm{T}},$$

and

$$X_t = C^{\mathrm{T}}C + A^{\mathrm{T}}V_{t-1}^{\mathrm{T}}\left[I + \mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}}\right]^{-1}V_{t-1}A. \tag{3.12}$$

Clearly $\mathcal{L}_{l\times m}(V_{t-1}B)$ is block-Toeplitz. Hence the results in Section 2 can be applied.

By Lemma 2.1, X_t can be computed by solving only $l+m$ rather than n equations.

Theorem 3.2. *Let*

$$\left[I + \mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}}\right]\begin{bmatrix}Q_{2,c}\\ Q_{2,b}\end{bmatrix} = \begin{bmatrix}0\\ I_l\end{bmatrix}, \qquad Q_{2,b}\in\mathbb{R}^{l\times l}, \tag{3.13a}$$

$$\left[I + \mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}}\right]\begin{bmatrix}Q_{3,t}\\ Q_{3,c}\end{bmatrix} = V_{t-1}B, \qquad Q_{3,t}\in\mathbb{R}^{l\times m}, \tag{3.13b}$$

and $W = I - \begin{bmatrix}Q_{3,t}\\ Q_{3,c}\end{bmatrix}^{\mathrm{T}} V_{t-1}B$. *Then the sequence X_t defined by (3.11) can be generated by*

$$X_t = \begin{bmatrix}C\\ \Xi_1\\ \Xi_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}I & & \\ & (I\otimes Q_{2,b})^{-1} & \\ & & (I\otimes W)^{-1}\end{bmatrix}\begin{bmatrix}C\\ \Xi_1\\ \Xi_2\end{bmatrix}, \tag{3.14}$$

where

$$\Xi_1 = \mathcal{U}_{l\times l}\left(\begin{bmatrix}Q_{2,c}\\ Q_{2,b}\end{bmatrix}\right)^{\mathrm{T}} V_{t-1}A, \qquad \Xi_2 = \mathcal{U}_{l\times m}\left(\begin{bmatrix}Q_{3,c}\\ 0\end{bmatrix}\right)^{\mathrm{T}} V_{t-1}A.$$

Proof. By Lemma 2.1 with $Y = -(Y^L)^{\mathrm{T}} \leftarrow 0, D_{t-1} = -(D_{t-1}^L)^{\mathrm{T}} \leftarrow V_{t-1}B$, we have

$$(I+T_tT_t^{\mathrm{T}})^{-1} = \mathcal{U}_{l\times l}\left(\begin{bmatrix}Q_2\\ Q_{2,b}\end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{l\times l}\left(\begin{bmatrix}Q_2\\ Q_{2,b}\end{bmatrix}\right)^{\mathrm{T}} + \mathcal{U}_{l\times m}\left(\begin{bmatrix}Q_3\\ 0\end{bmatrix}\right)(I\otimes W)^{-1}\mathcal{U}_{l\times m}\left(\begin{bmatrix}Q_3\\ 0\end{bmatrix}\right)^{\mathrm{T}},$$

where

$$\left[I + \mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}}\right]Q_3 = V_{t-1}B, \quad W = I - Q_3^{\mathrm{T}}V_{t-1}B,$$
$$(I+T_tT_t^{\mathrm{T}})\begin{bmatrix}Q_2\\ Q_{2,b}\end{bmatrix} = \begin{bmatrix}0\\ I_l\end{bmatrix}, \qquad Q_{2,b}\in\mathbb{R}^{l\times l}.$$

Note that

$$(I+T_tT_t^{\mathrm{T}})^{-1} = \begin{bmatrix}I & \\ & \left[I + \mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}}\right]^{-1}\end{bmatrix}.$$

Letting $Q_2 = \begin{bmatrix}Q_{2,t}\\ Q_{2,c}\end{bmatrix}$ where $Q_{2,t}\in\mathbb{R}^{l\times l}$, we have $Q_{2,t}=0$ and

$$\left[I + \mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}}\right]\begin{bmatrix}Q_{2,c}\\ Q_{2,b}\end{bmatrix} = \begin{bmatrix}0\\ I_l\end{bmatrix}.$$

Write $Q_3 = \begin{bmatrix}Q_{3,t}\\ Q_{3,c}\end{bmatrix}$ where $Q_{3,t}\in\mathbb{R}^{l\times m}$, and then it follows that

$$\begin{aligned}&\left[I + \mathcal{L}_{l\times m}(V_{t-1}B)\,\mathcal{L}_{l\times m}(V_{t-1}B)^{\mathrm{T}}\right]^{-1}\\ &\quad= \mathcal{U}_{l\times l}\left(\begin{bmatrix}Q_{2,c}\\ Q_{2,b}\end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{l\times l}\left(\begin{bmatrix}Q_{2,c}\\ Q_{2,b}\end{bmatrix}\right)^{\mathrm{T}} + \mathcal{U}_{l\times m}\left(\begin{bmatrix}Q_{3,c}\\ 0\end{bmatrix}\right)(I\otimes W)^{-1}\mathcal{U}_{l\times m}\left(\begin{bmatrix}Q_{3,c}\\ 0\end{bmatrix}\right)^{\mathrm{T}}.\end{aligned} \tag{3.15}$$

Then the result is a direct consequence of (3.12). □

Theorem 3.2 suggests a new algorithm, Algorithm 1, to approximate the solution of DAREs. The key is how to fast compute $Q_{*,*}$, or equivalently solve the linear systems (3.13), and compute the products of $\mathcal{U}_*(*)^{\mathrm{T}}V_{t-1}A$. Both are related to the manipulations on block Toeplitz matrices. It is well known that the fast Fourier transform (FFT) can be used to accelerate the calculation with Toeplitz matrices involved, see, e.g., [43, 28, 29] and the references therein.

Some remarks are given below to illustrate the algorithm.

Algorithm 1 FFT-based Toeplitz-structured Approximation (FTA) for DAREs

Input: A, B, C and t.

1: Compute sequentially $C \cdot A, CA \cdot A, \dots, CA^{t-3} \cdot A, CA^{t-2} \cdot A$, and form V_{t-1} by stacking C and the first $t-2$ terms vertically in order, and form $V_{t-1}A$ by stacking the $t-1$ terms vertically in order.

2: Compute $V_{t-1}B$.

3: Use Preconditioned Conjugate Gradient method to solve (3.13).

4: Compute $W = I - \begin{bmatrix} Q_{3,t} \\ Q_{3,c} \end{bmatrix}^{\mathrm{T}} V_{t-1}B$ and then the Cholesky factorizations of $Q_{2,b} = L_Q L_Q^{\mathrm{T}}$ and $W = L_W L_W^{\mathrm{T}}$.

5: Use fast multiplication to obtain $S_1 = \mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix} L_Q^{-\mathrm{T}}\right)^{\mathrm{T}} V_{t-1}A$ and $S_2 = \mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_{3,c} \\ 0 \end{bmatrix} L_W^{-\mathrm{T}}\right)^{\mathrm{T}} V_{t-1}A$, and form $S = \begin{bmatrix} C \\ S_1 \\ S_2 \end{bmatrix}$.

Output: S which satisfies $S^{\mathrm{T}}S \approx X_\star$.

Parameter and output

1. In order to use FFT, t is usually chosen as powers of 2, namely $t = 2^k$. There is no strategy to determine a proper t in advance. In practice, we may choose a heuristic k, for example 5–8. If the output is a good approximation of the solution, then we stop here; otherwise, we use the output as a new initial guess, and run another round to achieve a better approximation; the process is repeated until convergence, namely some criterion is satisfied. The details and the validity of implementing a new initial guess are discussed in Section 3.4.

2. Note that $\operatorname{rank}(S) = \operatorname{rank}(V_{2^k})$. Numerically V_{2^k} probably has rank much less than $2^k l$. An obvious clue is that V_{2^k} contains a power series of A performing on C, and as k goes larger and larger, the terms in it become more and more likely to be linearly dependent. This implies that (3.14) is not a compact form. To deal with this, some compression technique may be brought in. This idea needs more discussions on the convergence, which is given in Section 3.4.

3. In the output, we do not give an approximation of $X_\star$ directly but its factor, namely a $tl \times n$ matrix S. If some compression technique is used during the process, an approximation of S would have relatively small low row rank, say $r \ll n$. Then in practice we only need the products of X and other matrices, for example, in obtaining the optimal control (3.2). The setting $r < n$ makes multiplication with $X_\star$'s factor save time and space. This is also considered in many literatures, e.g., [11].

Time complexity Complexity for S, the factor of X_t:

1. Step 1, compute $V_{t-1}A$, namely $CA, CA^2, \dots, CA^{t-1}$, in $(t-1)(2n-1)nl$ flops.

2. Step 2, compute $V_{t-1}B$ in $(t-1)(2n-1)lm$ flops.

3. Step 3, use M-step PCG ($Ntl\ln l$ for fast multiplication), to compute $Q_{*,*}$, in $\mathrm{O}(MNtl[\ln^2 t + \ln(tl)](l+m))$ flops.

4. Step 4, compute W in $(2tl-1)\frac{m(m+1)}{2} + m$ flops, and $L_Q, L_W, \widetilde{D}_Q, \widetilde{D}_W$ in $\frac{2}{3}l(l-1)(l+4)$ flops.

5. Step 5, compute $\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix} L_Q^{-\mathrm{T}}, \begin{bmatrix} Q_{3,c} \\ 0 \end{bmatrix} L_W^{-\mathrm{T}}$ in $tl^3 + (t-1)lm^2$ flops; compute S_1, S_2 in $\mathrm{O}(2Nntl\ln l)$ flops.

6. To sum up, assuming $l \ll n, m \ll n$ and omitting lower order terms, the total complexity is $2ln^2t + 2lmnt + \mathrm{O}(MNl(l+m)t\ln^2 t + 2Nltn\ln l) = \mathrm{O}\left(t(n^2 + \ln^2 t)\right)$ flops.

7. Suppose A is sparse, and the number of nonzero entries is $\operatorname{nnz}(A)$. Only Step 1 is different, and the total complexity is $\mathrm{O}(lt\operatorname{nnz}(A) + lmnt + MNl(l+m)t\ln^2 t + 2Nltn\ln l) = \mathrm{O}\left(t(\operatorname{nnz}(A) + n + \ln^2 t)\right)$ flops.

Space complexity

1. Step 1, store $V_{t-1}A$ in $(t-1)nl$ units.

2. Step 2, store $V_{t-1}B$ in $(t-1)lm$ units.

3. Step 3, store $Q_{*,*}$ in $(t-1)l(l+m)$ units.

4. Step 4, store W and then L_W, L_Q in part of the storage for $V_{t-1}B$. (The storage is enough and no extra units are needed because the three matrices need in total $m(m+1)/2+l(l+1)/2$ units, which is less than $(t-1)lm$.)

5. Step 5, store S_2 in the storage for $V_{t-1}B$ and additional storage, consuming in total $(t-1)ln$ units; store S_1 in the storage for $V_{t-1}A$.

6. to sum up, the total storage is $(t-1)l(2n+l+m) = \mathrm{O}(tn)$ units.

3.4 Arbitrary initial term

In this subsection, we consider (3.8) with an arbitrary initial $X_0 = \Gamma^{\mathrm T}\Gamma \succeq 0$:

$$X_0 = \Gamma^{\mathrm T}\Gamma, \qquad X_{t+1} = \mathscr{D}(X_t) = C^{\mathrm T}C + A^{\mathrm T}X_t(I+BB^{\mathrm T}X_t)^{-1}A. \tag{3.8$'$}$$

Theorem 3.3. *Write*

$$\Upsilon_t = \begin{bmatrix} A^{t-1}B & \cdots & AB & B\end{bmatrix}.$$

Then the sequence $\{X_t\}$ generated by (3.8′) is given by

$$X_t = \begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}^{\mathrm T}\left(I+\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}^{\mathrm T}\right)^{-1}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}, \tag{3.16}$$

where V_t, T_t is defined by (3.10).

Remark 3.1. Note that (3.16) coincides with (3.11) at $\Gamma = 0$.

Proof. First examine X_1.

$$\begin{aligned}\begin{bmatrix} C \\ \Gamma A\end{bmatrix}^{\mathrm T}\left(I+\begin{bmatrix} 0 \\ \Gamma B\end{bmatrix}\begin{bmatrix} 0 \\ \Gamma B\end{bmatrix}^{\mathrm T}\right)^{-1}\begin{bmatrix} C \\ \Gamma A\end{bmatrix} &= C^{\mathrm T}C + A^{\mathrm T}\Gamma^{\mathrm T}(I+\Gamma BB^{\mathrm T}\Gamma^{\mathrm T})^{-1}\Gamma A \\ &= C^{\mathrm T}C + A^{\mathrm T}\Gamma^{\mathrm T}\Gamma(I+BB^{\mathrm T}\Gamma^{\mathrm T}\Gamma)^{-1}A = X_1.\end{aligned}$$

Then examine the recursion.

$$\begin{aligned}
&C^{\mathrm T}C + A^{\mathrm T}X_t(I+BB^{\mathrm T}X_t)^{-1}A \\
&= C^{\mathrm T}C + A^{\mathrm T}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}^{\mathrm T}\left(I+\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}^{\mathrm T}\right)^{-1}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}\left(I+BB^{\mathrm T}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}^{\mathrm T}\left(I+\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}^{\mathrm T}\right)^{-1}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}\right)^{-1}A \\
&\overset{(1.2)}{=} C^{\mathrm T}C + A^{\mathrm T}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}^{\mathrm T}\left(I+\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}^{\mathrm T}+\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}BB^{\mathrm T}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}^{\mathrm T}\right)^{-1}\begin{bmatrix} V_t \\ \Gamma A^t\end{bmatrix}A \\
&= \begin{bmatrix} C \\ V_tA \\ \Gamma A^{t+1}\end{bmatrix}^{\mathrm T}\begin{bmatrix} I & \\ & I+\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}\begin{bmatrix} T_t \\ \Gamma\Upsilon_t\end{bmatrix}^{\mathrm T}+\begin{bmatrix} V_tB \\ \Gamma A^tB\end{bmatrix}\begin{bmatrix} V_tB \\ \Gamma A^tB\end{bmatrix}^{\mathrm T}\end{bmatrix}^{-1}\begin{bmatrix} C \\ V_tA \\ \Gamma A^{t+1}\end{bmatrix} \\
&= \begin{bmatrix} V_{t+1} \\ \Gamma A^{t+1}\end{bmatrix}^{\mathrm T}\left(I+\begin{bmatrix} 0 & 0 \\ V_tB & T_t \\ \Gamma A^tB & \Gamma\Upsilon_t\end{bmatrix}\begin{bmatrix} 0 & 0 \\ V_tB & T_t \\ \Gamma A^tB & \Gamma\Upsilon_t\end{bmatrix}^{\mathrm T}\right)^{-1}\begin{bmatrix} V_{t+1} \\ \Gamma A^{t+1}\end{bmatrix} \\
&= \begin{bmatrix} V_{t+1} \\ \Gamma A^{t+1}\end{bmatrix}^{\mathrm T}\left(I+\begin{bmatrix} T_{t+1} \\ \Gamma\Upsilon_{t+1}\end{bmatrix}\begin{bmatrix} T_{t+1} \\ \Gamma\Upsilon_{t+1}\end{bmatrix}^{\mathrm T}\right)^{-1}\begin{bmatrix} V_{t+1} \\ \Gamma A^{t+1}\end{bmatrix} \\
&= X_{t+1}.
\end{aligned}$$

□

Theorem 3.4. *Let $Q_{2,c}, Q_{2,b}, Q_{3,t}, Q_{3,c}, W, \Xi_1, \Xi_2$ as in Theorem 3.2. Then the sequence X_t defined by (3.16) can be generated by*

$$X_t = \begin{bmatrix} C \\ \Xi_1 \\ \Xi_2 \\ \Xi_\Gamma\end{bmatrix}^{\mathrm T}\begin{bmatrix} I & & & \\ & (I\otimes Q_{2,b})^{-1} & & \\ & & (I\otimes W)^{-1} & \\ & & & W_\Gamma^{-1}\end{bmatrix}\begin{bmatrix} C \\ \Xi_1 \\ \Xi_2 \\ \Xi_\Gamma\end{bmatrix},$$

where

$$W_\Gamma = I + \Gamma \Upsilon_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} - \Xi_{1,\Gamma}^{\mathrm{T}} (I \otimes Q_{2,b})^{-1} \Xi_{1,\Gamma} - \Xi_{2,\Gamma}^{\mathrm{T}} (I \otimes W)^{-1} \Xi_{2,\Gamma},$$
$$\Xi_\Gamma = \Gamma A^t - \Xi_{1,\Gamma}^{\mathrm{T}} (I \otimes Q_{2,b})^{-1} \Xi_1 - \Xi_{2,\Gamma}^{\mathrm{T}} (I \otimes W)^{-1} \Xi_2,$$

and

$$\Xi_{1,\Gamma} = \mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix}\right)^{\mathrm{T}} \mathcal{L}_{l\times m}(V_{t-1}B)\Upsilon_{t-1}^{\mathrm{T}} A^{\mathrm{T}} \Gamma^{\mathrm{T}}, \qquad \Xi_{2,\Gamma} = \mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_{3,c} \\ 0 \end{bmatrix}\right)^{\mathrm{T}} \mathcal{L}_{l\times m}(V_{t-1}B)\Upsilon_{t-1}^{\mathrm{T}} A^{\mathrm{T}} \Gamma^{\mathrm{T}}.$$

Proof. By (3.16),

$$\begin{aligned}
X_t &= \begin{bmatrix} V_t \\ \Gamma A^t \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I + T_t T_t^{\mathrm{T}} & T_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} \\ \Gamma \Upsilon_t T_t^{\mathrm{T}} & I + \Gamma \Upsilon_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} \end{bmatrix}^{-1} \begin{bmatrix} V_t \\ \Gamma A^t \end{bmatrix} \\
&= (*)^{\mathrm{T}} \begin{bmatrix} I + T_t T_t^{\mathrm{T}} & \\ & I + \Gamma \Upsilon_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} - \Gamma \Upsilon_t T_t^{\mathrm{T}} (I + T_t T_t^{\mathrm{T}})^{-1} T_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} \end{bmatrix}^{-1} \begin{bmatrix} I & \\ -\Gamma \Upsilon_t T_t^{\mathrm{T}} (I + T_t T_t^{\mathrm{T}})^{-1} & I \end{bmatrix} \begin{bmatrix} V_t \\ \Gamma A^t \end{bmatrix} \\
&= \begin{bmatrix} V_t \\ \Xi_\Gamma \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} I + T_t T_t^{\mathrm{T}} & \\ & I + \Gamma \Upsilon_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} - \Gamma \Upsilon_t T_t^{\mathrm{T}} (I + T_t T_t^{\mathrm{T}})^{-1} T_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} \end{bmatrix}^{-1} \begin{bmatrix} V_t \\ \Xi_\Gamma \end{bmatrix} \\
&= V_t^{\mathrm{T}} (I + T_t T_t^{\mathrm{T}})^{-1} V_t + \Xi_\Gamma^{\mathrm{T}} \left[I + \Gamma \Upsilon_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} - \Gamma \Upsilon_t T_t^{\mathrm{T}} (I + T_t T_t^{\mathrm{T}})^{-1} T_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} \right]^{-1} \Xi_\Gamma,
\end{aligned}$$

where $\Xi_\Gamma = \Gamma A^t - \Gamma \Upsilon_t T_t^{\mathrm{T}} (I + T_t T_t^{\mathrm{T}})^{-1} V_t$ and $*$ is used to indicate the same part limited by the symmetry. Similarly to (3.12), by (3.15), writing $\mathcal{L} = \mathcal{L}_{l\times m}(V_{t-1}B)$, it can be simplified to

$$\begin{aligned}
X_t &= C^{\mathrm{T}} C + A^{\mathrm{T}} V_{t-1}^{\mathrm{T}} \left[I + \mathcal{L}\mathcal{L}^{\mathrm{T}} \right]^{-1} V_{t-1} A + \Xi_\Gamma^{\mathrm{T}} \left[I + \Gamma \Upsilon_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} - \Gamma A \Upsilon_{t-1} \mathcal{L}^{\mathrm{T}} (I + \mathcal{L}\mathcal{L}^{\mathrm{T}})^{-1} \mathcal{L} \Upsilon_{t-1}^{\mathrm{T}} A^{\mathrm{T}} \Gamma^{\mathrm{T}} \right]^{-1} \Xi_\Gamma \\
&= C^{\mathrm{T}} C + \Xi_1^{\mathrm{T}} (I \otimes Q_{2,b})^{-1} \Xi_1 + \Xi_2^{\mathrm{T}} (I \otimes W)^{-1} \Xi_2 \\
&\quad + \Xi_\Gamma^{\mathrm{T}} \left[I + \Gamma \Upsilon_t \Upsilon_t^{\mathrm{T}} \Gamma^{\mathrm{T}} - \Xi_{1,\Gamma}^{\mathrm{T}} (I \otimes Q_{2,b})^{-1} \Xi_{1,\Gamma} - \Xi_{2,\Gamma}^{\mathrm{T}} (I \otimes W)^{-1} \Xi_{2,\Gamma} \right]^{-1} \Xi_\Gamma,
\end{aligned}$$

where

$$\begin{aligned}
\Xi_\Gamma &= \Gamma A^t - \Gamma A \Upsilon_{t-1} \mathcal{L}^{\mathrm{T}} (I + \mathcal{L}\mathcal{L}^{\mathrm{T}})^{-1} V_{t-1} A \\
&= \Gamma A^t - \Xi_{1,\Gamma}^{\mathrm{T}} (I \otimes Q_{2,b})^{-1} \Xi_1 - \Xi_{2,\Gamma}^{\mathrm{T}} (I \otimes W)^{-1} \Xi_2.
\end{aligned}$$ □

Note that the product of two lower triangular block-Toeplitz matrices is still a lower triangular block-Toeplitz matrix. Hence writing $\Xi_1 = \begin{bmatrix} \Xi_{1,c} \\ \Xi_{1,b} \end{bmatrix}$ where $\Xi_{1,b} \in \mathbb{R}^{l\times n}$,

$$\begin{aligned}
\mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix}\right)^{\mathrm{T}} \mathcal{L}_{l\times m}(V_{t-1}B) &= \mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix}\right)^{\mathrm{T}} \left[I \otimes (CB) + \mathcal{L}_{l\times m}\left(\begin{bmatrix} 0 \\ V_{t-2}A \end{bmatrix}\right) I \otimes B \right] \\
&= \mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix}\right)^{\mathrm{T}} I \otimes (CB) + \mathcal{L}_{l\times n}\left(\begin{bmatrix} 0 \\ \Xi_{1,c} \end{bmatrix}\right) I \otimes B, \\
&= \mathcal{U}_{l\times l}\left((CB)^{\mathrm{T}} \begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix}\right)^{\mathrm{T}} + \mathcal{L}_{l\times m}\left(\begin{bmatrix} 0 \\ \Xi_{1,c} \end{bmatrix} B\right).
\end{aligned}$$

Similarly, writing $\Xi_2 = \begin{bmatrix} \Xi_{2,c} \\ \Xi_{2,b} \end{bmatrix}$ where $\Xi_{2,b} \in \mathbb{R}^{l\times n}$,

$$\mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_{3,c} \\ 0 \end{bmatrix}\right)^{\mathrm{T}} \mathcal{L}_{l\times m}(V_{t-1}B) = \mathcal{U}_{l\times l}\left((CB)^{\mathrm{T}} \begin{bmatrix} Q_{3,c} \\ 0 \end{bmatrix}\right)^{\mathrm{T}} + \mathcal{L}_{l\times m}\left(\begin{bmatrix} 0 \\ \Xi_{2,c} \end{bmatrix} B\right).$$

These can be used to reduce calculations for $\Xi_{1,\Gamma}$ and $\Xi_{2,\Gamma}$.

What kind of the initial Γ makes the iteration (3.8′) converge? Lemma 3.2 gives an easy sufficient condition, which can be immediately applied to Algorithm 1, and settles thorny situation where t must keep small, as is declared in the illustration on the parameter and output for Algorithm 1 in Section 3.3.

Lemma 3.2. *The unique positive semi-definite d-stabilizing solution $X_\star$ of the DARE (3.3) is an attractor (i.e., asymptotically stable fixed point) of the DRE (3.8) or (3.8′). Moreover, any symmetric matrix X satisfying either following condition lies in its attraction basin:*

1. $0 \preceq X \preceq X_\star$;
2. $\rho(B^{\mathrm{T}}(X - X_\star)B) \le 1 - \rho(A_{X_\star})$.

As a result, (3.8′) with the matrix above as its initial term converges to $X_\star$.

Proof. First calculate the differentials.

$$\begin{aligned}
\mathrm{d}\mathscr{D}(X) &= A^{\mathrm{T}}\,\mathrm{d}X(I+BB^{\mathrm{T}}X)^{-1}A + A^{\mathrm{T}}X\,\mathrm{d}\big((I+BB^{\mathrm{T}}X)^{-1}\big)\,A \\
&= A^{\mathrm{T}}\,\mathrm{d}X(I+BB^{\mathrm{T}}X)^{-1}A - A^{\mathrm{T}}X(I+BB^{\mathrm{T}}X)^{-1}BB^{\mathrm{T}}\,\mathrm{d}X(I+BB^{\mathrm{T}}X)^{-1}A \\
&= A^{\mathrm{T}}\left[I - X(I+BB^{\mathrm{T}}X)^{-1}BB^{\mathrm{T}}\right]\mathrm{d}X(I+BB^{\mathrm{T}}X)^{-1}A \\
&= A^{\mathrm{T}}(I+XBB^{\mathrm{T}})^{-1}\,\mathrm{d}X(I+BB^{\mathrm{T}}X)^{-1}A \\
&= A_X^{\mathrm{T}}\,\mathrm{d}XA_X,
\end{aligned}$$

where A_X is the closed loop matrix. In order to avoid the appearance of 4th-order tensor, we use vectorization to obtain

$$\mathrm{d}(\operatorname{vec}\mathscr{D}(x)) = A_X^{\mathrm{T}}\otimes A_X^{\mathrm{T}}\,\mathrm{d}(\operatorname{vec}X) \quad\text{and}\quad \frac{\mathrm{d}(\operatorname{vec}\mathscr{D}(x))}{\mathrm{d}(\operatorname{vec}X)} = A_X^{\mathrm{T}}\otimes A_X^{\mathrm{T}}.$$

Since $X_\star$ is d-stabilizing, $\rho(A_{X_\star}) < 1$ and thus the Fréchet derivative at $X_\star$ has norm less than 1, which guarantees $X_\star$ is an attractor.

For the first kind of matrices, by Lemma 3.1, $X_\star = \mathscr{D}^t(X_\star) \succeq \mathscr{D}^t(X) \succeq \mathscr{D}^t(0) \to X_\star$, which forces $\mathscr{D}^t(X) \to X_\star$. For the second one, the only thing to be needed is the Fréchet derivative at X has norm less than 1. Writing $X - X_\star = \Delta$ and using an norm induced by some vector norm,

$$\begin{aligned}
\|(I+BB^{\mathrm{T}}X)^{-1}A\| &= \|[I+(I+BB^{\mathrm{T}}X_\star)^{-1}BB^{\mathrm{T}}\Delta]^{-1}(I+BB^{\mathrm{T}}X_\star)^{-1}A\| \\
&\le \|[I+(I+BB^{\mathrm{T}}X_\star)^{-1}BB^{\mathrm{T}}\Delta]^{-1}\|\|(I+BB^{\mathrm{T}}X_\star)^{-1}A\| \\
&\le \frac{\|I\|}{1-\|(I+BB^{\mathrm{T}}X_\star)^{-1}BB^{\mathrm{T}}\Delta\|}\|(I+BB^{\mathrm{T}}X_\star)^{-1}A\| \\
&\le \frac{\|A_{X_\star}\|}{1-\|(I+BB^{\mathrm{T}}X_\star)^{-1}\|\|BB^{\mathrm{T}}\Delta\|} \\
&\le \frac{\|A_{X_\star}\|}{1-\|BB^{\mathrm{T}}\Delta\|},
\end{aligned}$$

which is less than 1, as long as $\rho(B^{\mathrm{T}}\Delta B) = \rho(BB^{\mathrm{T}}\Delta) \le 1 - \rho(A_{X_\star})$. □

According to Lemma 3.2, the compression technique can be used in Algorithm 1 without breaking its convergence. We roughly describe the process here: after performing Algorithm 1 for a small/mid t, a truncation technique (e.g., SVD/QR) is used on S to produce $S'_X = \Gamma$; then (3.16) is used to generate a new approximation; repeat this process until convergence. To decrease the number of calculations, the same t is used in each outer iteration.

4 CARE

Given a linear time-invariant control system in continuous-time:

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

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{l\times n}$. Suppose the following condition holds through out this section:

$$\boxed{(A,B)\text{ is c-stabilizable and }(C,A)\text{ is c-detectable,}}$$

or equivalently, $\operatorname{rank}(\begin{bmatrix}A-\lambda I & BB^{\mathrm{T}}\end{bmatrix}) = \operatorname{rank}(\begin{bmatrix}A^{\mathrm{T}}-\lambda I & C^{\mathrm{T}}C\end{bmatrix}) = n$ for any $\lambda \in \mathbb{C}\setminus\mathbb{C}_-$, where $\mathbb{C}_-$ is the open left half complex plane.

Its linear-quadratic optimal control can be expressed as

$$\arg\min_{u(t)} \int_0^\infty \left[y(t)^{\mathrm{T}}y(t) + u(t)^{\mathrm{T}}u(t)\right]\mathrm{d}t = -B^{\mathrm{T}}X_\star x(t),$$

where $X_\star$ is the unique symmetric positive semi-definite c-stabilizing solution X of the CARE [12, 14, 34, 39]:

$$\mathscr{C}(X) := A^{\mathrm{T}}X + XA - XBB^{\mathrm{T}}X + C^{\mathrm{T}}C = 0. \tag{4.1}$$

Here a solution X is called c-stabilizing, if the closed loop matrix $A_X = A - BB^{\mathrm{T}}X$ is c-stable, namely all of its eigenvalues lie in the open left half complex plane $\mathbb{C}_-$.

4.1 FTA and its equivalence to many other methods

Many numerical methods to solve CAREs are based on performing Cayley transformation on the associated Hamiltonian matrix

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

For example, the SDA (3.9) for DAREs is also valid for CAREs. If the initial terms are set by

$$A_0 = I + 2\gamma K_\gamma^{-\mathrm{T}}, \qquad G_0 = 2\gamma \widehat{A}_\gamma^{-1} BB^{\mathrm{T}} K_\gamma^{-1}, \qquad H_0 = 2\gamma K_\gamma^{-1} C^{\mathrm{T}} C \widehat{A}_\gamma^{-1},$$

where $\widehat{A}_\gamma = A - \gamma I, \gamma > 0$ and $K_\gamma = \widehat{A}_\gamma^{\mathrm{T}} + C^{\mathrm{T}}C\widehat{A}_\gamma^{-1}BB^{\mathrm{T}}$, then $I_n + G_k H_k$ are nonsingular for $k \ge 0$, and it holds that $A_k \to 0$, $G_k \to Y_\star$ (the solution to the dual CARE) and $H_k \to X_\star$, all quadratically [37]. The terms A_0, G_0, H_0 are given by the Cayley transformation $\mathcal{H} \mapsto (\mathcal{H} + \gamma I)(\mathcal{H} - \gamma I)^{-1}$.

Since the recursion of SDAs for DAREs and CAREs are the same, comparing the initial terms, it is clear that the SDA for CAREs is calculating a subsequence generated by this DRE:

$$X_0 = 0, \qquad X_{t+1} = H_0 + A_0^{\mathrm{T}} X_t \left(I + G_0 X_t\right)^{-1} A_0. \tag{4.2}$$

Write $Y_\gamma = C\widehat{A}_\gamma^{-1}B$, $B_\gamma = \sqrt{2\gamma}\widehat{A}_\gamma^{-1}B(I + Y_\gamma^{\mathrm{T}}Y_\gamma)^{-1/2}$, $C_\gamma = \sqrt{2\gamma}(I + Y_\gamma Y_\gamma^{\mathrm{T}})^{-1/2}C\widehat{A}_\gamma^{-1}$, $A_\gamma = I + 2\gamma\widehat{A}_\gamma^{-1} - B_\gamma Y_\gamma^{\mathrm{T}} C_\gamma$, and then

$$G_0 = 2\gamma\widehat{A}_\gamma^{-1}BB^{\mathrm{T}}(\widehat{A}_\gamma^{\mathrm{T}} + C^{\mathrm{T}}C\widehat{A}_\gamma^{-1}BB^{\mathrm{T}})^{-1} = 2\gamma\widehat{A}_\gamma^{-1}B(I + B^{\mathrm{T}}\widehat{A}_\gamma^{-\mathrm{T}}C^{\mathrm{T}}C\widehat{A}_\gamma^{-1}B)^{-1}B^{\mathrm{T}}\widehat{A}_\gamma^{-\mathrm{T}} = B_\gamma B_\gamma^{\mathrm{T}},$$

and similarly $H_0 = C_\gamma^{\mathrm{T}} C_\gamma, A_0 = A_\gamma$. The DRE (4.2) reads

$$X_0 = 0, \qquad X_{t+1} = C_\gamma^{\mathrm{T}} C_\gamma + A_\gamma^{\mathrm{T}} X_t \left(I + B_\gamma B_\gamma^{\mathrm{T}} X_t\right)^{-1} A_\gamma. \tag{4.3}$$

The form of (4.3) coincides with (3.8). Hence the discussions on DAREs can be adopted to CAREs.

Wong and Balakrishnan [44, 45] proposed the quadratic ADI method

$$\begin{aligned} X_0 &= 0, \\ X_{t+1/2}(A - \overline{\gamma}_{t+1}I - BB^{\mathrm{T}}X_t) &= -C^{\mathrm{T}}C - (A^{\mathrm{T}} + \overline{\gamma}_{t+1}I)X_t, \qquad \Re\gamma_{t+1} > 0, \\ (A^{\mathrm{T}} - \gamma_{t+1}I - X_{t+1/2}BB^{\mathrm{T}})X_{t+1} &= -C^{\mathrm{T}}C - X_{t+1/2}(A + \gamma_{t+1}I), \end{aligned}$$

and presented in [44, (10)]

$$X_{t+1} = C_{\gamma_{t+1}}^{\mathrm{T}} C_{\gamma_{t+1}} + A_{\gamma_{t+1}}^{\mathrm{T}} X_t \left(I + B_{\gamma_{t+1}} B_{\gamma_{t+1}}^{\mathrm{T}} X_t\right)^{-1} A_{\gamma_{t+1}},$$

which is the same as (4.3) as long as $\gamma_{t+1} = \gamma$.

Lin and Simoncini [38] developed the Cayley transformed Hamiltonian subspace iteration

$$\begin{aligned} \begin{bmatrix} M_{t+1} \\ N_{t+1} \end{bmatrix} &= (\mathcal{H} - \gamma_{t+1}I)^{-1}(\mathcal{H} + \overline{\gamma}_{t+1}I)\begin{bmatrix} I \\ -X_t \end{bmatrix}, \qquad \Re\gamma_{t+1} < 0, \\ X_{t+1} &= -N_{t+1}M_{t+1}^{-1}. \end{aligned}$$

Benner et al. [5] devoted the RADI method, originated from the incorporation technique (which we will illustrate later), and proved that if the initial approximation is 0 and the same shifts are used, the RADI method is equivalent to quadratic ADI method and Cayley transformed Hamiltonian subspace iteration, together with invariant subspace approach in [1, 4] for special cases.

From the analysis above, we can conclude that the FTA in this paper and the SDA are also equivalent to these methods under the same condition that the initial approximation is 0 and the shift γ is consistently used, in the sense that they all produce the same sequence (subsequence for SDA).

Clearly the FTA, Algorithm 1, can be performed on the corresponding DARE to obtain the solution of the CARE.

Rather than directly using the results in Section 3.2, we borrow the same analysis there and eventually obtain the following analogies of Theorems 3.1 and 3.2.

Theorem 4.1. *Let* $\widetilde{A} = I + 2\gamma\widehat{A}_\gamma^{-1}, \widetilde{B} = \sqrt{2\gamma}\widehat{A}_\gamma^{-1}B, \widetilde{C} = \sqrt{2\gamma}C\widehat{A}_\gamma^{-1}$. *Write*

$$\widetilde{V}_t = \begin{bmatrix} \widetilde{C} \\ \widetilde{C}\widetilde{A} \\ \widetilde{C}\widetilde{A}^2 \\ \vdots \\ \vdots \\ \widetilde{C}\widetilde{A}^{t-1} \end{bmatrix}, \qquad \widetilde{T}_t = \begin{bmatrix} Y_\gamma & & & & & \\ \widetilde{C}\widetilde{B} & Y_\gamma & & & & \\ \widetilde{C}\widetilde{A}\widetilde{B} & \widetilde{C}\widetilde{B} & \ddots & & & \\ \vdots & & \ddots & \ddots & & \\ \vdots & & & \widetilde{C}\widetilde{B} & Y_\gamma & \\ \widetilde{C}\widetilde{A}^{t-2}\widetilde{B} & \cdots & \cdots & \widetilde{C}\widetilde{A}\widetilde{B} & \widetilde{C}\widetilde{B} & Y_\gamma \end{bmatrix},$$

$\widetilde{T}_1 = Y_\gamma$. *Then the terms of the sequence* $\{X_t\}$ *generated by the DRE* (4.3) *are*

$$X_t = \widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t, \qquad t=1,2,\dots. \tag{4.4}$$

Moreover, $\{X_t\}$ *is monotonically nondecreasing, and* $X_t \to X_\star$*, the solution of CARE* (4.1).

Proof. The "moreover" part is the same as that for DAREs in Section 3.2 and hence omitted. Only (4.4) is proved here.

It is easy to verify that (4.4) is correct for $t=1$. Assuming (4.4) is correct for t, we are going to prove it is also correct for $t+1$. By the DRE (4.3),

$$\begin{aligned}
X_{t+1} &= \widetilde{C}^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}\widetilde{C} + (\widetilde{A}-\widetilde{B}Y_\gamma^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}\widetilde{C})^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}}\left(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}\right)^{-1}\widetilde{V}_t \\
&\quad \cdot\left(I+\widetilde{B}(I+Y_\gamma^{\mathrm{T}}Y_\gamma)^{-1}\widetilde{B}^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}}\left(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}\right)^{-1}\widetilde{V}_t\right)^{-1}(\widetilde{A}-\widetilde{B}Y_\gamma^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}\widetilde{C}) \\
&\overset{(1.2)}{=} \widetilde{C}^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}\widetilde{C} + (\widetilde{A}-\widetilde{B}Y_\gamma^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}\widetilde{C})^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}} \\
&\quad \cdot\left(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}+\widetilde{V}_t\widetilde{B}(I+Y_\gamma^{\mathrm{T}}Y_\gamma)^{-1}\widetilde{B}^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}}\right)^{-1}\widetilde{V}_t(\widetilde{A}-\widetilde{B}Y_\gamma^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}\widetilde{C}) \\
&= (*)^{\mathrm{T}}\begin{bmatrix} I+Y_\gamma Y_\gamma^{\mathrm{T}} & \\ & I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}+\widetilde{V}_t\widetilde{B}(I+Y_\gamma^{\mathrm{T}}Y_\gamma)^{-1}\widetilde{B}^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}}\end{bmatrix}^{-1}\begin{bmatrix}\widetilde{C} \\ \widetilde{V}_t\widetilde{A}-\widetilde{V}_t\widetilde{B}Y_\gamma^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}\widetilde{C}\end{bmatrix} \\
&= (*)^{\mathrm{T}}\begin{bmatrix} I+Y_\gamma Y_\gamma^{\mathrm{T}} & \\ & I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}+\widetilde{V}_t\widetilde{B}\widetilde{B}^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}}-\widetilde{V}_t\widetilde{B}Y_\gamma^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1}Y_\gamma\widetilde{B}^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}}\end{bmatrix}^{-1} \\
&\qquad\qquad \cdot\begin{bmatrix} I & \\ -\widetilde{V}_t\widetilde{B}Y_\gamma^{\mathrm{T}}(I+Y_\gamma Y_\gamma^{\mathrm{T}})^{-1} & I\end{bmatrix}\begin{bmatrix}\widetilde{C}\\ \widetilde{V}_t\widetilde{A}\end{bmatrix} \\
&= \begin{bmatrix}\widetilde{C}\\ \widetilde{V}_t\widetilde{A}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix} I+Y_\gamma Y_\gamma^{\mathrm{T}} & Y_\gamma\widetilde{B}^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}} \\ \widetilde{V}_t\widetilde{B}Y_\gamma^{\mathrm{T}} & I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}+\widetilde{V}_t\widetilde{B}\widetilde{B}^{\mathrm{T}}\widetilde{V}_t^{\mathrm{T}}\end{bmatrix}^{-1}\begin{bmatrix}\widetilde{C}\\ \widetilde{V}_t\widetilde{A}\end{bmatrix} \\
&= \begin{bmatrix}\widetilde{C}\\ \widetilde{V}_t\widetilde{A}\end{bmatrix}^{\mathrm{T}}\left(I+\begin{bmatrix} Y_\gamma & 0\\ \widetilde{V}_t\widetilde{B} & \widetilde{T}_t\end{bmatrix}\begin{bmatrix} Y_\gamma & 0\\ \widetilde{V}_t\widetilde{B} & \widetilde{T}_t\end{bmatrix}^{\mathrm{T}}\right)^{-1}\begin{bmatrix}\widetilde{C}\\ \widetilde{V}_t\widetilde{A}\end{bmatrix} \\
&= \widetilde{V}_{t+1}^{\mathrm{T}}(I+\widetilde{T}_{t+1}\widetilde{T}_{t+1}^{\mathrm{T}})^{-1}\widetilde{V}_{t+1}.
\end{aligned}$$

Here $*$ is still used to indicate the same part limited by the symmetry. □

Using the notations for Toeplitz matrices in Section 2, we have

$$\widetilde{T}_t = \mathcal{L}_{l\times m}\left(\begin{bmatrix} Y_\gamma \\ \widetilde{V}_{t-1}\widetilde{B}\end{bmatrix}\right) = I_t\otimes Y_\gamma + \begin{bmatrix} 0 & 0 \\ \mathcal{L}_{l\times m}(\widetilde{V}_{t-1}\widetilde{B}) & 0\end{bmatrix} = \begin{bmatrix} Y_\gamma & 0 \\ \widetilde{V}_{t-1}\widetilde{B} & \widetilde{T}_{t-1}\end{bmatrix}.$$

Theorem 4.2. *Let*

$$\left[I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}\right]\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix} = \begin{bmatrix} 0 \\ I_l\end{bmatrix}, \qquad Q_{2,b}\in\mathbb{R}^{l\times l}, \tag{4.5a}$$

$$\left[I+\widetilde{V}_{t-1}\widetilde{B}\widetilde{B}^{\mathrm{T}}\widetilde{V}_{t-1}^{\mathrm{T}}+\widetilde{T}_{t-1}\widetilde{T}_{t-1}^{\mathrm{T}}\right]Q_3 = \widetilde{V}_{t-1}\widetilde{B}, \tag{4.5b}$$

and $W = I - Q_3^{\mathrm{T}}\widetilde{V}_{t-1}\widetilde{B}$. *Then the sequence* X_t *defined by DRE* (4.4) *can be generated by*

$$X_t = \begin{bmatrix} \Xi_1 \\ \Xi_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix} (I\otimes Q_{2,b})^{-1} & \\ & (I\otimes[W+WY_\gamma^{\mathrm{T}}Y_\gamma W])^{-1}\end{bmatrix}\begin{bmatrix} \Xi_1 \\ \Xi_2\end{bmatrix}, \tag{4.6}$$

where

$$\Xi_1 = \mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix}\right)^{\mathrm{T}}\widetilde{V}_t, \qquad \Xi_2 = \mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_3 \\ 0\end{bmatrix}\right)^{\mathrm{T}}\widetilde{V}_t.$$

Proof. By Lemma 2.1 with $Y = -(Y^L)^{\mathrm{T}} \leftarrow Y_\gamma, D_{t-1} = -(D_{t-1}^L)^{\mathrm{T}} \leftarrow \widetilde{V}_{t-1}\widetilde{B}$,

$$\begin{aligned}
&(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1} \\
&= \mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix}\right)^{\mathrm{T}} + \mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_3 \\ 0\end{bmatrix}\right)(I\otimes[W+WY_\gamma^{\mathrm{T}}Y_\gamma W])^{-1}\mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_3 \\ 0\end{bmatrix}\right)^{\mathrm{T}}.
\end{aligned} \tag{4.7}$$

Then the result follows from (4.4). □

Theorem 4.2 suggests a similar algorithm, Algorithm 2, to approximate the solution of CAREs.

Algorithm 2 FFT-based Toeplitz-structured Approximation (FTA) for CAREs

Input: A, B, C and γ, t.

1: Compute $\widehat{A}_\gamma = A - \gamma I$ and generate the linear-system solver $\widehat{A}_\gamma^{-1}$ for the sparse A (or its PLU factorization $\widehat{A}_\gamma = PLU$ for the dense A).

2: Compute $(\texttt{tmp}) = U^{-1}L^{-1}P^{-1}B, \widetilde{C} = \sqrt{2\gamma}C\widehat{A}_\gamma^{-1}$by the linear solver or forward/backward substitution, and compute $\widetilde{B} = \sqrt{2\gamma}(\texttt{tmp}), Y_\gamma = C(\texttt{tmp})$.

3: Compute sequentially $\widetilde{C}\cdot\widetilde{A}, \widetilde{C}\widetilde{A}\cdot\widetilde{A}, \dots, \widetilde{C}\widetilde{A}^{t-3}\cdot\widetilde{A}, \widetilde{C}\widetilde{A}^{t-2}\cdot\widetilde{A}$ by the way $(\texttt{tmp})\widetilde{A} = (\texttt{tmp}) + 2\gamma(\texttt{tmp})\widehat{A}_\gamma^{-1}$ by the linear solver or forward/backward substitution, and form $\widetilde{V}_t$ by stacking $\widetilde{C}$ and the $t-1$ terms vertically in order, where the first $t-1$ terms consists of $\widetilde{V}_{t-1}$.

4: Compute $\widetilde{V}_{t-1}\widetilde{B}$ and form $\begin{bmatrix} Y_\gamma \\ \widetilde{V}_{t-1}\widetilde{B} \end{bmatrix}$.

5: Use Preconditioned Conjugate Gradient method to solve (4.5).

6: Compute $W = I - Q_3^{\mathrm{T}}\widetilde{V}_{t-1}\widetilde{B}, (\texttt{tmp}) = Y_\gamma W, \widetilde{W} = W + (\texttt{tmp})^{\mathrm{T}}(\texttt{tmp})$ and then the Cholesky factorizations of $Q_{2,b} = L_Q L_Q^{\mathrm{T}}$ and $\widetilde{W} = L_W L_W^{\mathrm{T}}$.

7: Use fast multiplication to obtain $S_1 = \mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix} L_Q^{-\mathrm{T}}\right)^{\mathrm{T}} \widetilde{V}_t$ and $S_2 = \mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix} L_W^{-\mathrm{T}}\right)^{\mathrm{T}} \widetilde{V}_t$, and form $S = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$.

Output: S which satisfies $S^{\mathrm{T}}S \approx X_\star$.

Some remarks are given below to illustrate the algorithm.

Parameter and output

1. Considerations similar to Algorithm 1 have to be made. Eq. (4.6) is not a compact form either, and some truncation/reduction/shrinking technique may be brought in.

Time complexity Complexity for S, the factor of X_t:

1. Step 1, compute $\widehat{A}_\gamma$ and its PLU factorization in $n + \frac{n(n-1)(4n+1)}{6}$ flops.
2. Step 2, compute $\widetilde{C}, \widetilde{B}, Y_\gamma$ in $2n^2m + 2n^2l + nm + nl + lm(2n-1)$ flops.
3. Step 3, compute $\widetilde{V}_t$, namely $\widetilde{C}\widetilde{A}, \widetilde{C}\widetilde{A}^2, \dots, \widetilde{C}\widetilde{A}^{t-1}$, in $(t-1)[2n^2l + nl + nl]$ flops.
4. Step 4, compute $\widetilde{V}_{t-1}\widetilde{B}$ in $(t-1)(2n-1)lm$ flops.
5. Step 5, Use M-step PCG ($Ntl\ln l$ for fast multiplication), to compute $Q_{*,*}$, in $\mathrm{O}(MNtl[\ln^2 t + \ln(tl)](l+m))$ flops.
6. Step 6, compute $W, \widetilde{W}$ in $(2tl-1)\frac{m(m+1)}{2} + m + (2m-1)lm + (2l-1)\frac{m(m+1)}{2}$ flops, and L_Q, L_W in $\frac{2}{3}l(l-1)(l+4)$ flops.
7. Step 7, compute $\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix} L_Q^{-\mathrm{T}}, \begin{bmatrix} Q_{3,c} \\ 0 \end{bmatrix} L_W^{-\mathrm{T}}$ in $tl^3 + (t-1)lm^2$ flops; compute S_1, S_2 in $\mathrm{O}(2Nntl\ln l)$ flops.
8. To sum up, assuming $l \ll n, m \ll n$ and omitting lower order terms, the total complexity is $\frac{2}{3}n^3 + 2(l+m)n^2 + 2lmn + 2ln^2t + 2lmnt + \mathrm{O}(MNl(l+m)t\ln^2 t + 2Nltn\ln l) = \frac{2}{3}n^3 + \mathrm{O}\left(t(n^2 + \ln^2 t)\right)$ flops.
9. Suppose A is sparse, and the number of nonzero entries is $\mathrm{nnz}(A)$. The PLU factorization in Step 1 can be replaced by an iterative solver with at most P steps, such as CG, MINRES and GMRES. The total complexity is $\mathrm{O}(P(m+l)\,\mathrm{nnz}(A) + lmn + lt\,\mathrm{nnz}(A) + lmnt + MNl(l+m)t\ln^2 t + 2Nltn\ln l) = \mathrm{O}\left(t(\mathrm{nnz}(A) + n + \ln^2 t)\right)$ flops. It is worthwhile to mention that computing $\widehat{A}_\gamma^{-1}$ or solving the corresponding linear systems is necessary for all methods like RADI and the Cayley transformed Hamiltonian subspace iteration.

Space complexity

1. The storage is similar to that of Algorithm 1.

4.2 Incorporation technique

In the following, we consider the incorporation technique (a.k.a. defect correction). The key idea is: once an approximate solution $\widetilde{X}$ is obtained, letting the difference from the exact solution $X_\star$ be $\varDelta$, namely $X_\star = \widetilde{X} + \varDelta$, the difference satisfies $A^{\mathrm{T}}(\widetilde{X}+\varDelta) + (\widetilde{X}+\varDelta)A + C^{\mathrm{T}}C - (\widetilde{X}+\varDelta)BB^{\mathrm{T}}(\widetilde{X}+\varDelta) = 0$, from which an approximation $\widetilde{\varDelta}$ can be generated and then $\widetilde{X} + \widetilde{\varDelta}$ should be an approximate solution to the original equation better than $\widetilde{X}$. More details can be found in [24, 5]. The following lemma is important as the guarantee of the validity of the incorporation technique.

Lemma 4.1 ([5, Theorem 1])**.** *Let $\widetilde{X}$ be an approximation to a solution to (4.1).*

1. *$\varDelta = X_\star - \widetilde{X}$ is a solution to the equation*
$$(A - BB^{\mathrm{T}}\widetilde{X})^{\mathrm{T}}\varDelta + \varDelta(A - BB^{\mathrm{T}}\widetilde{X}) + \mathscr{C}(\widetilde{X}) - \varDelta BB^{\mathrm{T}}\varDelta = 0. \tag{4.8}$$
2. *Conversely, if $\varDelta$ is a solution to (4.8), then $\widetilde{X} + \varDelta$ is a solution to (4.1). Moreover, if $\widetilde{X} \succeq 0$ and $\varDelta$ is a c-stabilizing solution to (4.8), then $\widetilde{X} + \varDelta$ is the c-stabilizing solution to (4.1).*
3. *If $\widetilde{X} \succeq 0, \mathscr{C}(\widetilde{X}) \succeq 0$, then $\varDelta$ is the unique c-stabilizing solution to (4.8).*
4. *If $\widetilde{X} \succeq 0, \mathscr{C}(\widetilde{X}) \succeq 0$, then $\varDelta \preceq X_\star$.*

To make the incorporation technique useful for Algorithm 2, the fundamental problem we face is the low rank factorization of $\mathscr{C}(\widetilde{X})$. Let $\widetilde{X} = X_t$ that we have obtained.

Theorem 4.3. *Let $\mathbf{1}_t \in \mathbb{R}^t$ be a vector with each entry one. Then for X_t defined by DRE (4.3), $\mathscr{C}(X_t) = C_t^{\mathrm{T}} C_t$, where*
$$C_0 = C, \qquad C_t = C + \sqrt{2\gamma}(\mathbf{1}_t^{\mathrm{T}} \otimes I_l)(I + \widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t. \tag{4.9}$$

Proof. Clearly,
$$\begin{aligned}
\mathscr{C}(X_t) &= A^{\mathrm{T}}X_t + X_tA + C^{\mathrm{T}}C - X_tBB^{\mathrm{T}}X_t \\
&= [\gamma I + 2\gamma(\widetilde{A}-I)^{-\mathrm{T}}]X_t + X_t[\gamma I + 2\gamma(\widetilde{A}-I)^{-1}] + 2\gamma(\widetilde{A}-I)^{-\mathrm{T}}\widetilde{C}^{\mathrm{T}}\widetilde{C}(\widetilde{A}-I)^{-1} \\
&\qquad - 2\gamma X_t(\widetilde{A}-I)^{-1}\widetilde{B}\widetilde{B}^{\mathrm{T}}(\widetilde{A}-I)^{-\mathrm{T}}X_t \\
&= 2\gamma\left[(\widetilde{A}-I)^{-\mathrm{T}}\widetilde{C}^{\mathrm{T}}\widetilde{C}(\widetilde{A}-I)^{-1} + (\widetilde{A}-I)^{-\mathrm{T}}X_t + X_t(\widetilde{A}-I)^{-1} + X_t - X_t(\widetilde{A}-I)^{-1}\widetilde{B}\widetilde{B}^{\mathrm{T}}(\widetilde{A}-I)^{-\mathrm{T}}X_t\right].
\end{aligned}$$

Note that by (4.4)
$$\begin{aligned}
X_t(\widetilde{A}-I)^{-1}\widetilde{B} &= \widetilde{V}_t^{\mathrm{T}}(I + \widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t(\widetilde{A}-I)^{-1}\widetilde{B} \\
&= \widetilde{V}_t^{\mathrm{T}}(I + \widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\begin{bmatrix}\widetilde{C}\\ \widetilde{C}\widetilde{A}\\ \vdots\\ \widetilde{C}\widetilde{A}^{t-1}\end{bmatrix}(\widetilde{A}-I)^{-1}\widetilde{B} \\
&= \widetilde{V}_t^{\mathrm{T}}(I + \widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\begin{bmatrix}Y_\gamma\\ \widetilde{C}\widetilde{B} + Y_\gamma\\ \vdots\\ \widetilde{C}\widetilde{A}^{t-2}\widetilde{B} + \cdots + \widetilde{C}\widetilde{B} + Y_\gamma\end{bmatrix} \\
&= \widetilde{V}_t^{\mathrm{T}}(I + \widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{T}_t\begin{bmatrix}I_m\\ I_m\\ \vdots\\ I_m\end{bmatrix} = \widetilde{V}_t^{\mathrm{T}}(I + \widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{T}_t(\mathbf{1}_t \otimes I_m),
\end{aligned}$$

and
$$X_t(\widetilde{A}-I)^{-1} = \widetilde{V}_t^{\mathrm{T}}(I + \widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\begin{bmatrix}\widetilde{C}\\ \widetilde{C}\widetilde{A}\\ \vdots\\ \widetilde{C}\widetilde{A}^{t-1}\end{bmatrix}(\widetilde{A}-I)^{-1}$$

$$
\begin{aligned}
&= \widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\begin{bmatrix}\widetilde{C}(\widetilde{A}-I)^{-1}\\ \widetilde{C}+\widetilde{C}(\widetilde{A}-I)^{-1}\\ \vdots\\ \widetilde{C}\widetilde{A}^{t-2}+\cdots+\widetilde{C}+\widetilde{C}(\widetilde{A}-I)^{-1}\end{bmatrix}\\
&= \widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\left((\mathbf{1}_t\otimes I_l)\widetilde{C}(\widetilde{A}-I)^{-1}+\mathcal{L}_{l\times n}\left(\begin{bmatrix}0\\ \widetilde{V}_{t-1}\end{bmatrix}\right)(\mathbf{1}_t\otimes I_l)\right)\\
&= \widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\left((\mathbf{1}_t\otimes I_l)\widetilde{C}(\widetilde{A}-I)^{-1}+\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)\widetilde{V}_t\right),
\end{aligned}
$$

of which the last equality is guaranteed by

$$
\mathcal{L}_{l\times n}\left(\begin{bmatrix}0\\ \widetilde{V}_{t-1}\end{bmatrix}\right)(\mathbf{1}_t\otimes I_l)=\begin{bmatrix}0\\ \widetilde{C}\\ \widetilde{C}\widetilde{A}+\widetilde{C}\\ \vdots\\ \widetilde{C}\widetilde{A}^{t-2}+\cdots+\widetilde{C}\end{bmatrix}=\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)\widetilde{V}_t.
$$

Hence

$$
\begin{aligned}
\frac{1}{2\gamma}\left(\mathscr{C}(X_t)-C_t^{\mathrm{T}}C_t\right) &= \underline{(\widetilde{A}-I)^{-\mathrm{T}}\widetilde{C}^{\mathrm{T}}\widetilde{C}(\widetilde{A}-I)^{-1}}\\
&\quad+\underline{(\widetilde{A}-I)^{-\mathrm{T}}\widetilde{C}^{\mathrm{T}}(\mathbf{1}_t^{\mathrm{T}}\otimes I_l)(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t}+\widetilde{V}_t^{\mathrm{T}}\,\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t\\
&\quad+\underline{\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}(\mathbf{1}_t\otimes I_l)\widetilde{C}(\widetilde{A}-I)^{-1}}+\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\,\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)\widetilde{V}_t\\
&\quad+\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t-\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{T}_t(\mathbf{1}_t\otimes I_m)(\mathbf{1}_t^{\mathrm{T}}\otimes I_m)\widetilde{T}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t\\
&\quad-\underline{(\widetilde{A}-I)^{-\mathrm{T}}\widetilde{C}^{\mathrm{T}}\widetilde{C}(\widetilde{A}-I)^{-1}}-\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}(\mathbf{1}_t\otimes I_l)(\mathbf{1}_t^{\mathrm{T}}\otimes I_l)(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t\\
&\quad-\underline{(\widetilde{A}-I)^{-\mathrm{T}}\widetilde{C}^{\mathrm{T}}(\mathbf{1}_t^{\mathrm{T}}\otimes I_l)(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t}-\underline{\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}(\mathbf{1}_t\otimes I_l)\widetilde{C}(\widetilde{A}-I)^{-1}}\\
&= \widetilde{V}_t^{\mathrm{T}}\,\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t+\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\,\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)\widetilde{V}_t\\
&\quad+\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t-\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{T}_t(\mathbf{1}_t\otimes I_m)(\mathbf{1}_t^{\mathrm{T}}\otimes I_m)\widetilde{T}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t\\
&\qquad-\widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}(\mathbf{1}_t\otimes I_l)(\mathbf{1}_t^{\mathrm{T}}\otimes I_l)(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t\\
&= \widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\Bigg[(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})\,\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)^{\mathrm{T}}+\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})\\
&\qquad+(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})-\widetilde{T}_t(\mathbf{1}_t\otimes I_m)(\mathbf{1}_t^{\mathrm{T}}\otimes I_m)\widetilde{T}_t^{\mathrm{T}}-(\mathbf{1}_t\otimes I_l)(\mathbf{1}_t^{\mathrm{T}}\otimes I_l)\Bigg](I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t\\
&= \widetilde{V}_t^{\mathrm{T}}(I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\Bigg[\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}\,\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)^{\mathrm{T}}+\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}\\
&\qquad+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}}-\widetilde{T}_t(\mathbf{1}_t\otimes I_m)(\mathbf{1}_t^{\mathrm{T}}\otimes I_m)\widetilde{T}_t^{\mathrm{T}}\Bigg](I+\widetilde{T}_t\widetilde{T}_t^{\mathrm{T}})^{-1}\widetilde{V}_t\\
&= 0,
\end{aligned}
$$

of which the last equality holds for

$$
\mathcal{L}_{l\times l}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_l\end{bmatrix}\right)\widetilde{T}_t=\widetilde{T}_t\,\mathcal{L}_{m\times m}\left(\begin{bmatrix}0\\ \mathbf{1}_{t-1}\otimes I_m\end{bmatrix}\right). \qquad \square
$$

According to Theorem 4.3, we are able to make incorporation easily and solve (4.8). One thing worth mentioning is that C_t is not difficult to calculate. Putting (4.7) into (4.9),

$$
C_t=C+\sqrt{2\gamma}\begin{bmatrix}\Xi_{1,I}\\ \Xi_{2,I}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}(I\otimes Q_{2,b})^{-1} & \\ & (I\otimes[W+WY_\gamma^{\mathrm{T}}Y_\gamma W])^{-1}\end{bmatrix}\begin{bmatrix}\Xi_1\\ \Xi_2\end{bmatrix},
$$

where Ξ_1, Ξ_2 is defined as in (4.6), and

$$\Xi_{1,I} = \mathcal{U}_{l\times l}\left(\begin{bmatrix} Q_{2,c} \\ Q_{2,b} \end{bmatrix}\right)^{\mathrm{T}} (\mathbf{1}_t \otimes I_l), \qquad \Xi_{2,I} = \mathcal{U}_{l\times m}\left(\begin{bmatrix} Q_{3,c} \\ 0 \end{bmatrix}\right)^{\mathrm{T}} (\mathbf{1}_t \otimes I_l).$$

Theorem 4.3 gives the detailed form of (4.8) at $\widetilde{X} = X_t$. Note that the sequence $\{X_t\}$ can be generated by the RADI method introduced in [5], if the initial approximation is 0 and the same shifts are used. The RADI process is actually making incorporation at each $t = 1$ iteratively. As a direct consequence, we have the following result.

Theorem 4.4. *For X_t defined by DRE (4.3) and C_t defined by (4.9),*

$$\Delta_t := X_{t+1} - X_t = 2\gamma(C_t A_{t,\gamma}^{-1})^{\mathrm{T}}(I + C_t A_{t,\gamma}^{-1} BB^{\mathrm{T}} A_{t,\gamma}^{-\mathrm{T}} C_t^{\mathrm{T}})^{-1} C_t A_{t,\gamma}^{-1}, \qquad A_{t,\gamma} = A - BB^{\mathrm{T}} X_t - \gamma I,$$

or equivalently, Δ_t is the approximate solution to (4.8) at $\widetilde{X} = X_t$ generated by DRE (4.3) on $t = 1$.

5 Experiments and discussions

In this section, we will provide several examples to illustrate the new algorithm FTA and compare it with some existing methods. As is stated in Section 4.1, many methods solve a CARE through an equivalent DARE. Hence here we only test the CARE of the form

$$A^{\mathrm{T}} X + XA - XBB^{\mathrm{T}} X + C^{\mathrm{T}} C = 0, \qquad A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{l\times n},$$

for the performance of the methods on CARE can be recognized as those on DARE. We will use these methods in the tests:

- FTA: our FFT-based Toeplitz-structured approximation with the incorporation technique;
- RKSM: rational Krylov subspace method [22, 18, 42];
- RADI+opt: RADI method [5], with the residual minimizing shifts;
- RADI+proj: RADI method with the residual Hamiltonian shifts;
- NK-ADI+GP: the Galerkin projected variant of Newton-Kleinman ADI method [6, 7, 10, 9];
- iNK-ADI+LS: the inexact variant of Newton-Kleinman ADI method with line search.

All experiments are done in MATLAB 2021a under the Windows 10 Professional 64-bit operating system on a PC with a Intel Core i7-8700 processor at 3.20GHz and 64GB RAM. The implementation of RKSM comes from the source codes from Simoncini's homepage[1] with some modifications; we use corresponding functions in the package M-M.E.S.S. version 2.1 [41] as the implementations of the last four methods.

The methods are intentionally chosen: RADI is the recommended method by the package M-M.E.S.S., and it is usually one of the fastest methods among the non-projective methods, and two different shift selection strategies are used, for there does not exist a definitely good one and both strategies are good in many tests; the two variants of NK-ADI are Newton-type methods; RKSM is a projection method. On the other hand, FTA, SDA, RADI, quadratic ADI and Cayley transformed Hamiltonian subspace iteration are theoretically equivalent if the shifts are the same; the last three are of the same type, so only one of them, namely RADI, is chosen; SDA is appropriate for small-to-mid scale dense problems, so we give up putting it into comparison.

Since classical performance indices behave very different in different methods, we directly use the accuracy vs. the running time to compare. The accuracy is measured by

$$\mathrm{NRes}(X) := \frac{\|\mathscr{C}(X)\|_{\mathrm{F}}}{\|\mathscr{C}(0)\|_{\mathrm{F}}} = \frac{\|A^{\mathrm{T}} X + XA - XBB^{\mathrm{T}} X + C^{\mathrm{T}} C\|_{\mathrm{F}}}{\|C^{\mathrm{T}} C\|_{\mathrm{F}}}.$$

In the following, three examples are tested, where the results are shown in Figure 5.1.

Example 5.1 (Rail)**.** The example is a version of the steel profile cooling model from the Oberwolfach Model Reduction Benchmark Collection, hosted at MORwiki [40]. The data include $A \preceq 0, E \succeq 0, B, C$ with $n = 79841, m = 7, l = 6$. Since we only focus on solving the CARE, E is simply dropped.

For the parameters, in FTA, we use a heuristic shift $\gamma = 8 \times 10^{-7}$ and in each incorporation step $\gamma \leftarrow \gamma/1.01$. Each of the other five methods has its own way to choose shifts, so we leave the task for their own. Similar arguments apply for the following examples.

[1] http://www.dm.unibo.it/~simoncin/software.html

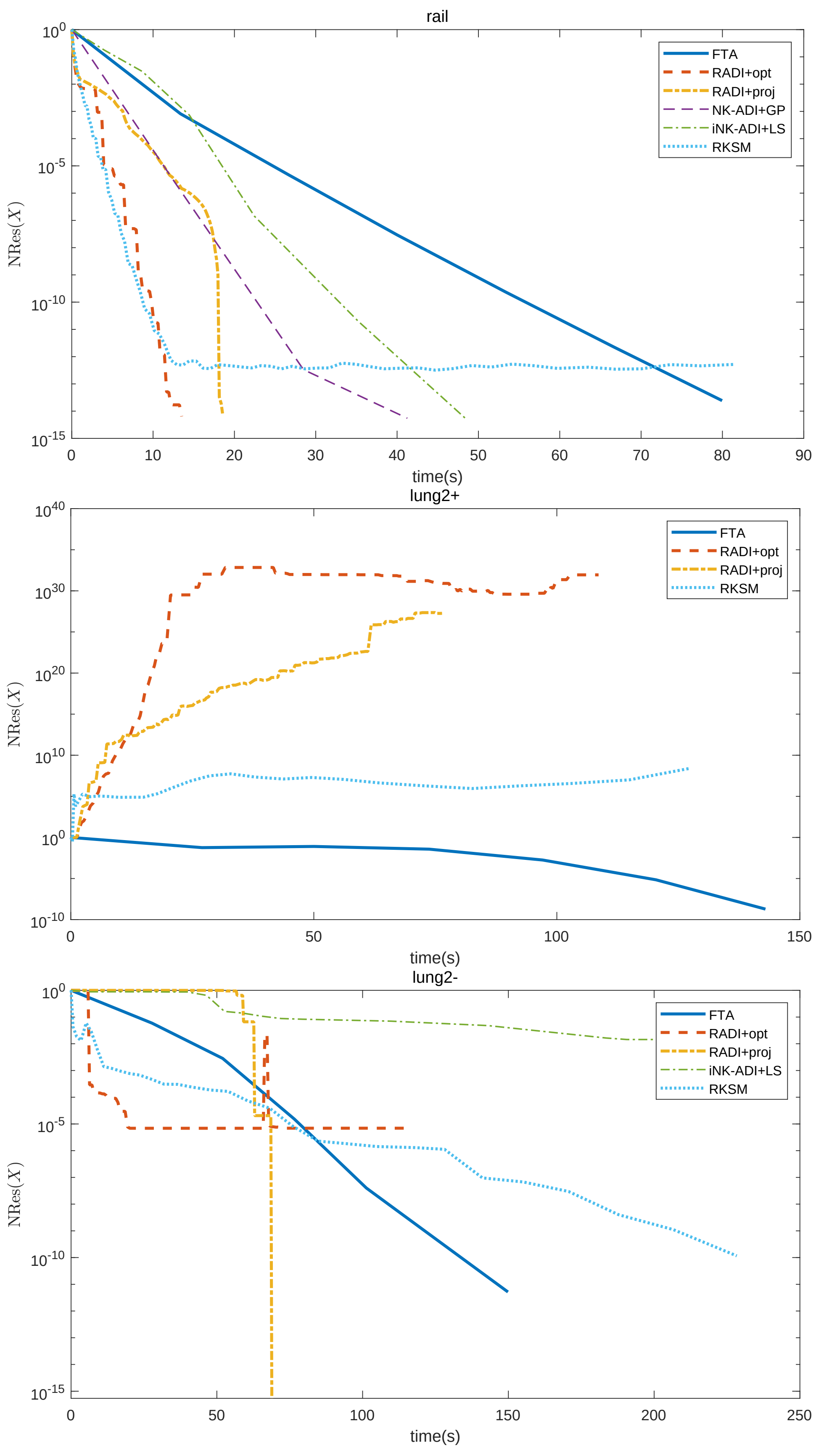

Figure 5.1: accuracy vs. time

In this example $A \prec 0$ and thus c-stable, which implies the properties of this problem are good. This results in the fact that all methods converge. We can see that the FTA is the slowest one among all the six methods. This phenomenon is reasonable. FTA and RADI are theoretically equivalent, while the only difference is that RADI has much more chances to choose different shifts to accelerate its convergence. Good shifts largely accelerate its convergence, and on the opposite, bad shifts would slow it down. RKSM and NK-ADI also benefit from the choice of shifts.

Example 5.2 (Lung2+)**.** The example is generated in this way: A is the matrix `lung2` in the SuiteSparse Matrix Collection [16] (formerly the University of Florida Sparse Matrix Collection), modelling temperature and water vapor transport in the human lung; B, C are generated by MATLAB function `rand`. Here $n = 109460, m = 10, l = 10$.

For the parameters, in FTA, we use a heuristic shift $\gamma = 5 \times 10^3$ and in each incorporation step $\gamma \leftarrow \gamma/1.01$.

In this example A is nonsymmetric and the eigenvalues of A lie in the right half plane, namely A is c-anti-stable, or $-A$ is c-stable. None of RADI+opt, RADI+proj, and RKSM converges. NK-ADI+GP and iNK-ADI+LS both report that non-stable Ritz values were detected and terminated the process. Only FTA produces a good approximate solution. This tells that the other five methods strongly rely on the stability of A. For example, a sufficient condition for achieving the convergence is that A is stable and the shifts γ_k satisfy the non-Blaschke condition $\sum_{k=1}^{+\infty} \frac{\Re(\gamma_k)}{1+|\gamma_k|^2} = -\infty$. However, the FTA works well even for the case that A is not stable, which implies that in this sense the FTA is more robust with respect to the spectrum of A.

Example 5.3 (Lung2$-$)**.** The example is almost the same with Example 5.2 except that the matrix `lung2` is used as $-A$ rather than A.

For the FTA, we still use a heuristic shift $\gamma = 5 \times 10^3$ and in each incorporation step $\gamma \leftarrow \gamma/1.01$.

In this example A is nonsymmetric and c-stable. Note that the only difference between RADI+opt and RADI+proj is the different shift selection strategies. RADI+opt tends to converge fast but finally stays at a low accuracy; RADI+proj becomes convergent very late but soonly converges in a very short time. The phenomenon illustrates that the choice of shifts fatally affects its speed of convergence. The iNK-ADI+LS converges very slowly, while the NK-ADI+GP reports that non-stable Ritz values were detected again. The RKSM converges in a fairly good speed but finally slow down. The FTA converges steadily in a predictable speed.

In another view, compared with Example 5.2, the running time of the FTA is nearly the same for different A's, so the running time is predictable and can be estimated in advance.

Summarizing the numerical results, we see that the FTA has two significant features:

1. FTA is robust in some sense and it converges no matter how the property of A is;
2. FTA has a steady convergence rate and the execution time is predictable, although in good cases it may converge slowly compared with other methods.

Moreover, it is easy to see that if A is dense, FTA needs the LU/PLU factorization only several times, while the other methods need as many as number of iterations, according to the number of used shifts.

6 Conclusion

We have presented our FFT-based Toeplitz-structured approximation method for computing the stabilizing solution of large-scale algebraic Riccati equations with low-rank structure. It is shown that the closed form given by operator theory under good assumptions is also valid for the general case, which is proved by matrix analysis. It is quite natural to ask whether the closed form can be directly produced by the analysis of unbounded linear operators, which would be a difficult task for future work. On the numerical front, our method works robust in some sense and few parameters are needed. As the readers may see, there is still possibility to improve the behavior by adopting more techniques. However, to keep this paper compact and concentrated, we leave it for another work.

A Displacement rank and Toeplitz matrix

In order to prove Lemma 2.1, we first give a few but urgent tips to the displacement rank and Toeplitz matrices, and interested readers are referred to the review paper [33] and the references therein.

For any matrix $R \in \mathbb{R}^{pn \times pn}$, its ($\pm$)-*displacement rank* $\alpha_\pm(R, p)$ with respect to block size $p \times p$, is defined by

$$\alpha_+(R,p) := \operatorname{rank}_p(R - Z_{n,p} R Z_{n,p}^{\mathrm{T}}), \qquad \alpha_-(R,p) := \operatorname{rank}_p(R - Z_{n,p}^{\mathrm{T}} R Z_{n,p}),$$

where $Z_{n,p} = \begin{bmatrix} 0 & 0 \\ I_{(n-1)p} & 0 \end{bmatrix}_{pn \times pn}$, and $\operatorname{rank}_p(\cdot)$ is considered as the rank of the linear transformation on the module $\mathbb{R}^{np \times p}$ over the ring $\mathbb{R}^{p \times p}$. For the case $p = 1$, $\operatorname{rank}_p(\cdot) = \operatorname{rank}(\cdot)$, the ordinary rank of matrices in $\mathbb{R}^{n \times n}$.

The definition is based on the following result, namely Lemma A.1.

Lemma A.1 ([32]). *Given $R_1, R_2 \in \mathbb{R}^{pn\times p}$ and $R \in \mathbb{R}^{pn\times pn}$, then*

$$R - Z_{n,p} R Z_{n,p}^{\mathrm{T}} = R_1 R_2^{\mathrm{T}} \iff R = \mathcal{L}_{p\times p}(R_1)\,\mathcal{L}_{p\times p}(R_2)^{\mathrm{T}},$$
$$R - Z_{n,p}^{\mathrm{T}} R Z_{n,p} = R_1 R_2^{\mathrm{T}} \iff R = \mathcal{U}_{p\times p}(R_1)\,\mathcal{U}_{p\times p}(R_2)^{\mathrm{T}}.$$

Lemma A.1 implies that for a matrix its displacement rank is related to how it can be expressed as a sum of products of block-Toeplitz matrices, as is shown in Lemma A.2.

Lemma A.2 ([32, 31, 33]). *Given a matrix $R \in \mathbb{R}^{pn\times pn}$.*

1. *Its $(+)$-displacement rank $\alpha_+(R,p)$ is the smallest integer β such that R can be written in the form*

$$R = \sum_{i=1}^{\beta} \mathcal{L}_{p\times p}(R_i)\,\mathcal{U}_{p\times p}(\widetilde{R}_i), \tag{A.2a}$$

where $R_i, \widetilde{R}_i \in \mathbb{R}^{pn\times p}$.

2. *Its $(-)$-displacement rank $\alpha_-(R,p)$ is the smallest integer β such that R can be written in the form*

$$R = \sum_{i=1}^{\beta} \mathcal{U}_{p\times p}(R_i)\,\mathcal{L}_{p\times p}(\widetilde{R}_i), \tag{A.2b}$$

where $R_i, \widetilde{R}_i \in \mathbb{R}^{pn\times p}$.

3. *If R is symmetric and positive semidefinite, (A.2a) and (A.2b) can be replaced respectively by*

$$R = \sum_{i=1}^{\beta} \mathcal{L}_{p\times p}(R_i)\,\mathcal{L}_{p\times p}(R_i)^{\mathrm{T}}, \qquad \text{and} \qquad R = \sum_{i=1}^{\beta} \mathcal{U}_{p\times p}(R_i)\,\mathcal{U}_{p\times p}(R_i)^{\mathrm{T}}.$$

4. *If R is nonsingular, then $\alpha_+(R,p) = \alpha_-(R^{-1},p), \alpha_-(R,p) = \alpha_+(R^{-1},p)$.*

Lemma A.2 demonstrates the relation between the displacement ranks of a matrix and its inverse, which is actually the theoretical foundation of the fast and superfast algorithms on Toeplitz matrices.

The following result, namely Lemma A.3, gives an expression of the inverse related to the displacement rank.

Lemma A.3 ([20]). *Given $R \in \mathbb{R}^{pn\times pn}$, suppose*

1. *R is nonsingular, and $R^{-1} = \begin{bmatrix} Q_{1,t} & Q_1^L \\ Q_1 & * \end{bmatrix} = \begin{bmatrix} * & Q_2 \\ Q_2^L & Q_{2,b} \end{bmatrix}$ where $Q_{1,t}, Q_{2,b} \in \mathbb{R}^{p\times p}$ are nonsingular;*

2. *$R - Z_{n,p} R Z_{n,p}^{\mathrm{T}} = \begin{bmatrix} * & * \\ * & D_1 \Sigma D_2^{\mathrm{T}} \end{bmatrix}$ where $D_1, D_2 \in \mathbb{R}^{p(n-1)\times p\alpha}$ and Σ is a diagonal matrix whose diagonal entries are ± 1;*

3. *writing $R = \begin{bmatrix} * & * \\ * & R_s \end{bmatrix}$ where $R_s \in \mathbb{R}^{p(n-1)\times p(n-1)}$, there exists Q_3, Q_3^L such that $R_s Q_3 = D_1, Q_3^L R_s = D_2^{\mathrm{T}}$.*

Then

$$\begin{aligned} R^{-1} = &-\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_1 \\ 0 \end{bmatrix}\right)(I\otimes Q_{1,t})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_1^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} + \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2^L & Q_{2,b} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &+ \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes W)^{-1}\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}, \end{aligned} \tag{A.3a}$$

or alternatively,

$$\begin{aligned} R^{-1} = &\mathcal{L}_{p\times p}\left(\begin{bmatrix} Q_{1,t} \\ Q_1 \end{bmatrix}\right)(I\otimes Q_{1,t})^{-1}\mathcal{L}_{p\times p}\left(\begin{bmatrix} Q_{1,t} & Q_1^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} - \mathcal{L}_{p\times p}\left(\begin{bmatrix} 0 \\ Q_2 \end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{L}_{p\times p}\left(\begin{bmatrix} 0 & Q_2^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &- \mathcal{L}_{p\times p\alpha}\left(\begin{bmatrix} 0 \\ Q_3 \end{bmatrix}\right)(I\otimes W)^{-1}\mathcal{L}_{p\times p\alpha}\left(\begin{bmatrix} 0 & Q_3^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}, \end{aligned} \tag{A.3b}$$

where $W = \Sigma - Q_3^L D_1$.

Moreover, if R is symmetric, then there exists a factorization to make $D_1 = D_2$; for that case, (A.3) can be rewritten by $Q_1^L = Q_1^{\mathrm{T}}, Q_2^L = Q_2^{\mathrm{T}}, Q_3^L = Q_3^{\mathrm{T}}$.

Remark A.1. Item 2 of Lemma A.3 implies that $R - Z_{n,p}^{\mathrm{T}} R Z_{n,p} = \begin{bmatrix} -D_1 \Sigma D_2^{\mathrm{T}} & * \\ * & * \end{bmatrix}$.

Note that (A.3) presents a sum of $\alpha + 2$ products of block-Toeplitz matrices, in which the number of terms may not be the smallest one, namely $\alpha_{\mp}(R, p)$.

In the following, we will derive a sum of the $\alpha_+(R, p) = \alpha_-(R^{-1}, p)$ terms, to coincide with Lemma A.2, of which the form is called a *shortest sum*. Using the same way a sum of $\alpha_-(R, p) = \alpha_+(R^{-1}, p)$ terms can also be derived, so we omit the details.

Write $R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_s \end{bmatrix}$, and then $R - Z_{n,p} R Z_{n,p}^{\mathrm{T}} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & D_1 \Sigma D_2^{\mathrm{T}} \end{bmatrix}$. Thus, $\alpha \le \alpha_+(R, p) \le \alpha + 2$.

On the other hand, by (1.3), under sufficient nonsingular conditions, it is easy to have

$$\begin{aligned} \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_s \end{bmatrix}^{-1} &= \begin{bmatrix} R_{11}^{-1} + R_{11}^{-1} R_{12} (R_s - R_{21} R_{11}^{-1} R_{12})^{-1} R_{21} R_{11}^{-1} & -R_{11}^{-1} R_{12} (R_s - R_{21} R_{11}^{-1} R_{12})^{-1} \\ -(R_s - R_{21} R_{11}^{-1} R_{12})^{-1} R_{21} R_{11}^{-1} & (R_s - R_{21} R_{11}^{-1} R_{12})^{-1} \end{bmatrix} \\ &= \begin{bmatrix} (R_{11} - R_{12} R_s^{-1} R_{21})^{-1} & -(R_{11} - R_{12} R_s^{-1} R_{21})^{-1} R_{12} R_s^{-1} \\ -R_s^{-1} R_{21} (R_{11} - R_{12} R_s^{-1} R_{21})^{-1} & R_s^{-1} + R_s^{-1} R_{21} (R_{11} - R_{12} R_s^{-1} R_{21})^{-1} R_{12} R_s^{-1} \end{bmatrix}. \end{aligned}$$

Compared with the conditions,

$$Q_1 = -R_s^{-1} R_{21} Q_{1,t}, \quad Q_1^L = -Q_{1,t} R_{12} R_s^{-1}, \quad Q_{1,t} = (R_{11} - R_{12} R_s^{-1} R_{21})^{-1}.$$

If $\alpha_+(R, p) = \alpha$, then it has to hold that $R - Z_{n,p} R Z_{n,p}^{\mathrm{T}} = \begin{bmatrix} S_1^{\mathrm{T}} \Sigma^{-1} S_2 & S_1^{\mathrm{T}} D_2^{\mathrm{T}} \\ D_1 S_2 & D_1 \Sigma D_2^{\mathrm{T}} \end{bmatrix}$ for some $S_1, S_2 \in \mathbb{R}^{p\alpha \times p}$. Clearly S_1, S_2 are of full column rank for R is nonsingular. Noticing $\Sigma^{-1} = \Sigma$, we have

$$\begin{aligned} Q_1 &= -R_s^{-1} D_1 S_2 Q_{1,t} = -Q_3 S_2 Q_{1,t}, \\ Q_1^L &= -Q_{1,t} S_1^{\mathrm{T}} D_2^{\mathrm{T}} R_s^{-1} = -Q_{1,t} S_1^{\mathrm{T}} Q_3^L, \\ Q_{1,t} &= (S_1^{\mathrm{T}} \Sigma^{-1} S_2 - S_1^{\mathrm{T}} D_2^{\mathrm{T}} R_s^{-1} D_1 S_2)^{-1} = (S_1^{\mathrm{T}} \Sigma^{-1} S_2 - S_1^{\mathrm{T}} Q_3^L D_1 S_2)^{-1} = (S_1^{\mathrm{T}} W S_2)^{-1}. \end{aligned}$$

Thus

$$\begin{aligned} &\mathcal{U}_{p\times p} \left(\begin{bmatrix} Q_1 \\ 0 \end{bmatrix} \right) (I \otimes Q_{1,t})^{-1} \mathcal{U}_{p\times p} \left(\begin{bmatrix} Q_1^L & 0 \end{bmatrix}^{\mathrm{T}} \right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p} \left(\begin{bmatrix} -Q_3 S_2 Q_{1,t} \\ 0 \end{bmatrix} \right) (I \otimes Q_{1,t})^{-1} \mathcal{U}_{p\times p} \left(\begin{bmatrix} -Q_{1,t} S_1^{\mathrm{T}} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}} \right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p\alpha} \left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix} \right) (I \otimes S_2 Q_{1,t}) (I \otimes Q_{1,t})^{-1} (I \otimes Q_{1,t} S_1^{\mathrm{T}}) \mathcal{U}_{p\times p\alpha} \left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}} \right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p\alpha} \left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix} \right) (I \otimes S_2 Q_{1,t} S_1^{\mathrm{T}}) \mathcal{U}_{p\times p\alpha} \left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}} \right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p\alpha} \left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix} \right) (I \otimes S_2 (S_1^{\mathrm{T}} W S_2)^{-1} S_1^{\mathrm{T}}) \mathcal{U}_{p\times p\alpha} \left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}} \right)^{\mathrm{T}}. \end{aligned}$$

Note that

$$\left[W^{-1} - S_2 \left(S_1^{\mathrm{T}} W S_2 \right)^{-1} S_1^{\mathrm{T}} \right] W S_2 = 0.$$

Complement S_2 to a nonsingular matrix $\begin{bmatrix} S_2 & S_2^c \end{bmatrix}$, and then

$$\begin{aligned} \left[W^{-1} - S_2 \left(S_1^{\mathrm{T}} W S_2 \right)^{-1} S_1^{\mathrm{T}} \right] W \begin{bmatrix} S_2 & S_2^c \end{bmatrix} &= \begin{bmatrix} S_2 & S_2^c \end{bmatrix} \begin{bmatrix} 0 & -\left(S_1^{\mathrm{T}} W S_2 \right)^{-1} S_1^{\mathrm{T}} W S_2^c \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} S_2 & S_2^c \end{bmatrix} \begin{bmatrix} -\left(S_1^{\mathrm{T}} W S_2 \right)^{-1} S_1^{\mathrm{T}} W S_2^c \\ I \end{bmatrix} \begin{bmatrix} 0 & I \end{bmatrix}, \end{aligned}$$

whose rank is $p(\alpha - 1)$. Write

$$W_1 = \begin{bmatrix} S_2 & S_2^c \end{bmatrix} \begin{bmatrix} -\left(S_1^{\mathrm{T}} W S_2 \right)^{-1} S_1^{\mathrm{T}} W S_2^c \\ I_{p(\alpha-1)} \end{bmatrix} \in \mathbb{R}^{p\alpha \times p(\alpha-1)}, \qquad W_1^L = \begin{bmatrix} 0 & I_{p(\alpha-1)} \end{bmatrix} \begin{bmatrix} S_2 & S_2^c \end{bmatrix}^{-1} W^{-1} \in \mathbb{R}^{p(\alpha-1)\times p\alpha},$$

and then $W^{-1} - S_2 \left(S_1^{\mathrm{T}} W S_2\right)^{-1} S_1^{\mathrm{T}} = W_1 W_1^L$. Hence

$$\begin{aligned} R^{-1} &= \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2^L & Q_{2,b} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} + \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes W_1 W_1^L)\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2^L & Q_{2,b} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} + \mathcal{U}_{p\times p(\alpha-1)}\left(\begin{bmatrix} Q_3 W_1 \\ 0 \end{bmatrix}\right)\mathcal{U}_{p\times p(\alpha-1)}\left(\begin{bmatrix} W_1^L Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}. \end{aligned} \tag{A.4}$$

If $\alpha_+(R,p) = \alpha+1$, then it holds that

$$R - Z_{n,p} R Z_{n,p}^{\mathrm{T}} = \begin{bmatrix} S_1^{\mathrm{T}}\Sigma^{-1}S_2 & S_1^{\mathrm{T}} D_2^{\mathrm{T}} \\ D_1 S_2 & D_1 \Sigma D_2^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} S_3 & D_3^{\mathrm{T}} \\ 0 & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} S_1^{\mathrm{T}}\Sigma^{-1}S_2 & S_1^{\mathrm{T}} D_2^{\mathrm{T}} \\ D_1 S_2 & D_1 \Sigma D_2^{\mathrm{T}} \end{bmatrix} + \begin{bmatrix} S_3 & 0 \\ D_3 & 0 \end{bmatrix}$$

for some $S_1, S_2 \in \mathbb{R}^{p\alpha\times p}$, $S_3 \in \mathbb{R}^{p\times p}$ and $D_3 \in \mathbb{R}^{p(n-1)\times p}$.

Consider the former one. Then,

$$\begin{aligned} Q_1 &= -R_s^{-1} D_1 S_2 Q_{1,t} = -Q_3 S_2 Q_{1,t}, \\ Q_1^L &= -Q_{1,t}(S_1^{\mathrm{T}} D_2^{\mathrm{T}} + D_3^{\mathrm{T}}) R_s^{-1} = -Q_{1,t} S_1^{\mathrm{T}} Q_3^L - Q_{1,t} D_3^{\mathrm{T}} R_s^{-1}, \\ Q_{1,t} &= (S_1^{\mathrm{T}}\Sigma^{-1}S_2 + S_3 - (S_1^{\mathrm{T}} D_2^{\mathrm{T}} + D_3^{\mathrm{T}}) R_s^{-1} D_1 S_2)^{-1} \\ &= (S_1^{\mathrm{T}}\Sigma^{-1}S_2 + S_3 - D_3^{\mathrm{T}} Q_3 S_2 - S_1^{\mathrm{T}} Q_3^L D_1 S_2)^{-1} \\ &= (S_3 - D_3^{\mathrm{T}} Q_3 S_2 + S_1^{\mathrm{T}} W S_2)^{-1}. \end{aligned}$$

Thus,

$$\begin{aligned} &\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_1 \\ 0 \end{bmatrix}\right)(I\otimes Q_{1,t})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_1^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p}\left(\begin{bmatrix} -Q_3 S_2 Q_{1,t} \\ 0 \end{bmatrix}\right)(I\otimes Q_{1,t})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} -Q_{1,t} S_1^{\mathrm{T}} Q_3^L - Q_{1,t} D_3^{\mathrm{T}} R_s^{-1} & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes S_2 Q_{1,t})(I\otimes Q_{1,t})^{-1}(I\otimes Q_{1,t} S_1^{\mathrm{T}})\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &\quad + \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes S_2 Q_{1,t})(I\otimes Q_{1,t})^{-1}(I\otimes Q_{1,t})\mathcal{U}_{p\times p}\left(\begin{bmatrix} D_3^{\mathrm{T}} R_s^{-1} & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes S_2 Q_{1,t} S_1^{\mathrm{T}})\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &\quad + \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes S_2 Q_{1,t})\mathcal{U}_{p\times p}\left(\begin{bmatrix} D_3^{\mathrm{T}} R_s^{-1} & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes S_2(S_3 - D_3^{\mathrm{T}} Q_3 S_2 + S_1^{\mathrm{T}} W S_2)^{-1} S_1^{\mathrm{T}})\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &\qquad + \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes S_2 Q_{1,t})\mathcal{U}_{p\times p}\left(\begin{bmatrix} D_3^{\mathrm{T}} R_s^{-1} & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}. \end{aligned}$$

Since

$$\begin{aligned} W^{-1} - S_2(S_3 - D_3^{\mathrm{T}} Q_3 S_2 + S_1^{\mathrm{T}} W S_2)^{-1} S_1^{\mathrm{T}} &= W^{-1}\left(W - W S_2(S_3 - D_3^{\mathrm{T}} Q_3 S_2 + S_1^{\mathrm{T}} W S_2)^{-1} S_1^{\mathrm{T}} W\right) W^{-1} \\ &\overset{(1.3)}{=} W^{-1}\left(W^{-1} + S_2(S_3 - D_3^{\mathrm{T}} Q_3 S_2)^{-1} S_1^{\mathrm{T}}\right)^{-1} W^{-1} \\ &= \left(W + W S_2(S_3 - D_3^{\mathrm{T}} Q_3 S_2)^{-1} S_1^{\mathrm{T}} W\right)^{-1} =: W_1^{-1}, \end{aligned}$$

we have

$$\begin{aligned} R^{-1} &= \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2^L & Q_{2,b} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} + \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes W_1)^{-1}\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &\qquad - \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes S_2 Q_{1,t})\mathcal{U}_{p\times p}\left(\begin{bmatrix} D_3^{\mathrm{T}} R_s^{-1} & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &= \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2 \\ Q_{2,b} \end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2^L & Q_{2,b} \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\ &\quad + \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right)(I\otimes W_1)^{-1}\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L - W_1 S_2 Q_{1,t} D_3^{\mathrm{T}} R_s^{-1} & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}. \end{aligned} \tag{A.5}$$

Similarly, for the latter one,

$$R^{-1} = \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2^L & Q_{2,b}\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} + \mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3 - R_s^{-1}D_3Q_{1,t}S_1^{\mathrm{T}}W_1 \\ 0\end{bmatrix}\right)(I\otimes W_1)^{-1}\mathcal{U}_{p\times p\alpha}\left(\begin{bmatrix} Q_3^L & 0\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{A.6}$$

where $W_1 = W + WS_2(S_3 - S_1^{\mathrm{T}}Q_3^LD_3)^{-1}S_1^{\mathrm{T}}W$.

To sum up, we have Lemma A.4.

Lemma A.4. *Given $R\in\mathbb{R}^{pn\times pn}$, suppose the conditions in Lemma A.3 hold. Then $\alpha\le\alpha_+(R,p)\le\alpha+2$, and the following statements hold.*

1. *if $\alpha_+(R,p)=\alpha$, then (A.4) is a shortest sum.*
2. *if $\alpha_+(R,p)=\alpha+1$, then (A.5) or (A.6) is a shortest sum.*
3. *if $\alpha_+(R,p)=\alpha+2$, then (A.3a) is a shortest sum.*

Moreover, if R is symmetric, then there exists a factorization to make $D_1 = D_2$; for that case, (A.4) to (A.6) and (A.3a) can be rewritten by $Q_1^L = Q_1^{\mathrm{T}}, Q_2^L = Q_2^{\mathrm{T}}, Q_3^L = Q_3^{\mathrm{T}}, S_1 = S_2, D_3 = 0$.

In the end, we prove Lemma 2.1.

Proof of Lemma 2.1. First consider the case $p_1 = p_2 = p$. Since

$$I - T_tT_t^L = \begin{bmatrix} I - YY^L & -YD_{t-1}^L \\ -D_{t-1}Y^L & I - D_{t-1}D_{t-1}^L - T_{t-1}T_{t-1}^L\end{bmatrix},$$

and

$$\begin{aligned}(I - T_tT_t^L) - Z_{t,p}(I - T_tT_t^L)Z_{t,p}^{\mathrm{T}} &= I - Z_{t,p}Z_{t,p}^{\mathrm{T}} - T_tT_t^L + Z_{t,p}T_tT_t^LZ_{t,p}^{\mathrm{T}} \\ &= \begin{bmatrix} I & \\ & 0\end{bmatrix} - \begin{bmatrix} Y & 0 \\ D_{t-1} & T_{t-1}\end{bmatrix}\begin{bmatrix} Y^L & D_{t-1}^L \\ 0 & T_{t-1}^L\end{bmatrix} + \begin{bmatrix} 0 & 0 \\ T_{t-1} & 0\end{bmatrix}\begin{bmatrix} 0 & T_{t-1}^L \\ 0 & 0\end{bmatrix} \\ &= \begin{bmatrix} I - YY^L & -YD_{t-1}^L \\ -D_{t-1}Y^L & -D_{t-1}D_{t-1}^L\end{bmatrix},\end{aligned}$$

we have $\alpha_+(I - T_tT_t^L, p) = 2$. By Lemma A.4, since $\alpha = 1$, the case falls in Item 2 with substitutions

$$D_3 \leftarrow 0, S_3 \leftarrow I, D_1 \leftarrow D_{t-1}, \varSigma \leftarrow -I, D_2^{\mathrm{T}} \leftarrow D_{t-1}^L, S_1^{\mathrm{T}} \leftarrow Y, S_2 \leftarrow Y^L.$$

Then (A.5) (or equivalently (A.6)) becomes

$$(I - T_tT_t^L)^{-1} = \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_2^L & Q_{2,b}\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} + \mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_3 \\ 0\end{bmatrix}\right)(I\otimes[W + WY^LYW])^{-1}\mathcal{U}_{p\times p}\left(\begin{bmatrix} Q_3^L & 0\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}},$$

where $Q_2, Q_{2,b}, Q_2^L, Q_3, Q_3^L, W$ is as in (2.1).

Then consider the case $p_1 > p_2$. Complement Y to a $p_1\times p_1$ matrix $\widetilde{Y} = \begin{bmatrix} Y & 0\end{bmatrix}$ and similarly for $\widetilde{D}_{t-1} = \begin{bmatrix} D_{t-1} & 0\end{bmatrix}, \widetilde{Y}^L = \begin{bmatrix} Y^L \\ 0\end{bmatrix}, \widetilde{D}_{t-1}^L = \begin{bmatrix} D_{t-1}^L \\ 0\end{bmatrix}$. Immediately we are able to use the result above on the case $p_1 = p_2$ to obtain $\left[I - \mathcal{L}_{p_1\times p_1}\left(\begin{bmatrix}\widetilde{Y} \\ \widetilde{D}_{t-1}\end{bmatrix}\right)\mathcal{L}_{p_1\times p_1}\left(\begin{bmatrix}\widetilde{Y}^L & \widetilde{D}_{t-1}^L\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}\right]^{-1}$. Note that

$$\mathcal{L}_{p_1\times p_1}\left(\begin{bmatrix}\widetilde{Y} \\ \widetilde{D}_{t-1}\end{bmatrix}\right)\mathcal{L}_{p_1\times p_1}\left(\begin{bmatrix}\widetilde{Y}^L & \widetilde{D}_{t-1}^L\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} = \begin{bmatrix} * & 0 & * & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ * & 0 & * & 0 & \cdots\end{bmatrix}\begin{bmatrix} * & \cdots & * \\ 0 & \cdots & 0 \\ * & \cdots & * \\ 0 & \cdots & 0 \\ \vdots & & \vdots\end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} * & * & \cdots \\ \vdots & \vdots & \\ * & * & \cdots \end{bmatrix} \begin{bmatrix} * & \cdots & * \\ * & \cdots & * \\ \vdots & & \vdots \end{bmatrix} \\
&= \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y \\ D_{t-1} \end{bmatrix}\right) \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y^L & D_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}.
\end{aligned}
$$

Thus, $\widetilde{Q}_2 = Q_2, \widetilde{Q}_2^L = Q_2^L, \widetilde{Q}_{2,b} = Q_{2,b}$, and $\widetilde{Q}_3 = \begin{bmatrix} Q_3 & 0 \end{bmatrix}, \widetilde{Q}_3^L = \begin{bmatrix} Q_3^L \\ 0 \end{bmatrix}$. Therefore,

$$
\begin{aligned}
\widetilde{W} &= -I - \begin{bmatrix} Q_3^L \\ 0 \end{bmatrix} \begin{bmatrix} D_{t-1} & 0 \end{bmatrix} = \begin{bmatrix} -I - Q_3^L D_{t-1} & 0 \\ 0 & -I \end{bmatrix} = \begin{bmatrix} W & \\ & -I \end{bmatrix}, \\
\widetilde{W}\widetilde{Y}^L\widetilde{Y}\widetilde{W} &= \begin{bmatrix} W & \\ & -I \end{bmatrix} \begin{bmatrix} Y^L \\ 0 \end{bmatrix} \begin{bmatrix} Y & 0 \end{bmatrix} \begin{bmatrix} W & \\ & -I \end{bmatrix} = \begin{bmatrix} WY^LYW & \\ & 0 \end{bmatrix}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} \widetilde{Q}_3 \\ 0 \end{bmatrix}\right) (I \otimes [\widetilde{W} + \widetilde{W}\widetilde{Y}^L\widetilde{Y}\widetilde{W}])^{-1} \mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} \widetilde{Q}_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} \\
&= \begin{bmatrix} * & 0 & * & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ * & 0 & * & 0 & \cdots \end{bmatrix} \begin{bmatrix} * & & & & \\ & -I & & & \\ & & * & & \\ & & & -I & \\ & & & & \ddots \end{bmatrix} \begin{bmatrix} * & \cdots & * \\ 0 & \cdots & 0 \\ * & \cdots & * \\ 0 & \cdots & 0 \\ \vdots & & \vdots \end{bmatrix} \\
&= \begin{bmatrix} * & * & \cdots \\ \vdots & \vdots & \\ * & * & \cdots \end{bmatrix} \begin{bmatrix} * & & \\ & * & \\ & & \ddots \end{bmatrix} \begin{bmatrix} * & \cdots & * \\ * & \cdots & * \\ \vdots & & \vdots \end{bmatrix} \\
&= \mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3 \\ 0 \end{bmatrix}\right) (I \otimes [W + WY^LYW])^{-1} \mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3^L & 0 \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}.
\end{aligned}
$$

Finally consider the case $p_1 < p_2$. Complement Y to a $p_2 \times p_2$ matrix $\widetilde{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix}$ and similarly for $\widetilde{D}_{t-1}^{\mathrm{T}} = \begin{bmatrix} * & 0 & * & 0 & \cdots \end{bmatrix}$ where $D_{t-1}^{\mathrm{T}} = \begin{bmatrix} * & * & \cdots \end{bmatrix}$, and $\widetilde{Y}^L = \begin{bmatrix} Y^L & 0 \end{bmatrix}, \widetilde{D}_{t-1}^L = \begin{bmatrix} * & 0 & * & 0 & \cdots \end{bmatrix}$ where $D_{t-1}^L = \begin{bmatrix} * & * & \cdots \end{bmatrix}$. To make things clear, two permutations P, P_s are used to make $P \begin{bmatrix} \widetilde{Y} \\ \widetilde{D}_{t-1} \end{bmatrix} = \begin{bmatrix} Y \\ D_{t-1} \\ 0 \end{bmatrix}, P_s \widetilde{D}_{t-1} = \begin{bmatrix} D_{t-1} \\ 0 \end{bmatrix}$. So $\begin{bmatrix} \widetilde{Y}^L & \widetilde{D}_{t-1}^L \end{bmatrix} P^{\mathrm{T}} = \begin{bmatrix} Y^L & D_{t-1}^L & 0 \end{bmatrix}, \widetilde{D}_{t-1}^L P_s^{\mathrm{T}} = \begin{bmatrix} D_{t-1}^L & 0 \end{bmatrix}$, and

$$
P\,\mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y} \\ \widetilde{D}_{t-1} \end{bmatrix}\right) = \begin{bmatrix} \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y \\ D_{t-1} \end{bmatrix}\right) \\ 0 \end{bmatrix}, \mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y}^L & \widetilde{D}_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} P^{\mathrm{T}} = \begin{bmatrix} \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y^L & D_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} & 0 \end{bmatrix}.
$$

Then we use the result above on the case $p_1 = p_2$ to obtain $\left[I - \mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y} \\ \widetilde{D}_{t-1} \end{bmatrix}\right) \mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y}^L & \widetilde{D}_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}\right]^{-1}$. Note that

$$
P\left[I - \mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y} \\ \widetilde{D}_{t-1} \end{bmatrix}\right) \mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y}^L & \widetilde{D}_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}\right] P^{\mathrm{T}} = \begin{bmatrix} I - \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y \\ D_{t-1} \end{bmatrix}\right) \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y^L & D_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} & \\ & I \end{bmatrix}.
$$

Thus,

$$
\begin{aligned}
P\begin{bmatrix} \widetilde{Q}_2 \\ \widetilde{Q}_{2,b} \end{bmatrix} &= P\left[I - \mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y} \\ \widetilde{D}_{t-1} \end{bmatrix}\right) \mathcal{L}_{p_2\times p_2}\left(\begin{bmatrix} \widetilde{Y}^L & \widetilde{D}_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}\right]^{-1} P^{\mathrm{T}} P \begin{bmatrix} 0 \\ I_{p_2} \end{bmatrix} \\
&= \begin{bmatrix} \left[I - \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y \\ D_{t-1} \end{bmatrix}\right) \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y^L & D_{t-1}^L \end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}\right]^{-1} & \\ & I \end{bmatrix} \begin{bmatrix} 0 \\ \begin{bmatrix} I_{p_1} & 0 \end{bmatrix} \\ 0 \\ \begin{bmatrix} 0 & I_{p_2-p_1} \end{bmatrix} \end{bmatrix}
\end{aligned}
$$

$$= \begin{bmatrix} \left[I - \mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y \\ D_{t-1}\end{bmatrix}\right)\mathcal{L}_{p_1\times p_2}\left(\begin{bmatrix} Y^L & D_{t-1}^L\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}}\right]^{-1}\begin{bmatrix} 0 \\ I_{p_1}\end{bmatrix} & 0 \\ 0 & \begin{bmatrix} 0 \\ I_{p_2-p_1}\end{bmatrix}\end{bmatrix} = \begin{bmatrix} \begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix} & 0 \\ 0 & \begin{bmatrix} 0 \\ I_{p_2-p_1}\end{bmatrix}\end{bmatrix},$$

and similarly, $\begin{bmatrix}\widetilde{Q}_2^L & \widetilde{Q}_{2,b}\end{bmatrix} P^{\mathrm{T}} = \begin{bmatrix}\begin{bmatrix} Q_2^L & Q_{2,b}\end{bmatrix} & 0 \\ 0 & \begin{bmatrix} 0 & I_{p_2-p_1}\end{bmatrix}\end{bmatrix}$. Therefore, $\widetilde{Q}_{2,b} = \begin{bmatrix} Q_{2,b} & 0 \\ 0 & I_{p_2-p_1}\end{bmatrix}$ and

$$\begin{aligned}
&P\mathcal{U}_{p_2\times p_2}\left(\begin{bmatrix}\widetilde{Q}_2 \\ \widetilde{Q}_{2,b}\end{bmatrix}\right) P^{\mathrm{T}} P (I\otimes \widetilde{Q}_{2,b})^{-1} P^{\mathrm{T}} P \mathcal{U}_{p_2\times p_2}\left(\begin{bmatrix}\widetilde{Q}_2^L & \widetilde{Q}_{2,b}\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} P^{\mathrm{T}} \\
&= \begin{bmatrix}\mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix}\right) & 0 \\ 0 & I\end{bmatrix}\begin{bmatrix}(I\otimes Q_{2,b})^{-1} & \\ & I\end{bmatrix}\begin{bmatrix}\mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} Q_2^L & Q_{2,b}\end{bmatrix}^{\mathrm{T}}\right) & 0 \\ 0 & I\end{bmatrix}^{\mathrm{T}} \\
&= \begin{bmatrix}\mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} Q_2 \\ Q_{2,b}\end{bmatrix}\right)(I\otimes Q_{2,b})^{-1}\mathcal{U}_{p_1\times p_1}\left(\begin{bmatrix} Q_2^L & Q_{2,b}\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} & \\ & I\end{bmatrix}.
\end{aligned}$$

Similarly,

$$P_s\widetilde{Q}_3 = P_s\left[I - \widetilde{D}_{t-1}\widetilde{D}_{t-1}^L - \widetilde{T}_{t-1}\widetilde{T}_{t-1}^L\right]^{-1} P_s^{\mathrm{T}} P_s \widetilde{D}_{t-1} = \begin{bmatrix}\left[I - D_{t-1}D_{t-1}^L - T_{t-1}T_{t-1}^L\right]^{-1} & \\ & I\end{bmatrix}\begin{bmatrix} D_{t-1} \\ 0\end{bmatrix} = \begin{bmatrix} Q_3 \\ 0\end{bmatrix},$$

and similarly, $\widetilde{Q}_3^L P_s^{\mathrm{T}} = \begin{bmatrix} Q_3^L & 0\end{bmatrix}$, and then

$$\begin{aligned}
\widetilde{W} &= -I - \widetilde{Q}_3^L P_s^{\mathrm{T}} P_s \widetilde{D}_{t-1} = -I - Q_3^L D_{t-1} = W, \\
\widetilde{W}\widetilde{Y}^L\widetilde{Y}\widetilde{W} &= W\begin{bmatrix} Y^L & 0\end{bmatrix}\begin{bmatrix} Y \\ 0\end{bmatrix} W = WY^LYW.
\end{aligned}$$

Hence

$$\begin{aligned}
&P\mathcal{U}_{p_2\times p_2}\left(\begin{bmatrix}\widetilde{Q}_3 \\ 0\end{bmatrix}\right)(I\otimes[\widetilde{W} + \widetilde{W}\widetilde{Y}^L\widetilde{Y}\widetilde{W}])^{-1}\mathcal{U}_{p_2\times p_2}\left(\begin{bmatrix}\widetilde{Q}_3^L & 0\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} P^{\mathrm{T}} \\
&= \begin{bmatrix}\mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3 \\ 0\end{bmatrix}\right) \\ 0\end{bmatrix}(I\otimes[W + WY^LYW])^{-1}\begin{bmatrix}\mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3^L & 0\end{bmatrix}^{\mathrm{T}}\right) \\ 0\end{bmatrix}^{\mathrm{T}} \\
&= \begin{bmatrix}\mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3 \\ 0\end{bmatrix}\right)(I\otimes[W + WY^LYW])^{-1}\mathcal{U}_{p_1\times p_2}\left(\begin{bmatrix} Q_3^L & 0\end{bmatrix}^{\mathrm{T}}\right)^{\mathrm{T}} & \\ & 0\end{bmatrix}. \qquad \square
\end{aligned}$$

References

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

[2] Brian D.O. Anderson. Second-order convergent algorithms for the steady-state Riccati equation. *Internat. J. Control*, 28(2):295–306, 1978.

[3] Panos J. Antsaklis and Anthony N. Michel. *A Linear Systems Primer*. Birkhäuser, Boston, 2007.

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

[5] P. Benner, Z. Bujanović, P. Kürschner, and J. Saak. RADI: a low-rank ADI-type algorithm for large-scale algebraic Riccati equations. *Numer. Math.*, 138:301–330, 2018.

[6] P. Benner, M. Heinkenschloss, J. Saak, and H. K. Weichelt. An inexact low-rank Newton-ADI merhod for large-scale Riccati equations. *Appl. Numer. Math.*, 108:125–142, 2016.

[7] P. Benner, J.-R. Li, and T. Penzl. Numerical solution of large Lyapunov equations, Riccati equations, and linear-quadratic control problems. *J. Numer. Lin. Alg. Appl.*, pages 755–777, 2008.

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

[9] P. Benner and J. Saak. A Newton-Galerkin-ADI method for large-scale algebraic Riccati equations. In *Applied Linear Algebra*. GAMM Workshop Applied and Numerical Linear Algebra, May 2010.

[10] P. Benner and J. Saak. Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey. *GAMM-Mitteilungen*, 36:32–52, 2013.

[11] Peter Benner, Zvonimir Bujanović, Patrick Kürschner, and Jens Saak. A numerical comparison of different solvers for large-scale, continuous-time algebraic Riccati equations and LQR problems. *SIAM J. Sci. Comput.*, 42(2):A957–A996, 2020.

[12] D. A. Bini, B. Iannazzo, and B. Meini. *Numerical Solution of Algebraic Riccati Equations*, volume 9 of *Fundamentals of Algorithm.* SIAM Publications, Philadelphia, 2012.

[13] T. Çimen. State-dependent Riccati equations. *IFAC Proc. Vol.*, 41:3761–3775, 2008.

[14] Eric K.-W. Chu, H. Y. Fan, and W.-W. Lin. A structure-preserving doubling algorithm for continuous-time algebraic Riccati equations. *Linear Algebra Appl.*, 396:55–80, 2005.

[15] Eric K.-W. Chu, H.-Y. Fan, W.-W. Lin, and C.-S. Wang. A structure-preserving doubling algorithm for periodic discrete-time algebraic Riccati equations. *Internat. J. Control*, 77:767–788, 2004.

[16] Timothy A. Davis and Yifan Hu. The university of Florida sparse matrix collection. *ACM Trans. Math. Software*, 38(1):Article 1, 2011. 25 pages.

[17] L. Dieci. Numerical integration of the differential Riccati equation and some related issues. *SIAM J. Numer. Anal.*, 29:781–815, 1992.

[18] V. Druskin and V. Simoncini. Adaptive rational Krylov subspaces for large-scale dynamical systems. *Sys. Contr. Lett.*, 60:546–560, 2011.

[19] V. Druskin, V. Simoncini, and M. Zaslavsky. Adaptive tangential interpolation in rational Krylov subspaces for MIMO dynamical systems. *SIAM J. Matrix Anal. Appl.*, 35:476–498, 2014.

[20] B. Friedlander, M. Morf, T. Kailath, and L. Ljung. New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices. *Linear Algebra Appl.*, 27:31–60, 1979.

[21] Z.-C. Guo, E. K.-W. Chu, X. Liang, and W.-W. Lin. A decoupled form of the structure-preserving doubling algorithm with low-rank structures. *ArXiv e-prints*, may 2020. 18 pages, arXiv: 2005.08288.

[22] Stefan Güttel. Rational Krylov approximation of matrix functions: numerical methods and optimal pole selection. *GAMM-Mitteilungen*, 36:8–31, 2013.

[23] M. Heyouni and K. Jbilou. An extended block Arnoldi algorithm for large-scale solutions of continuous-time algebraic Riccati equation. *Electron. Trans. Numer. Anal.*, 33:53–62, 2009.

[24] T.-M. Huang, R.-C. Li, and W.-W. Lin. *Structure-Preserving Doubling Algorithms for Nonlinear Matrix Equations.* Fundamentals of Algorithms. SIAM, Philadelphia, 2018.

[25] T.-M. Huang and W.-W. Lin. Structured doubling algorithms for weakly stabilizing Hermitian solutions of algebraic Riccati equations. *Linear Algebra Appl.*, 430:1452–1478, 2009.

[26] Vlad Ionescu, Cristian Oară, and Martin Weiss. *Generalized Riccati Theory and Robust Control: A Popov Function Approach.* John Wiley & Sons, Chichester, UK, 1999.

[27] K. Jbilou. An Arnoldi based algorithm for large algebraic Riccati equations. *Appl. Math. Lett.*, 19:437–444, 2006.

[28] Xiaoqing Jin. *Developments and Applications of Block Toeplitz Iterative Solvers.* Science Press, Beijing, 2002.

[29] Xiaoqing Jin. *Preconditioning Techniques for Toeplitz Systems.* Higher Education Press, Beijing, 2010.

[30] Thomas Kailath and Joohwan Chun. Generalized Gohberg-Semencul formulas for matrix inversion. *Oper. Theory: Adv. Appl.*, 40:231–246, 1989.

[31] Thomas Kailath and Joohwan Chun. Generalized displacement stucture for block-Toeplitz, Toeplitz-block, and Toeplitz-derived matrices. *SIAM J. Matrix Anal. Appl.*, 15(1):114–128, 1994.

[32] Thomas Kailath, Sun-Yuan Kung, and Martin Morf. Displacement ranks of matrices and linear equations. *J. Math. Anal. and Appl.*, 68:395–407, 1979.

[33] Thomas Kailath and Ali H. Sayed. Displacement structure: theory and applications. *SIAM Rev.*, 37(3):297–386, 1995.

[34] P. Lancaster and L. Rodman. Solutions of continuous and discrete time algebraic riccati equations: A review. In S. Bittanti, A. J. Laub, and J. C. Willems, editors, *The Riccati Equations*, pages 11–52. Springer-Verlag, Berlin, Heidelberg, 1991.

[35] P. Lancaster and L. Rodman. *Algebraic Riccati Equations.* The clarendon Press, Oxford Sciece Publications, New York, 1995.

[36] T. Li, E. K.-W. Chu, W.-W. Lin, and P. C.-Y. Weng. Solving large-scale continuous-time algebraic Riccati equations by doubling. *J. Comput. Appl. Math.*, 237:373–383, 2013.

[37] W.-W. Lin and S.-F. Xu. Convergence analysis of structure-preserving doubling algirithm for Riccati-type matrix equations. *SIAM J. Matrix Anal. Appl.*, 28:26–39, 2006.

[38] Y. Lin and V. Simoncini. A new subspace iteration method for the algebraic Riccati equation. *J. Numer. Lin. Alg. Appl.*, 22(1):26–47, 2015.

[39] V. L. Mehrmann. The autonomous linear quadratic control problems. In *Lecture Notes in Control and Information Sciences*, volume 163. Springer-Verlag, Berlin, 1991.

[40] Oberwolfach Benchmark Collection. Steel profile. hosted at MORwiki – Model Order Reduction Wiki, 2005.

[41] J. Saak, M. Köhler, and P. Benner. M-M.E.S.S.-2.1 – the matrix equations sparse solvers library, April 2021. see also: https://www.mpi-magdeburg.mpg.de/projects/mess.

[42] V. Simoncini, D. Szyld, and M. Monsalve. On two numerical methods for the solution of large-scale algebraic Riccati equations. *IMA J. Numer. Anal.*, 34:904–920, 2014.

[43] Gilbert Strang. A proposal for Toeplitz matrix calculations. *Stud. Appl. Math.*, 74:171–176, 1986.

[44] N. Wong and V. Balakrishnan. Quadratic alternating direction implicit iteration for the fast solution of algebraic Riccati equations. In *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems*, pages 373–376, 2005.

[45] N. Wong and V. Balakrishnan. Fast positive-real balanced truncation via quadratic alternating direction implicit iteration. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, 26(9):1725–1731, 2007.