Electronic Transactions on Numerical Analysis.
Volume 54, pp. 68–88, 2021.

ISSN 1068–9613.
DOI: 10.1553/etna_vol54s68

ETNA
Kent State University and
Johann Radon Institute (RICAM)

ON THE SOLUTION OF THE NONSYMMETRIC T-RICCATI EQUATION*

PETER BENNER† AND DAVIDE PALITTA†

Abstract. The nonsymmetric T-Riccati equation is a quadratic matrix equation where the linear part corresponds to the so-called T-Sylvester or T-Lyapunov operator that has previously been studied in the literature. It has applications in macroeconomics and policy dynamics. So far, it presents an unexplored problem in numerical analysis, and both theoretical results and computational methods are lacking in the literature. In this paper we provide some sufficient conditions for the existence and uniqueness of a nonnegative minimal solution, namely the solution with component-wise minimal entries. Moreover, the efficient computation of such a solution is analyzed. Both the small-scale and large-scale settings are addressed, and Newton-Kleinman-like methods are derived. The convergence of these procedures to the minimal solution is proven, and several numerical results illustrate the computational efficiency of the proposed methods.

Key words. T-Riccati equation, M-matrices, minimal nonnegative solution, Newton-Kleinman method

AMS subject classifications. 65F30, 15A24, 49M15, 39B42, 40C05

1. Introduction.

In this paper, we consider the nonsymmetric T-Riccati operator

$$\mathcal{R}_T : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}, \quad \mathcal{R}_T(X) := DX + X^T A - X^T BX + C,$$

where $A, B, C, D \in \mathbb{R}^{n\times n}$, and we provide sufficient conditions for the existence and uniqueness of a minimal solution $X_{\min} \in \mathbb{R}^{n\times n}$ to

$$\mathcal{R}_T(X) = 0. \tag{1.1}$$

The solution of the nonsymmetric T-Riccati equation (1.1) plays a role in solving dynamics generalized equilibrium (DSGE) problems [9, 22, 25]. "DSGE modeling is a method in macroeconomics that attempts to explain economic phenomena, such as economic growth and business cycles, and the effects of economic policy"[1]. Equations of the form (1.1) appear in certain procedures for solving DSGE models using perturbation-based methods [9, 25].

Taking inspiration from the (inexact) Newton-Kleinman method for standard algebraic Riccati equations, we illustrate efficient numerical procedures for solving (1.1). Both the small-scale and large-scale settings are addressed. In particular, in the latter framework, we assume the matrices A and D to be such that the matrix-vector products Av and Dw require $\mathcal{O}(n)$ floating point operations (flops) for any $v, w \in \mathbb{R}^n$. This is the case, for instance, when A and D are sparse. Moreover, we suppose B and C to be of low rank. These hypotheses mimic the usual assumptions adopted when dealing with large-scale standard algebraic Riccati equations; see, e.g., [2, 5, 7, 10, 13, 17, 18, 19, 23, 24, 21] and the recent survey paper [3].

The following is a synopsis of the paper. In Section 2 we present the result about the existence and uniqueness of a minimal solution $X_{\min}$ to (1.1), namely the nonnegative solution with component-wise minimal entries. A Newton-Kleinman method for the computation of such a $X_{\min}$ is derived in Section 3, and its convergence features are proven in Section 3.1. The large-scale setting is addressed in Section 3.2, where the convergence of an inexact Newton-Kleinman method equipped with a specific line search is illustrated. Some implementation details of the latter procedure are discussed in Section 3.3. Several numerical results showing

*Received July 3, 2020. Accepted September 20, 2020. Published online on November 20, 2020. Recommended by Dario Bini. This work was supported by the Australian Research Council (ARC) Discovery Grant No. DP1801038707.

†Department Computational Methods in Systems and Control Theory (CSC), Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany ({benner,palitta}@mpi-magdeburg.mpg.de).

[1] https://en.wikipedia.org/wiki/Dynamic_stochastic_general_equilibrium

ETNA
Kent State University and
Johann Radon Institute (RICAM)

the effectiveness of the proposed approaches are reported in Section 4, while our conclusions are given in Section 5.

Throughout the paper we adopt the following notation: The matrix inner product is defined as $\langle X, Y\rangle_F := \operatorname{trace}(Y^T X)$ so that the induced norm is $\|X\|^2 = \langle X, X\rangle_F$. I_n denotes the identity matrix of order n, and the subscript is omitted whenever the dimension of I is clear from the context. The brackets $[\cdot]$ are used to concatenate matrices of conforming dimensions. In particular, a MATLAB-like notation is adopted, where $[M, N]$ denotes the matrix obtained by augmenting M with N. $A \geq 0$ ($A > 0$) indicates a nonnegative (positive) matrix, that is, a matrix whose entries are all nonnegative (positive). Clearly, $A \leq 0$ ($A < 0$) if $-A \geq 0$ ($-A > 0$), and $A \geq B$ if $A - B \geq 0$. Moreover, we recall that a matrix A is a Z-matrix if all its off-diagonal entries are nonpositive. It is easy to show that a Z-matrix can be written in the form $A = sI - N$, where $s \in \mathbb{R}$ and $N \geq 0$. If $s \geq \rho(N)$, where $\rho(\cdot)$ denotes the spectral radius, then A is called M-matrix.

Furthermore, we will always suppose that the following assumption holds.

ASSUMPTION 1.1. *We assume that*

- *B is nonnegative ($B \geq 0$) and C is nonpositive ($C \leq 0$).*
- *$I \otimes D + (A^T \otimes I)\Pi$ is a nonsingular M-matrix where $\otimes$ denotes the Kronecker product while $\Pi \in \mathbb{R}^{n^2 \times n^2}$ is a permutation matrix given by $\Pi := \sum_{i=1}^{n} \sum_{j=1}^{n} E_{i,j} \otimes E_{j,i}$.*

The matrix $E_{i,j} \in \mathbb{R}^{n \times n}$ in Assumption 1.1 is the matrix whose (i, j)-th entry is 1 while all the others are zero.

Notice that $I \otimes D + (A^T \otimes I)\Pi$ being a nonsingular M-matrix implies that the T-Sylvester operator $\mathcal{S}_T : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$, $\mathcal{S}_T(X) := DX + X^T A$, has a nonnegative inverse, i.e., $\mathcal{S}_T^{-1}(X) \geq 0$ for $X \geq 0$. For the standard Sylvester operator $\mathcal{S} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$, $\mathcal{S}(X) := DX + XA$, this is guaranteed by assuming A, D to be nonsingular M-matrices; see, e.g., [10, Theorem A.20]. Another important consequence of Assumption 1.1 is the monotonicity of $\mathcal{S}_T$, i.e., $\mathcal{S}_T(X) \geq 0$ implies $X \geq 0$; see, e.g., [8].

It is not easy to analyze the impact that Assumption 1.1 has on the coefficient matrices A and D. Nevertheless, a careful inspection of the ordering of the entries of $I \otimes D + (A^T \otimes I)\Pi$ shows that if the latter is a singular M-matrix, then $A \leq 0$. Indeed, every entry of A appears, at least once, as an off-diagonal entry in $I \otimes D + (A^T \otimes I)\Pi$, and since the off-diagonal components of an M-matrix are nonpositive, it must hold that $A \leq 0$.

2. Existence and uniqueness of a minimal solution. In this section we provide sufficient conditions for the existence and uniqueness of a minimal solution $X_{\min}$ of (1.1). Our result relies on the following fixed-point iteration:

$$\begin{aligned} X_0 &= 0, \\ DX_{k+1} + X_{k+1}^T A &= X_k^T B X_k - C, \quad k \geq 0. \end{aligned} \tag{2.1}$$

THEOREM 2.1. *The iterates computed by the fixed-point iteration (2.1) are such that*

$$X_{k+1} \geq X_k, \quad k \geq 0,$$

and, if there exists a nonnegative matrix Y such that $\mathcal{R}_T(Y) \geq 0$, then $X_k \leq Y$ for any $k \geq 0$. Moreover, under this assumption, the sequence $\{X_k\}_{k \geq 0}$ converges, and its limit is the minimal nonnegative solution $X_{\min}$ to (1.1).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Proof. We first show that $X_{k+1} \geq X_k$ for any $k \geq 0$ by induction on k. For $k = 0$, we have $X_1 = \mathcal{S}_T^{-1}(-C) \geq 0 = X_0$ as $C \leq 0$. We now assume that $X_{\bar{k}} \geq X_{\bar{k}-1}$ for a certain $\bar{k} > 0$, and we show that $X_{\bar{k}+1} \geq X_{\bar{k}}$. We have

$$\begin{aligned} X_{\bar{k}+1} &= \mathcal{S}_T^{-1}(X_{\bar{k}}^T B X_{\bar{k}} - C) \\ &= \mathcal{S}_T^{-1}(X_{\bar{k}}^T B X_{\bar{k}}) + \mathcal{S}_T^{-1}(-C) = \mathcal{S}_T^{-1}(X_{\bar{k}}^T B X_{\bar{k}}) + X_1 + X_{\bar{k}} - X_{\bar{k}} \\ &= \mathcal{S}_T^{-1}(X_{\bar{k}}^T B X_{\bar{k}}) + X_1 + X_{\bar{k}} - \mathcal{S}_T^{-1}(X_{\bar{k}-1}^T B X_{\bar{k}-1} - C) \\ &= \mathcal{S}_T^{-1}(X_{\bar{k}}^T B X_{\bar{k}} - X_{\bar{k}-1}^T B X_{\bar{k}-1}) + X_{\bar{k}}. \end{aligned}$$

Clearly, $X_{\bar{k}}^T \geq X_{\bar{k}-1}^T$, as $X_{\bar{k}} \geq X_{\bar{k}-1}$ by the induction hypothesis. Therefore, recalling that $B \geq 0$, we have

$$X_{\bar{k}}^T B X_{\bar{k}} - X_{\bar{k}-1}^T B X_{\bar{k}-1} \geq 0,$$

so that $X_{\bar{k}+1} \geq X_{\bar{k}}$.

We now suppose that there exists a nonnegative Y such that $\mathcal{R}_T(Y) \geq 0$, and we show that $X_k \leq Y$ for any $k \geq 0$ by induction on k once again. The result is straightforward for $k = 0$ as $X_0 = 0$. We now assume that $X_{\bar{k}} \leq Y$ for a certain $\bar{k} > 0$, and we show that $X_{\bar{k}+1} \leq Y$. Since $X_{\bar{k}} \leq Y$ and $B \geq 0$, $X_{\bar{k}}^T B X_{\bar{k}} \leq Y^T B Y$ so that $-X_{\bar{k}}^T B X_{\bar{k}} \geq -Y^T B Y$. Thus, we can write

$$0 \leq DY + Y^T A - Y^T B Y + C \leq DY + Y^T A - X_{\bar{k}}^T B X_{\bar{k}} + C,$$

and since by definition $-X_{\bar{k}}^T B X_{\bar{k}} + C = -D X_{\bar{k}+1} - X_{\bar{k}+1}^T A$, we get

$$0 \leq DY + Y^T A - D X_{\bar{k}+1} - X_{\bar{k}+1}^T A.$$

This means that $\mathcal{S}_T(Y - X_{\bar{k}+1}) \geq 0$, which implies $Y \geq X_{\bar{k}+1}$ thanks to the monotonicity of $\mathcal{S}_T$.

In conclusion, $\{X_k\}_{k \geq 0}$ is a nondecreasing, nonnegative sequence bounded from above, therefore it has a finite limit $\lim_{k \to +\infty} X_k = X_{\min} \geq 0$. Taking the limit on both sides of (2.1) shows that $X_{\min}$ is a solution of the equation $\mathcal{R}_T(X) = 0$. Moreover, $X_{\min}$ is the minimal nonnegative solution, as we have proven that $X_{\min} \leq Y$ for any nonnegative Y such that $\mathcal{R}_T(Y) \geq 0$. □

A similar result has been shown in [15, Theorem 2.3] for the (standard) nonsymmetric Riccati equation.

3. The (inexact) Newton-Kleinman method. Due to its possible slow convergence rate, the fixed-point iteration (2.1) may not be attractive for the actual computation of the minimal solution $X_{\min}$, and a Newton-Kleinman-like method can be more effective for this task.

The k-th iteration of the Newton method is defined as

$$\mathcal{R}_T'[X](X_{k+1} - X_k) = -\mathcal{R}_T(X_k),$$

where $\mathcal{R}_T'[X]$ denotes the Fréchet derivative of $\mathcal{R}_T$ at X. For the nonsymmetric T-Riccati operator, we have

$$\mathcal{R}_T'[X](Y) = DY + Y^T A - Y^T B X - X^T B Y = (D - X^T B)Y + Y^T(A - BX),$$

and therefore the $(k+1)$-st iterate of the Newton method is the solution of the T-Sylvester equation

$$(D - X_k^T B)X_{k+1} + X_{k+1}^T(A - BX_k) = -X_k^T B X_k - C. \tag{3.1}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Depending on the problem size n, different state-of-the-art methods can be employed for the solution of equations (3.1); see, e.g., [11, 12].

If n is moderate, say $n \leq \mathcal{O}(10^3)$, then dense methods based on some decomposition of the coefficient matrices can be employed to solve the T-Sylvester equations in (3.1). For instance, in [11, Section 3] an algorithm based on the generalized Schur decomposition of the pair (D, A^T) is presented for efficiently solving a T-Sylvester equation of the form $DX + X^T A = C$.

If the problem dimension does not allow for dense matrix operations, then equations (3.1) must be solved iteratively. The iterative solution of the T-Sylvester equations may introduce some inexactness in the Newton scheme leading to the so-called inexact Newton-Kleinman method and affecting the convergence features of the latter. By using tools similar to the ones presented in [5], in Section 3.2 we show how a specific line search guarantees the convergence of the inexact Newton method.

However, we first need to guarantee that the sequence $\{X_k\}_{k\geq 0}$ generated by (3.1) is well-defined and that it converges to $X_{\min}$; this is the topic of the next section.

3.1. A convergence result. In this section we prove the convergence properties of the Newton-Kleinman method (3.1). To this end, we first recall a couple of classic results about M-matrices; see, e.g., [8, Chapter 6].

LEMMA 3.1. *Let A be a Z-matrix. Then A is a nonsingular M-matrix if and only if there exists a nonnegative vector v such that $Av > 0$. Moreover, if A is a nonsingular M-matrix and $B \geq A$ is a Z-matrix, then B is also a nonsingular M-matrix.*

To prove convergence of the Newton method to the minimal nonnegative solution $X_{\min}$ to (1.1), we also need the following lemma.

LEMMA 3.2. *Assume that there exists a matrix $\overline{Y}$ such that $\mathcal{R}_T(\overline{Y}) > 0$. Then it follows that $I \otimes (D - X_{\min}^T B) + ((A - BX_{\min})^T \otimes I)\Pi$ is a nonsingular M-matrix.*

Proof. Since Assumption 1.1 holds, we have that $I \otimes D + (A^T \otimes I)\Pi = rI_{n^2} - N$, with $N \geq 0$, $r > \rho(N)$, and we can write

$$\begin{aligned} I \otimes (D - X_{\min}^T B) + ((A - BX_{\min})^T \otimes I)\Pi \\ &= I \otimes D + (A^T \otimes I)\Pi - (I \otimes X_{\min}^T B + (BX_{\min})^T \otimes I)\Pi) \\ &= rI - \underbrace{(N + (I \otimes X_{\min}^T B + (BX_{\min})^T \otimes I)\Pi)}_{\geq 0}, \end{aligned}$$

as B, $X_{\min} \geq 0$. Notice that $X_{\min} \geq 0$ exists since the hypotheses of Theorem 2.1 are fulfilled. Therefore, $I \otimes (D - X_{\min}^T B) + ((A - BX_{\min})^T \otimes I)\Pi$ is a Z-matrix.

Moreover,

$$\begin{aligned} (D - X_{\min}^T B)(\overline{Y} - X_{\min}) + (\overline{Y} - X_{\min})^T (A - BX_{\min}) \\ &= D\overline{Y} - X_{\min}^T B\overline{Y} - DX_{\min} + X_{\min}^T BX_{\min} \\ &\quad + \overline{Y}^T A - \overline{Y}^T BX_{\min} - X_{\min}^T A + X_{\min}^T BX_{\min}. \end{aligned}$$

Since $\mathcal{R}_T(X_{\min}) = 0$, it follows that $-DX_{\min} - X_{\min}^T A + X_{\min}^T BX_{\min} = C$. Moreover, adding and subtracting $\overline{Y}^T B\overline{Y}$, we get

$$\begin{aligned} (D - X_{\min}^T B)(\overline{Y} - X_{\min}) + (\overline{Y} - X_{\min})^T (A - BX_{\min}) \\ &= \mathcal{R}_T(\overline{Y}) + (\overline{Y} - X_{\min})^T B(\overline{Y} - X_{\min}). \end{aligned}$$

To conclude, we notice that $\overline{Y} - X_{\min} \geq 0$, as $X_{\min}$ is the minimal solution to (1.1) and $\mathcal{R}_T(\overline{Y}) > 0$. Therefore,

$$(D - X_{\min}^T B)(\overline{Y} - X_{\min}) + (\overline{Y} - X_{\min})^T (A - BX_{\min}) \geq \mathcal{R}_T(\overline{Y}) > 0.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

This means that $\mathrm{vec}(\overline{Y} - X_{\min})$ is a nonnegative vector such that

$$(I \otimes (D - X_{\min}^T B) + ((A - BX_{\min})^T \otimes I)\Pi)\mathrm{vec}(\overline{Y} - X_{\min}) > 0,$$

and $I \otimes (D - X_{\min}^T B) + ((A - BX_{\min})^T \otimes I)\Pi$ is thus a nonsingular M-matrix thanks to Lemma 3.1. □

THEOREM 3.3. *If the assumptions of Lemma 3.2 hold, then the sequence* $\{X_k\}_{k\geq 0}$ *computed by the Newton method* (3.1) *with* $X_0 = 0$ *is well-defined and* $X_k \leq X_{k+1} \leq X_{\min}$ *for any* $k \geq 0$. *Moreover,* $\{X_k\}_{k\geq 0}$ *converges to the minimal nonnegative solution* $X_{\min}$ *to* (1.1).

Proof. For the Newton method (3.1) with $X_0 = 0$, the matrix X_1 is given by

$$DX_1 + X_1^T A = -C.$$

Since the T-Sylvester operator $\mathcal{S}_T$ has a nonnegative inverse by Assumption 1.1 and $-C \geq 0$, the first iterate X_1 is nonnegative. Therefore the statements

$$X_k \leq X_{k+1}, \quad X_k \leq X_{\min}, \quad I \otimes (D - X_k^T B) + ((A - BX_k)^T \otimes I)\Pi \text{ is an M-matrix},$$

hold for $k = 0$. We now assume that they hold for a certain $\bar{k} > 0$, and we show them for $\bar{k} + 1$. We start proving that $X_{\bar{k}+1} \geq X_{\bar{k}}$. By definition, we have

$$(D - X_{\bar{k}}^T B)X_{\bar{k}+1} + X_{\bar{k}+1}^T(A - BX_{\bar{k}}) = -X_{\bar{k}}^T BX_{\bar{k}} - C, \tag{3.2}$$

so that

$$(D - X_{\bar{k}}^T B)(X_{\bar{k}+1} - X_{\bar{k}}) + (X_{\bar{k}+1} - X_{\bar{k}})^T(A - BX_{\bar{k}}) = -DX_{\bar{k}} - X_{\bar{k}}^T A + X_{\bar{k}}^T BX_{\bar{k}} - C.$$

We can write

$$\begin{aligned}
&- DX_{\bar{k}} - X_{\bar{k}}^T A + X_{\bar{k}}^T BX_{\bar{k}} - C \\
&\quad = -(D - X_{\bar{k}-1}^T B)X_{\bar{k}} - X_{\bar{k}}^T(A - BX_{\bar{k}-1}) - X_{\bar{k}-1}^T BX_{\bar{k}} \\
&\qquad - X_{\bar{k}}^T BX_{\bar{k}-1} + X_{\bar{k}}^T BX_{\bar{k}} - C \\
&\quad = X_{\bar{k}-1}^T BX_{\bar{k}-1} + C - X_{\bar{k}-1}^T BX_{\bar{k}} - X_{\bar{k}}^T BX_{\bar{k}-1} + X_{\bar{k}}^T BX_{\bar{k}} - C \\
&\quad = (X_{\bar{k}} - X_{\bar{k}-1})^T B(X_{\bar{k}} - X_{\bar{k}-1}) \geq 0,
\end{aligned}$$

since $X_{\bar{k}} \geq X_{\bar{k}-1}$ and $B \geq 0$. If $\mathcal{S}_T^{(k)}(X) := (D - X_{\bar{k}}^T B)X + X^T(A - BX_{\bar{k}})$, then $(\mathcal{S}_T^{(k)})^{-1}$ is nonnegative as the matrix $I \otimes (D - X_{\bar{k}}^T B) + ((A - BX_{\bar{k}})^T \otimes I)\Pi$ is a nonsingular M-matrix by the induction hypothesis. Therefore, $X_{\bar{k}+1} - X_{\bar{k}} \geq 0$.

We now show that $X_{k+1} \leq X_{\min}$. Considering again (3.2), we see that

$$\begin{aligned}
&(D - X_{\bar{k}}^T B)(X_{\bar{k}+1} - X_{\min}) + (X_{\bar{k}+1} - X_{\min})^T(A - BX_{\bar{k}}) \\
&\quad = -DX_{\min} - X_{\min}^T A + X_{\bar{k}}^T BX_{\min} + X_{\min}^T BX_{\bar{k}} - X_{\bar{k}}^T BX_{\bar{k}} - C.
\end{aligned}$$

We change sign and, by adding and subtracting $X_{\min}^T BX_{\min}$ on the right-hand side, we get

$$\begin{aligned}
&(D - X_{\bar{k}}^T B)(X_{\min} - X_{\bar{k}+1}) + (X_{\min} - X_{\bar{k}+1})^T(A - BX_{\bar{k}}) \\
&\qquad = DX_{\min} + X_{\min}^T A - X_{\bar{k}}^T BX_{\min} - X_{\min}^T BX_{\bar{k}} + X_{\bar{k}}^T BX_{\bar{k}} \\
&\qquad\quad + C + X_{\min}^T BX_{\min} - X_{\min}^T BX_{\min} \\
&\qquad = (X_{\min} - X_{\bar{k}})^T B(X_{\min} - X_{\bar{k}}) \geq 0,
\end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

where we have used the fact that $\mathcal{R}_T(X_{\min}) = 0$, $X_{\min} \geq X_{\bar{k}}$, and $B \geq 0$. Since $\mathcal{S}_T^{(k)}$ has a nonnegative inverse we conclude that $X_{\min} - X_{\bar{k}+1} \geq 0$.

The last statement that we have to prove is that the matrix

$$I \otimes (D - X_{\bar{k}+1}^T B) + ((A - BX_{\bar{k}+1})^T \otimes I)\Pi$$

is a nonsingular M-matrix. Since Assumption 1.1 holds, we have the representation $I \otimes D + (A^T \otimes I)\Pi = rI_{n^2} - N$, with $N \geq 0$, $r > \rho(N)$, and we can write

$$\begin{aligned} &I \otimes (D - X_{\bar{k}+1}^T B) + ((A - BX_{\bar{k}+1})^T \otimes I)\Pi \\ &\quad = I \otimes D + (A^T \otimes I)\Pi - (I \otimes X_{\bar{k}+1}^T B + (BX_{\bar{k}+1})^T \otimes I)\Pi) \\ &\quad = rI - \underbrace{(N + (I \otimes X_{\bar{k}+1}^T B + (BX_{\bar{k}+1})^T \otimes I)\Pi)}_{\geq 0}, \end{aligned}$$

as B, $X_{\bar{k}+1} \geq 0$. Therefore, $I \otimes (D - X_{\bar{k}+1}^T B) + ((A - BX_{\bar{k}+1})^T \otimes I)\Pi$ is a Z-matrix. Moreover,

$$\begin{aligned} &I \otimes (D - X_{\bar{k}+1}^T B) + ((A - BX_{\bar{k}+1})^T \otimes I)\Pi \\ &\qquad \geq I \otimes (D - X_{\min}^T B) + ((A - BX_{\min})^T \otimes I)\Pi \end{aligned}$$

since $X_{\bar{k}+1} \leq X_{\min}$ and $I \otimes (D - X_{\min}^T B) + ((A - BX_{\min})^T \otimes I)\Pi$ is a nonsingular M-matrix by Lemma 3.2. The matrix $I \otimes (D - X_{\bar{k}+1}^T B) + ((A - BX_{\bar{k}+1})^T \otimes I)$ is thus a nonsingular M-matrix by Lemma 3.1.

In conclusion, the Newton sequence $\{X_k\}_{k \geq 0}$ is well-defined, nondecreasing, and bounded from above. Therefore, $\{X_k\}_{k \geq 0}$ has a finite limit X_*, and, by taking the limit on both sides of (3.1), it is easy to show that it is also a solution of $\mathcal{R}_T(X) = 0$. Moreover, we can show by induction that $X_k \leq H$ for any $k \geq 0$ and $H \geq 0$ with $\mathcal{R}_T(H) \geq 0$. Since the inequality is preserved for $k \to +\infty$, we conclude that $X_* \leq H$, and X_* is thus the minimal solution of $\mathcal{R}_T(X) = 0$, i.e., $X_* = X_{\min}$. □

3.2. The large-scale setting. In this section, we consider T-Riccati equations of large dimension. In this setting, unless the data A, B, C, and D are equipped with some particular structure, equation (1.1) is not numerically tractable. For instance, the solution X would be, in general, a dense $n \times n$ matrix that cannot be stored. Therefore, as already mentioned, we assume that the matrices A and D are such that the matrix-vector products Av and Dw are computable in $\mathcal{O}(n)$ flops for any $v, w \in \mathbb{R}^n$. This is the case when, for instance, A and D are sparse. Moreover, we assume B and C to be of low rank, namely $B = B_1 B_2^T$, B_1, $B_2 \in \mathbb{R}^{n \times p}$, and $C = C_1^T C_2$, C_1, $C_2 \in \mathbb{R}^{q \times n}$, where $p + q \ll n$.

In case of algebraic Riccati equations, the assumptions above lead to a solution Z that is often numerically rank deficient [1], and low-rank approximations of the form $\overline{Z}\,\overline{Z}^T \approx Z$, $\overline{Z} \in \mathbb{R}^{n \times t}$, $t \ll n$, are therefore expected to be accurate. We think that also in the case of the nonsymmetric T-Riccati equation it is possible to show that the singular values of the solution X to (1.1) show a fast decay, and low-rank approximations can thus be sought. This may be proven by combining the arguments in [1] with bounds for the decay of the singular values of the solution of certain T-Sylvester equations [12]. However, this is beyond the scope of this paper. We restrict ourselves to illustrate how low-rank approximations turn out to be sufficiently accurate in the examples we tested.

Equation (1.1) can be written as

$$\mathcal{R}_T(X) = DX + X^T A - X^T B_1 B_2^T X + C_1^T C_2 = 0, \tag{3.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

and low-rank approximations to X are sought, namely we aim to compute and store only a couple of low-rank matrices $P_1, P_2 \in \mathbb{R}^{n\times t}$, $t \ll n$, such that $P_1 P_2^T \approx X$.

The results presented in the previous section are still valid in the large-scale setting for equation (3.3). The Newton method can still be applied, and the $(k+1)$-st iterate can be computed by solving the equation

$$(D - X_k^T B_1 B_2^T) X_{k+1} + X_{k+1}^T (A - B_1 B_2^T X_k) = -X_k B_1 B_2^T X_k - C_1^T C_2. \tag{3.4}$$

However, due to the large dimension of the problem, the exact solution to (3.4) cannot be computed, and only an approximation $\widetilde{X}_{k+1} \approx X_{k+1}$ can be constructed by, e.g., the projection methods presented in [12].

The iterative solution of equations (3.3) introduces some inexactness in the Newton scheme leading to an inexact Newton method. The convergence result stated in Theorem 2.1 no longer holds for the inexact variant of the Newton procedure, and a line search has to be performed to ensure the convergence of the overall scheme. This procedure is similar to the one presented in [5] for the algebraic Riccati equations.

Given a matrix $X_k \in \mathbb{R}^{n\times n}$, $\alpha > 0$, and $\eta_k \in (0,1)$, we want to compute a matrix $S_k \in \mathbb{R}^{n\times n}$ such that

$$\|\mathcal{R}_T'[X_k](S_k) + \mathcal{R}_T(X_k)\| \le \eta_k \|\mathcal{R}_T(X_k)\|, \tag{3.5}$$

and then define the next iterate of the inexact Newton-Kleinman scheme as

$$X_{k+1} := X_k + \lambda_k S_k, \tag{3.6}$$

where the step size $\lambda_k > 0$ is bounded away from zero and such that

$$\|\mathcal{R}_T(X_k + \lambda_k S_k)\| \le (1 - \lambda_k \alpha)\|\mathcal{R}_T(X_k)\|. \tag{3.7}$$

If we define the Newton step residual

$$\mathcal{R}_T'[X_k](S_k) + \mathcal{R}_T(X_k) =: L_{k+1}, \tag{3.8}$$

then equation (3.5) can be written as $\|L_{k+1}\| \le \eta_k \|\mathcal{R}_T(X_k)\|$. Moreover, explicitly writing the terms on the left-hand side in (3.8) yields

$$(D - X_k^T B_1 B_2^T)(X_k + S_k) + (X_k + S_k)^T (A - B_1 B_2^T X_k) + X_k^T B_1 B_2^T X_k + C_1^T C_2 = L_{k+1},$$

so that the matrix $\widetilde{X}_{k+1} := X_k + S_k$ is the solution of the T-Sylvester equation

$$(D - X_k^T B_1 B_2^T)\widetilde{X}_{k+1} + \widetilde{X}_{k+1}^T (A - B_1 B_2^T X_k) = -X_k^T B_1 B_2^T X_k - C_1^T C_2 + L_{k+1}. \tag{3.9}$$

The matrix L_{k+1} is never computed, and the notation in (3.9) is only used to indicate that $\widetilde{X}_{k+1}$ is an inexact solution to the equation (3.1) such that the residual norm $\|L_{k+1}\|$ is sufficiently small. In particular, this means that the iterative routine employed to solve (3.4) must be able to provide us with the residual norm $\|L_{k+1}\|$. When this is sufficiently small, the approximate solution $\widetilde{X}_{k+1}$ is accepted. Once $\widetilde{X}_{k+1}$ is computed, we recover S_k by $S_k = \widetilde{X}_{k+1} - X_k$, and the new iterate can be defined as in (3.6).

The T-Riccati residual at X_{k+1} can be written as

$$\mathcal{R}_T(X_{k+1}) = \mathcal{R}_T(X_k + \lambda_k S_k) = (1 - \lambda_k)\mathcal{R}_T(X_k) + \lambda_k L_{k+1} - \lambda_k^2 S_k^T B_1 B_2^T S_k,$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

and, if $\eta_k \leq \bar{\eta} < 1$ and $\alpha \in (0, 1-\bar{\eta})$, we have

$$\begin{aligned}\|\mathcal{R}_T(X_k + \lambda S_k)\| &\leq (1-\lambda)\|\mathcal{R}_T(X_k)\| + \lambda\|L_{k+1}\| + \lambda^2\|S_k^T B_1 B_2^T S_k\| \\ &\leq (1-\alpha\lambda)\|\mathcal{R}_T(X_k)\|,\end{aligned}$$

for all $\lambda \in (0, (1-\alpha-\bar{\eta})\frac{\|\mathcal{R}_T(X_k)\|}{\|S_k^T B_1 B_2^T S_k\|}]$. In particular, the sufficient decrease condition (3.7) is satisfied for all λ in the latter interval.

For the actual computation of the step size λ_k, we mimic the derivation given in [5, Section 3] for the algebraic Riccati equation, and we exploit the expression of the residual norm $\|\mathcal{R}_T(X_k + \lambda S_k)\|^2$ in terms of a quartic polynomial p_k in λ. In particular,

$$p_k(\lambda) := \|\mathcal{R}_T(X_k + \lambda S_k)\|^2 = (1-\lambda)^2\alpha_k + \lambda^2\beta_k + \lambda^4\delta_k + 2\lambda(1-\lambda)\gamma_k - 2\lambda^2(1-\lambda)\epsilon_k - 2\lambda^3\xi_k,$$

where

$$\begin{aligned}\alpha_k &= \|\mathcal{R}_T(X_k)\|^2, & \beta_k &= \|L_{k+1}\|^2, \\ \gamma_k &= \langle \mathcal{R}_T(X_k), L_{k+1}\rangle_F, & \delta_k &= \|S_k^T B_1 B_2^T S_k\|^2, \\ \epsilon_k &= \langle \mathcal{R}_T(X_k), S_k^T B_1 B_2^T S_k\rangle_F, & \xi_k &= \langle L_{k+1}, S_k^T B_1 B_2^T S_k\rangle_F.\end{aligned} \tag{3.10}$$

The first derivative of $p_k(\lambda)$ is given by

$$p_k'(\lambda) = -2(1-\lambda)\alpha_k + 2\lambda\beta_k + 4\lambda^3\delta_k + 2(1-2\lambda)\gamma_k - 2\lambda(2-3\lambda)\epsilon_k - 6\lambda^2\xi_k,$$

so that

$$\begin{aligned}p_k'(0) &= -2\alpha_k + 2\gamma_k = -2\|\mathcal{R}_T(X_k)\|^2 + 2\langle \mathcal{R}_T(X_k), L_{k+1}\rangle_F + \|L_{k+1}\|^2 - \|L_{k+1}\|^2 \\ &= -\|\mathcal{R}_T(X_k) - L_{k+1}\|^2 - \|\mathcal{R}_T(X_k)\|^2 + \|L_{k+1}\|^2 \leq (\eta_k^2 - 1)\|\mathcal{R}_T(X_k)\|^2 < 0,\end{aligned}$$

as $\eta_k \in (0,1)$, and S_k is thus a descent direction.

The step size λ_k can be computed by exploiting the expression of the T-Riccati residual norm in terms of $p_k(\lambda)$. For $\theta_k := \min\{1, (1-\alpha-\bar{\eta})\sqrt{\alpha_k/\delta_k}\}$, we propose to compute λ_k as

$$\lambda_k := \operatorname*{argmin}_{(0,\theta_k]} p_k(\lambda). \tag{3.11}$$

The choice of the interval $(0, \theta_k]$ is motivated by the fact that if X_k and $\widetilde{X}_{k+1}$ are nonnegative matrices, then also $X_{k+1} = X_k + \lambda_k(\widetilde{X}_{k+1} - X_k)$ is nonnegative. Moreover, the sufficient decrease condition is satisfied for $\lambda_k \in (0, \theta_k]$.

Clearly (3.11) is not the only way to compute λ_k. For instance, in [5, Section 3.2], a step size computation based on the Armijo rule is explored in case of the inexact Newton-Kleinman method applied to the algebraic Riccati equation, and such approach can be adapted to our setting as well. The inexact Newton-Kleinman method with line search is summarized in Algorithm 1.

Notice that we can employ the line search also in the small-scale setting. However, in this case, we are often able to solve the T-Sylvester equations in line 3 of Algorithm 1 with very high accuracy so that $\|L_{k+1}\| = 0$ for all k, and the computation of the step size λ_k simplifies accordingly. Indeed, it is easy to show that the quartic polynomial $p_k(\lambda)$ has a local minimizer in $(0, 2]$ for all k, and we can replace the computation of the step size (3.11) by $\lambda_k := \operatorname{argmin}_{(0,2]} p_k(\lambda)$; see [4] for a line search technique within the *exact* Newton-Kleinman method for algebraic Riccati equations.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 1: Inexact Newton-Kleinman method with line search ($X_0 = 0$).

input : $A, D \in \mathbb{R}^{n\times n}$, $B_1, B_2 \in \mathbb{R}^{n\times p}$, $C_1, C_2 \in \mathbb{R}^{q\times n}$, $\varepsilon > 0$, $\bar{\eta} \in (0,1)$, $\alpha \in (0, 1-\bar{\eta})$.
output : $X_k \in \mathbb{R}^{n\times n}$ approximate solution to (1.1).
for $k = 0, 1, \ldots,$ *till convergence* **do**
  if $\|\mathcal{R}_T(X_k)\| < \varepsilon \cdot \|C_1^T C_2\|$ **then**
1 **Stop** and return X_k
  end
2 Select $\eta_k \in (0, \bar{\eta}]$.
3 Compute $\widetilde{X}_{k+1}$ s.t.

$$(D - X_k^T B_1 B_2^T)\widetilde{X}_{k+1} + \widetilde{X}_{k+1}^T (A - B_1 B_2^T X_k)^T = -X_k^T B_1 B_2^T X_k - C_1^T C_2 + L_{k+1},$$

  where $\|L_{k+1}\| \le \eta_k \|\mathcal{R}_T(X_k)\|$.
4 Set $S_k = \widetilde{X}_{k+1} - X_k$.
5 Compute $\lambda_k > 0$ as in (3.11).
6 Set $X_{k+1} = X_k + \lambda_k S_k$.
end

In the next theorem we show convergence of Algorithm 1 to the minimal solution $X_{\min}$ under some suitable assumptions.

THEOREM 3.4. *Let Assumption 1.1 and Lemma 3.2 hold, and assume that, for all $k \ge 0$, there exists a matrix $\widetilde{X}_{k+1}$ satisfying (3.9), where $\|L_{k+1}\| \le \eta_k \|\mathcal{R}_T(X_k)\|$.*

(i) If the step sizes λ_k are bounded away from zero, $\lambda_k \ge \lambda_{\min} > 0$ for all k, then $\|\mathcal{R}_T(X_k)\| \to 0$.

(ii) If, in addition to (i), the matrices L_{k+1} are nonpositive for all $k \ge 0$ and such that $-\mathcal{R}_T(X_k) + L_{k+1} \ge 0$, then the sequence $\{X_k\}_{k\ge 0}$ generated by the inexact Newton-Kleinman method with $X_0 = 0$ is well-defined and $X_k \le X_{k+1} \le X_{\min}$. Moreover, $\{X_k\}_{k\ge 0}$ converges to the minimal solution $X_{\min}$ of (3.3).

Proof. The sufficient decrease condition (3.7) implies that, for any $\ell \ge 0$,

$$\begin{aligned}\|\mathcal{R}_T(X_0)\| \ge \|\mathcal{R}_T(X_0)\| - \|\mathcal{R}_T(X_{\ell+1})\| &= \sum_{k=0}^{\ell} (\|\mathcal{R}_T(X_k)\| - \|\mathcal{R}_T(X_{k+1})\|) \\ &\ge \sum_{k=0}^{\ell} \lambda_k \alpha \|\mathcal{R}_T(X_k)\| \ge 0.\end{aligned}$$

Taking the limit $\ell \to +\infty$ and using the fact that $\lambda_k \ge \lambda_{\min} > 0$ for all k, we have $\|\mathcal{R}_T(X_k)\| \to 0$.

The proof of (ii) is given by induction on k. For $k = 0$ we have

$$D\widetilde{X}_1 + \widetilde{X}_1^T A = -C_1^T C_2 + L_1, \quad \|L_1\| \le \eta_0 \|C_1^T C_2\|.$$

Since $I \otimes D + (A^T \otimes I)\Pi$ is a nonsingular M-matrix by assumption, and $C_1^T C_2 \le 0$ and $L_1 \ge 0$, it follows that the matrix $\widetilde{X}_1$ is nonnegative. Then $X_1 := \lambda_0 \widetilde{X}_1 \ge 0$ as $\lambda_0 = \operatorname{argmin}_{(0,\theta_0]} p_0(\lambda) > 0$. Moreover, $\mathcal{R}_T(X_0) = C_1^T C_2 \le 0$. Therefore, the properties

$$X_k \le X_{k+1}, \qquad X_k \le X_{\min}, \qquad \mathcal{R}_T(X_k) \le 0, \tag{3.12}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

and

$$I \otimes (D - X_k^T B_1 B_2^T) + ((A - B_1 B_2^T X_k)^T \otimes I)\Pi \quad \text{being a nonsingular M-matrix} \tag{3.13}$$

hold for $k = 0$.

We now assume that the same holds also for a certain $\bar{k} > 0$, and we show properties (3.12), (3.13) for $\bar{k} + 1$. We have

$$\begin{aligned}(D - X_{\bar{k}}^T B_1 B_2^T)\widetilde{X}_{\bar{k}+1} + \widetilde{X}_{\bar{k}+1}^T(A - B_1 B_2^T X_{\bar{k}}) &= -X_{\bar{k}}^T B_1 B_2^T X_{\bar{k}} - C_1^T C_2 + L_{\bar{k}+1},\\ \|L_{\bar{k}+1}\| &\le \eta_{\bar{k}} \|\mathcal{R}_T(X_{\bar{k}})\|,\end{aligned}$$

so that

$$(D - X_{\bar{k}}^T B_1 B_2^T)(\widetilde{X}_{\bar{k}+1} - X_{\bar{k}}) + (\widetilde{X}_{\bar{k}+1} - X_{\bar{k}})^T(A - B_1 B_2^T X_{\bar{k}}) = -\mathcal{R}_T(X_k) + L_{\bar{k}+1}.$$

Since $I \otimes (D - X_{\bar{k}}^T B_1 B_2^T) + ((A - B_1 B_2^T X_{\bar{k}})^T \otimes I)\Pi$ is a nonsingular M-matrix by the induction hypothesis and the right-hand side in the above expression is nonnegative by assumption, we have $\widetilde{X}_{\bar{k}+1} \ge X_{\bar{k}}$. Then, $X_{\bar{k}+1} = (1 - \lambda_{\bar{k}})X_{\bar{k}} + \lambda_{\bar{k}}\widetilde{X}_{\bar{k}+1} \ge X_{\bar{k}}$, since $\lambda_{\bar{k}} \in (0, 1]$.

We now show that $X_{\bar{k}+1} \le X_{\min}$. To this end we can show that $\widetilde{X}_{\bar{k}+1} \le X_{\min}$ since $X_{\bar{k}+1} \le \widetilde{X}_{\bar{k}+1}$. Indeed,

$$X_{\bar{k}+1} = (1 - \lambda_{\bar{k}})X_{\bar{k}} + \lambda_{\bar{k}}\widetilde{X}_{\bar{k}+1} \le (1 - \lambda_{\bar{k}})\widetilde{X}_{\bar{k}+1} + \lambda_{\bar{k}}\widetilde{X}_{\bar{k}+1} = \widetilde{X}_{\bar{k}+1}.$$

We have

$$\begin{aligned}&(D - X_{\bar{k}}^T B_1 B_2^T)(\widetilde{X}_{\bar{k}+1} - X_{\min}) + (\widetilde{X}_{\bar{k}+1} - X_{\min})^T(A - B_1 B_2^T X_{\bar{k}})\\ &\qquad = -DX_{\min} - X_{\min}^T A + X_{\min}^T B_1 B_2^T X_{\bar{k}}\\ &\qquad\quad + X_{\bar{k}}^T B_1 B_2^T X_{\min} - X_{\bar{k}}^T B_1 B_2^T X_{\bar{k}} - C_1^T C_2 + L_{\bar{k}+1},\end{aligned}$$

and by changing the sign, adding and subtracting $X_{\min}^T B_1 B_2^T X_{\min}$ on the right-hand side, we get

$$\begin{aligned}&(D - X_{\bar{k}}^T B_1 B_2^T)(X_{\min} - \widetilde{X}_{\bar{k}+1}) + (X_{\min} - \widetilde{X}_{\bar{k}+1})^T(A - B_1 B_2^T X_{\bar{k}})\\ &\qquad = (X_{\min} - X_{\bar{k}})^T B_1 B_2^T (X_{\min} - X_{\bar{k}}) - L_{k+1},\end{aligned}$$

where we used the fact that $\mathcal{R}_T(X_{\min}) = 0$. Since $X_{\min} \ge X_{\bar{k}}$ by the induction hypothesis and $B_1 B_2^T \ge 0$ and $L_{k+1} \le 0$, it follows that the right-hand side in the above equation is nonnegative so that $\widetilde{X}_{\bar{k}+1} \le X_{\min}$ thanks to the fact that the induction hypothesis states that $I \otimes (D - X_{\bar{k}}^T B_1 B_2^T) + ((A - B_1 B_2^T X_{\bar{k}})^T \otimes I)\Pi$ is a nonsingular M-matrix.

To show that $I \otimes (D - X_{\bar{k}+1}^T B_1 B_2^T) + ((A - B_1 B_2^T X_{\bar{k}+1})^T \otimes I)\Pi$ is a nonsingular M-matrix, we can use the same argument as in the proof of Theorem 2.1, as $X_{\bar{k}+1} \le X_{\min}$.

The last statement we have to show is $\mathcal{R}_T(X_{\bar{k}+1}) \le 0$. We can write

$$\begin{aligned}\mathcal{R}_T(X_{\bar{k}+1}) = (D - X_{\bar{k}+1}^T B_1 B_2^T)(X_{\bar{k}+1} - X_{\min}) + (X_{\bar{k}+1} - X_{\min})^T(A - B_1 B_2^T X_{\bar{k}+1})\\ - DX_{\min} - X_{\min}^T A - + X_{\min}^T B_1 B_2^T X_{\bar{k}+1} + X_{\bar{k}+1}^T B_1 B_2^T X_{\min}\\ - X_{\bar{k}+1}^T B_1 B_2^T X_{\bar{k}+1} - C_1^T C_2.\end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Since $X_{\bar{k}+1} - X_{\min} \leq 0$ and $I \otimes (D - X_{\bar{k}+1}^T B_1 B_2^T) + ((A - B_1 B_2^T X_{\bar{k}+1})^T \otimes I)\Pi$ is a nonsingular M-matrix, $(D - X_{\bar{k}+1}^T B_1 B_2^T)(X_{\bar{k}+1} - X_{\min}) + (X_{\bar{k}+1} - X_{\min})^T (A - B_1 B_2^T X_{\bar{k}+1}) \leq 0$, and we have

$$\begin{aligned}\mathcal{R}_T(X_{\bar{k}+1}) &\leq -DX_{\min} - X_{\min}^T A + X_{\min}^T B_1 B_2^T X_{\bar{k}+1} + X_{\bar{k}+1}^T B_1 B_2^T X_{\min} \\ &\quad - X_{\bar{k}+1}^T B_1 B_2^T X_{\bar{k}+1} - C_1^T C_2 \\ &= -(X_{\min} - X_{\bar{k}+1})^T B_1 B_2^T (X_{\min} - X_{\bar{k}+1}) \leq 0,\end{aligned}$$

as $X_{\min} \geq X_{\bar{k}+1}$ and $B_1 B_2^T \geq 0$.

In conclusion, the sequence $\{X_k\}_{k\geq 0}$ computed by the inexact Newton-Kleinman method with $X_0 = 0$ and equipped with the line search (3.11) is well-defined, nondecreasing, and bounded from above. Therefore, $\{X_k\}_{k\geq 0}$ has a finite limit X_* that is also a solution of the T-Riccati equation since

$$0 = \lim_{k\to+\infty} \|\mathcal{R}_T(X_k)\| = \|\mathcal{R}_T(\lim_{k\to+\infty} X_k)\| = \|\mathcal{R}_T(X_*)\|.$$

Moreover, it is easy to show that $X_* \leq H$ for every nonnegative solution H to (1.1). Indeed, by using the same argument that we employed to prove $X_k \leq X_{\min}$ for all k, we can show that $X_k \leq H$ for all k. Since the inequality is preserved at the limit, we have $X_* \leq H$, hence $X_* = X_{\min}$. □

The assumption on the sign of L_{k+1} may remind the reader of the hypothesis made in [13] for proving convergence of the inexact Newton-Kleinman method applied to the standard algebraic Riccati equation. Indeed, in [13, Theorem 4.4], the matrix L_{k+1} is supposed to be positive semidefinite for all k[2]. However, as outlined in [5], this condition is hard to meet in practice, and in [5, Theorem 10] a different approach is used for showing convergence of the inexact Newton scheme. In our setting we do not see any particular drawback in assuming L_{k+1} nonpositive for every k. Moreover, if the projection method presented in [12] is employed for the computation of $\widetilde{X}_{k+1}$, then the nonpositivity of L_{k+1} may be further explored by exploiting the explicit form of this residual matrix given in [12, Proposition 4.3]. However, this is beyond the scope of this paper. We also point out that the assumption $-\mathcal{R}_T(X_k) + L_{k+1} \geq 0$ is somehow consistent with having $\|L_{k+1}\| \leq \eta_k \|\mathcal{R}_T(X_k)\|$. Indeed, both $-\mathcal{R}_T(X_k)$ and $-L_{k+1}$ are nonnegative matrices so that $-\mathcal{R}_T(X_k) + L_{k+1} \geq 0$ implies $\|L_{k+1}\| \leq \|\mathcal{R}_T(X_k)\|$.

To conclude, we would like to mention that the assumptions on L_{k+1} in Theorem 3.4 are automatically satisfied if $\|L_{k+1}\| = 0$, for all $k \geq 0$, as in the small-scale setting. Therefore, Theorem 3.4 shows convergence to the minimal nonnegative solution of the *exact* Newton-Kleinman method equipped with a line search. The latter procedure may improve the convergence rate of the exact Newton-Kleinman method, especially for the first iterations, as shown in [4] for the standard algebraic Riccati equation; see Example 4.2 in Section 4.

3.3. Implementation details. In this section, we present some details for an efficient implementation of Algorithm 1 in case of high-dimensional problems.

First of all, we recall that the computation of the Frobenius norm of low-rank matrices does not need to assemble any $n \times n$ dense matrix. For instance, only $q \times q$ matrices are actually involved in the computation of $\|C_1^T C_2\|$ as

$$\|C_1^T C_2\|^2 = \text{trace}(C_2^T C_1 C_1^T C_2) = \text{trace}((C_1 C_1^T)(C_2 C_2^T)).$$

[2]Notice that in [13] the authors were looking for a maximal solution so that they required L_{k+1} to be positive semidefinite. Here we are computing the minimal solution, and we thus ask L_{k+1} to be nonpositive.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The most expensive part of Algorithm 1 is the solution of the large-scale T-Sylvester equations in line 3. These equations can be solved, e.g., by employing the projection method presented in [12]. Given the T-Sylvester equation

$$DX + X^T A = -C_1^T C_2,$$

an approximate solution $X_m \in \mathbb{R}^{n\times n}$ of the form $X_m = V_m Y_m W_m^T \approx X$ is constructed, where the orthonormal columns of $V_m, W_m \in \mathbb{R}^{n\times \ell}$ span suitable subspaces $\mathcal{K}_{V_m}$ and $\mathcal{K}_{W_m}$, respectively, i.e., $\mathcal{K}_{V_m} = \text{Range}(V_m)$ and $\mathcal{K}_{W_m} = \text{Range}(W_m)$. We will always assume that V_m and W_m have full rank so that $\dim(\mathcal{K}_{V_m}) = \dim(\mathcal{K}_{W_m}) = \ell$. If this is not the case, then deflation strategies as the ones presented in [16] can be implemented to overcome the possible linear dependence of the spanning vectors. The $\ell \times \ell$ matrix Y_m is computed by imposing a Petrov-Galerkin condition on the residual matrix $R_m = DX_m + X_m^T A + C_1^T C_2$ with respect to the space $\mathcal{K}_{W_m} \otimes \mathcal{K}_{W_m}$. This condition is equivalent to computing Y_m by solving the reduced T-Sylvester equation

$$(W_m^T D V_m) Y_m + Y_m^T (V_m^T A W_m) = -(W_m^T C_1^T)(C_2 W_m); \tag{3.14}$$

see [12, Section 3]. Equation (3.14) can be solved by employing, e.g., Algorithm 3.1 presented in [11], as the small dimension of the coefficient matrices allows for the computation of the generalized Schur decomposition of the pair $(W_m^T D V_m, (V_m^T A W_m)^T)$.

The effectiveness of the projection framework presented in [12] is strongly related to the choice of the approximation spaces $\mathcal{K}_{V_m}$ and $\mathcal{K}_{W_m}$. In [12] it is shown how the selection of these spaces may depend on the location of the spectrum $\Lambda(A^{-T}D)$ of $A^{-T}D$. In particular, if $\Lambda(A^{-T}D)$ is strictly contained in the unit disk, it is suggested to select

$$\begin{aligned} \mathcal{K}_{V_m} &= \mathbf{K}_m^{\square}(A^{-T}D, A^{-T}[C_1^T, C_2^T]), \qquad \text{and} \\ \mathcal{K}_{W_m} &= A^T \cdot \mathcal{K}_{V_m} = \mathbf{K}_m^{\square}(DA^{-T}, [C_1^T, C_2^T]), \end{aligned}$$

where

$$\begin{aligned} &\mathbf{K}_m^{\square}(A^{-T}D, A^{-T}[C_1^T, C_2^T]) \\ &\quad = \text{Range}\big([A^{-T}[C_1^T, C_2^T], A^{-T}DA^{-T}[C_1^T, C_2^T], \ldots, (A^{-T}D)^{m-1}A^{-T}[C_1^T, C_2^T]]\big) \end{aligned}$$

is the block Krylov subspace generated by $A^{-T}D$ and $A^{-T}[C_1^T, C_2^T]$. On the other hand, if $\Lambda(A^{-T}D)$ is well outside the unit disk, then the roles of A and D are reversed, and we can choose

$$\begin{aligned} \mathcal{K}_{V_m} &= \mathbf{K}_m^{\square}(D^{-1}A^T, D^{-1}[C_1^T, C_2^T]), \quad \text{and} \\ \mathcal{K}_{W_m} &= D \cdot \mathcal{K}_{V_m} = \mathbf{K}_m^{\square}(A^T D^{-1}, [C_1^T, C_2^T]). \end{aligned}$$

However, in general, the spectrum of $A^{-T}D$ is neither strictly contained in the unit disk nor well outside it, and the employment of the extended Krylov subspaces

$$\begin{aligned} \mathcal{K}_{V_m} &= \mathbf{EK}_m^{\square}(A^{-T}D, A^{-T}[C_1^T, C_2^T]), \quad \text{and} \\ \mathcal{K}_{W_m} &= A^T \cdot \mathbf{EK}_m^{\square}(A^{-T}D, A^{-T}[C_1^T, C_2^T]), \end{aligned} \tag{3.15}$$

where

$$\begin{aligned} \mathbf{EK}_m^{\square}(A^{-T}D, A^{-T}[C_1^T, C_2^T]) := \mathbf{K}_m^{\square}(A^{-T}D, A^{-T}[C_1^T, C_2^T]) \\ + \mathbf{K}_m^{\square}(D^{-1}A^T, D^{-1}[C_1^T, C_2^T]), \end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

is recommended in this case. It has been shown that the projection method based on the extended Krylov subspaces (3.15) performs quite well in most of the results reported in [12, Section 7], and if this procedure fails to converge, then also the projection schemes based on the block Krylov subspaces above fail as well. Therefore, we also adopt the extended Krylov subspaces (3.15) as approximation spaces in the solution of the sequence of T-Sylvester equations (3.4) arising from the inexact Newton-Kleinman scheme.

The coefficient matrix defining the equations in (3.4) are of the form $D - X_k^T B_1 B_2^T$ and $A - B_1 B_2^T X_k$ so that the spaces

$$\mathbf{EK}_m^{\square}((A - B_1 B_2^T X_k)^{-T}(D - X_k^T B_1 B_2^T), (A - B_1 B_2^T X_k)^{-T}[C_1^T, C_2^T, X_k^T B_1, X_k^T B_2]),$$

and

$$\begin{aligned}&(A - B_1 B_2^T X_k)^T \cdot \mathbf{EK}_m^{\square}((A - B_1 B_2^T X_k)^{-T}(D - X_k^T B_1 B_2^T),\\ &(A - B_1 B_2^T X_k)^{-T}[C_1^T, C_2^T, X_k^T B_1, X_k^T B_2]),\end{aligned}$$

have to be computed at each Newton step $k \geq 0$. Such constructions require to solve linear systems of the form $(A + MN^T)z = y$, where $M, N \in \mathbb{R}^{n \times p}$ are low-rank, and the Sherman-Morrison-Woodbury (SMW) formula

$$(A - MN^T)^{-1} = A^{-1} + A^{-1}M(I - N^T A^{-1} M)^{-1} N^T A^{-1},$$

can be employed to this end; see, e.g., [14, Equation (2.1.4)].

Algorithm 2 summarizes the projection method for the solution of the $(k+1)$-st T-Sylvester equation (3.4), where we suppose that the k-th iterate X_k is given in low-rank format, namely $X_k = P_{1,k} P_{2,k}^T$, $P_k \in \mathbb{R}^{n \times t_k}$, $t_k \ll n$.

To compute the residual norm $\|L_{k+1}\|$ we do not need to construct the dense $n \times n$ residual matrix

$$\begin{aligned}L_{k+1} = &(D - P_{2,k}\boldsymbol{\alpha} B_2^T)(V_m Y_m W_m^T)\\ &+ (V_m Y_m W_m^T)^T (A - B_1 \boldsymbol{\beta}^T P_{2,k}) + P_{2,k}\boldsymbol{\alpha}\boldsymbol{\beta}^T P_{2,k} + C_1^T C_2.\end{aligned}$$

Indeed, it is easy to show that

$$\|L_{k+1}\| = \|\boldsymbol{\tau}_{m+1,m}(e_m^T \otimes I_{4(p+q)}) Y_m\|,$$

where $\boldsymbol{\tau}_{m+1,m} := W_{m+1}^T (D - P_{2,k}\boldsymbol{\alpha} B_2^T) V_m$ and $e_m \in \mathbb{R}^m$ is the m-th canonical basis vector of $\mathbb{R}^m$; see [12, Proposition 5.1]. At the k-th iteration of the Newton-Kleinman scheme, we can set $\varepsilon = \eta_k \|\mathcal{R}_T(X_k)\|$ as inner tolerance for Algorithm 2.

The computation of the coefficients in (3.10) needed for calculating the step size λ_k can be carried out at low cost. Indeed, even if it is not evident, all the quantities in (3.10) consist of inner products with low-rank matrices, and they are thus cheap to evaluate as recalled at the beginning of this section. In particular, if $X_k = P_{1,k} P_{2,k}^T$ is the k-th iterate of the Newton-Kleinman scheme and $\widetilde{X}_{k+1} = \widetilde{P}_{1,k+1} \widetilde{P}_{2,k+1}^T$ is the matrix computed by Algorithm 2, then we can write

$$\begin{aligned}\mathcal{R}_T(X_k) &= DX_k + X_k^T A - X_k^T B_1 B_2^T X_k + C_1 C_2^T\\ &= [DP_{1,k}, P_{2,k}, -P_{2,k}(P_{1,k}^T B_1), C_1^T][P_{2,k}, A^T P_{1,k}, P_{2,k}(P_{1,k}^T B_2), C_2^T]^T,\\ L_{k+1} &= (D - X_k^T B_1 B_2^T)\widetilde{X}_{k+1} + \widetilde{X}_{k+1}^T (A - B_1 B_2^T X_k) + X_k^T B_1 B_2^T X_k + C_1^T C_2\\ &= [D\widetilde{P}_{1,k+1}, -P_{2,k}\boldsymbol{\alpha}\widetilde{\boldsymbol{\beta}}^T, \widetilde{P}_{2,k+1}, \widetilde{P}_{2,k+1}, \widetilde{P}_{2,k+1}\widetilde{\boldsymbol{\alpha}}^T, C_1^T]\cdot\\ &\quad [\widetilde{P}_{2,k+1}, \widetilde{P}_{2,k+1}, A^T\widetilde{P}_{1,k+1}, -P_{2,k}(\boldsymbol{\beta}\widetilde{\boldsymbol{\alpha}}^T), \widetilde{P}_{2,k+1}\widetilde{\boldsymbol{\beta}}^T, C_2^T]^T,\\ S_k &= \widetilde{X}_{k+1} - X_k = [\widetilde{P}_{1,k+1}, -P_{1,k}][\widetilde{P}_{2,k+1}, P_{2,k}]^T,\end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 2: Extended Krylov subspace method for T-Sylvester equations.

input : $A, D \in \mathbb{R}^{n\times n}$, $B_1, B_2 \in \mathbb{R}^{n\times p}$, $C_1, C_2 \in \mathbb{R}^{q\times n}$, $P_{1,k}, P_{2,k} \in \mathbb{R}^{n\times t_k}$, $t_k \ll n$, $\varepsilon > 0$, $m_{\max} > 0$

output : $\widetilde{P}_{1,k+1}, \widetilde{P}_{2,k+1} \in \mathbb{R}^{n\times t_{k+1}}$, $t_{k+1} \ll n$, s.t. $\widetilde{X}_{k+1} = \widetilde{P}_{1,k+1}\widetilde{P}_{2,k+1}^T$ is an approximate solution to (3.4)

Set $\boldsymbol{\alpha} = P_{1,k}^T B_1$ and $\boldsymbol{\beta} = P_{1,k}^T B_2$.
Set $H = [C_1^T, C_2^T, P_{2,k}\boldsymbol{\alpha}, P_{2,k}\boldsymbol{\beta}]$.
Perform economy-size QR, $[H, (A - B_1(B_2^T P_k)P_k^T)^{-1}H] = [\mathcal{V}_1^{(1)}, \mathcal{V}_1^{(2)}]\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}, \in \mathbb{R}^{4(q+p)\times 4(q+p)}$.
Set $V_1 = [\mathcal{V}_1^{(1)}, \mathcal{V}_1^{(2)}]$.
$W_1 \leftarrow$ orthonormalize the columns of $(A - B_1\boldsymbol{\beta}^T P_{2,k}^T)^T V_1$.
for $m = 1, 2, \ldots,$ *till* $m_{\max}$ **do**
  Compute next basis block $\mathcal{V}_{m+1}$ as in [12] and set $V_{m+1} = [V_m, \mathcal{V}_{m+1}]$.
  $\mathcal{W}_{m+1} \leftarrow$ orthonormalize the columns of $(A - B_1\boldsymbol{\beta}^T P_{1,k}^T)^T\mathcal{V}_{m+1}$ w.r.t. W_m.
  Set $W_{m+1} = [W_m, \mathcal{W}_{m+1}]$.
  Update $T_m = W_m^T(D - P_{2,k}\boldsymbol{\alpha}B_2^T)V_m$, $K_m = V_m^T(A - B_1\boldsymbol{\beta}^T P_{2,k}^T)W_m$ as in [12].
  Update $G_1 = W_m^T[C_1^T, P_{2,k}\boldsymbol{\alpha}]$ and $G_2 = W_m^T[C_2^T, P_{2,k}\boldsymbol{\beta}^T]$.
  Solve $T_m Y_m + Y_m^T K_m = -G_1 G_2^T$.
  if $\|L_{k+1}\| = \|(D - P_{2,k}\boldsymbol{\alpha}B_2^T)(V_m Y_m W_m^T) + (V_m Y_m W_m^T)^T(A - B_1\boldsymbol{\beta}^T P_{2,k}) + P_{2,k}\boldsymbol{\alpha}\boldsymbol{\beta}^T P_{2,k} + C_1^T C_2\| \le \varepsilon$ **then**
    Break and go to **13**.
  end
end
Factorize Y_m, and retain $\widehat{Y}_{1,m}, \widehat{Y}_{2,m} \in \mathbb{R}^{4m(q+p)\times t_{k+1}}$, $t_{k+1} \le 4m(q+p)$, $\widehat{Y}_{1,m}\widehat{Y}_{2,m}^T \approx Y_m$.
Set $\widetilde{P}_{1,k+1} = V_m\widehat{Y}_{1,m}$, $\widetilde{P}_{2,k+1} = W_m\widehat{Y}_{2,m}$.

where $\boldsymbol{\alpha} = P_{1,k}^T B_1$, $\boldsymbol{\beta} = P_{1,k}^T B_2$, $\widetilde{\boldsymbol{\alpha}} = \widetilde{P}_{1,k+1}^T B_1$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{P}_{1,k+1}^T B_2$.

4. Numerical examples. In this section we report some results regarding the numerical solution of the nonsymmetric T-Riccati equation (1.1). Different instances of (1.1) are considered, and both the small-scale and the large-scale scenario are addressed.

When n is moderate, the T-Sylvester equations arising from the Newton-Kleinman scheme (3.1) are solved by means of Algorithm 3.1 presented in [11]. We show that even when equations (3.1) are solved exactly, a line search can improve the convergence rate of the Newton-Kleinman scheme by maintaining a monotone decrease in the residual norm. We always set the threshold for the relative residual norm to be equal to 10^{-12} for small n. Moreover, we report the number of iterations, i.e., the number of T-Sylvester equations solved, to achieve such accuracy, as well as the final relative residual norm and the overall computational time in seconds.

For problems of large dimensions, the inexact Newton-Kleinman method is employed in the solution of (1.1) together with Algorithm 2 as inner solver. The tolerance for the outer relative residual norm achieved by the Newton scheme is set to 10^{-6} while the one for the inner solver changes as the iterations proceed accordingly to the discussion in Section 3.3, where $\eta_k = 1/(1+k^3)$. Such a value of η_k has been proposed, e.g., in [5] in the context of inexact

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Newton methods for algebraic Riccati equations, and it leads to a superlinear convergence of the Newton procedure.

Also in the large-scale setting we report the total number of T-Sylvester equations that need to be solved to get the desired accuracy, along with the average number of inner iterations, the final relative residual norm, and the computational time for solving the problem. Moreover, since the memory requirements are one of the main issue in the numerical solution of large-scale matrix equations, we also document the storage demand of the solution process, which corresponds to the dimension of the largest spaces (3.15) constructed. The rank of the final numerical solution is reported to show that, at least in the tested examples, a low-rank approximate solution to (1.1) can be sought. All results were obtained with MATLAB R2017b [20] on a Dell machine with two 2GHz processors and 128 GB of RAM.

We would like to mention that, while in Example 4.1, Assumption 1.1 is fulfilled, the coefficient matrices considered in Example 4.2 and 4.3 do not necessarily give rise to a nonsingular M-matrix $I \otimes D + (A^T \otimes I)\Pi$. Nevertheless, we are still able to compute an approximate solution that provides a small relative residual norm. This proof of concept numerically shows that the class of T-Riccati equations which admits a solution may be larger than the one considered in this paper, and further studies are necessary.

EXAMPLE 4.1. In the first example we consider a slight modification of Example 1 in [6]. In particular, we define the $n \times n$ matrices

$$D = \begin{bmatrix} 4 & -1 & & & \\ & 4 & -1 & & \\ & & \ddots & \ddots & \\ & & & 4 & -1 \\ & & & & 4 \end{bmatrix}, \quad A = \begin{bmatrix} -1 & -1 & & & \\ & -1 & -1 & & \\ & & \ddots & \ddots & \\ & & & -1 & -1 \\ & & & & -1 \end{bmatrix}, \quad E = \begin{bmatrix} -1 & -1 & & & \\ & -1 & -1 & & \\ & & \ddots & \ddots & \\ & & & -1 & -1 \\ & & & & -0.9 \end{bmatrix},$$

together with $B = -A/\|A\|$, and $C = E/\|E\|$.

We consider moderate values of n to be able to verify that the conditions in Assumption 1.1 are fulfilled. Clearly, $B \geq 0$ and $C \leq 0$ for all n. Moreover, for all the tested values of n, the matrix $I \otimes D + (A^T \otimes I)\Pi$ is a nonsingular M-matrix. Indeed, such a matrix is a Z-matrix as all its off-diagonal entries are nonpositive. Moreover, its spectrum is contained in the right half open plane[3], and $I \otimes D + (A^T \otimes I)\Pi$ is thus a nonsingular M-matrix.

In Figure 4.1 (left) we report the results for different values of n along with the smallest eigenvalue of the corresponding matrix $I \otimes D + (A^T \otimes I)\Pi$. In particular, we document only the results obtained by the Newton method without line search since, for this example, the latter technique does not significantly improve the converge behavior of the Newton scheme.

As predicted by the analysis in the previous sections, the Newton method is able to compute a nonnegative solution for all tested n. In Figure 4.1 (right) we display the computed solution in a logarithmic scale for the case $n = 100$.

EXAMPLE 4.2. In this example, we consider the same coefficient matrices as in [12, Numerical test 7.1]. In particular, the matrices $D, A \in \mathbb{R}^{n \times n}$ come from the finite difference discretization on the unit square of the 2-dimensional differential operators

$$\mathcal{L}_D(u) = -u_{xx} - u_{yy} + y(1-x)u_x + \gamma u, \quad \text{and} \quad \mathcal{L}_A(u) = -u_{xx} - u_{yy},$$

respectively, with $\gamma = 10^4$. Homogeneous Dirichlet boundary conditions are considered.

We first tackle the case of moderate problem dimensions and choose $B, C \in \mathbb{R}^{n \times n}$ to be full random matrices. In Table 4.1 we report the results for different n.

[3] This is a necessary and sufficient condition for a real Z-matrix to be a nonsingular M-matrix; see, e.g., [8].

ETNA
Kent State University and
Johann Radon Institute (RICAM)

n	Its	Rel. Res	Time (secs)	$\lambda_{\min}$
100	3	5.08e-13	0.42	1+1e-3
300	3	1.42e-14	2.95	1+1e-4
500	3	1.88e-14	9.41	1+4e-5

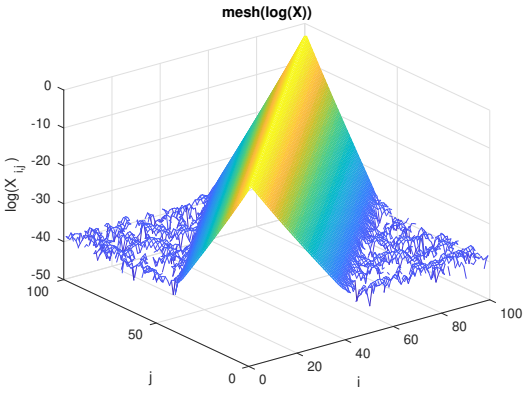

FIG. 4.1. *Example 4.1. Left: Results for different values of (moderate) n. Right: Computed solution in a logarithmic scale for $n = 100$.*

TABLE 4.1
Example 4.2. Results for different values of (moderate) n.

	n	Its	Rel. Res	Time (secs)
w/o line search	324	8	8.51e-15	11.28
w/ line search		5	2.99e-14	7.54
w/o line search	784	10	8.62e-14	99.94
w/ line search		8	2.32e-14	73.73

For this example, the exact line search discussed at the end of Section 3.2 is effective in decreasing the number of iterations necessary to achieve the prescribed accuracy leading to a speed-up of the solution process. In particular, a small step size λ_1 is computed at the first iteration avoiding an increment in the relative residual norm and allowing us to faster reach the region where quadratic convergence occurs. This is apparent from Figure 4.2, where the relative residual norms produced by the Newton-Kleinman method with and without line search are plotted for the case $n = 784$. We can appreciate how a monotone decrease in the relative residual norm is obtained if the line search is performed.

In the large-scale setting, we consider low-rank matrices $B = B_1 B_2^T$, $B_1, B_2 \in \mathbb{R}^{n\times p}$, and $C = C_1 C_2^T$, $C_1, C_2 \in \mathbb{R}^{n\times q}$, such that B_i, C_i have unit norm and random entries for $i = 1, 2$. The matrices A and D are as before. In Table 4.2 we report the results for different values of p, q, and n.

TABLE 4.2
Example 4.2. Results for different values of p, q and n.

n	p	q	Its (inner)	Mem.	Rank(X)	Rel. Res.	Time (secs)
	1	1	13 (6.46)	160	28	8.33e-7	15.65
10,000	1	5	6 (6.66)	624	87	5.14e-7	52.15
	5	10	6 (6.00)	1,560	186	4.39e-7	110.12
	1	1	15 (10.60)	352	26	5.18e-7	69.19
22,500	1	5	convergence not achieved				
	5	10	convergence not achieved				
	1	1	convergence not achieved				
32,400	1	5	convergence not achieved				
	5	10	convergence not achieved				

ETNA
Kent State University and
Johann Radon Institute (RICAM)

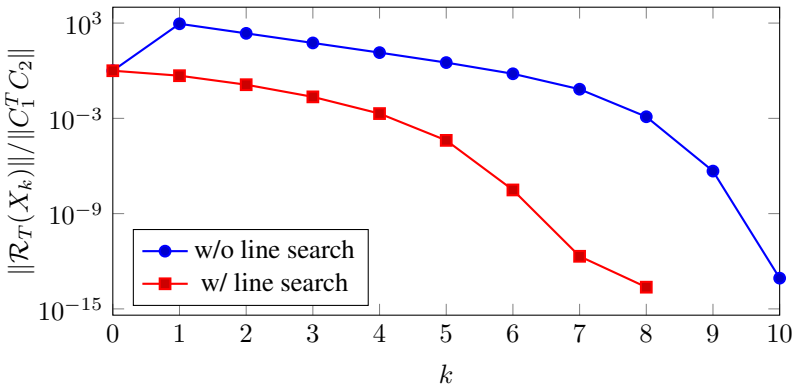

FIG. 4.2. *Example 4.2. Relative residual norms produced by the Newton-Kleinman with and without line search for $n = 784$.*

We notice that for the largest values of n, the inexact Newton-Kleinman method does not always achieve the desired accuracy in terms of the relative residual norm. Indeed, for a certain $k > 0$, Algorithm 2 does not manage to solve the k-th equation (3.1) of the Newton-Kleinman scheme[4], and we thus stop the process. In Figure 4.3 (left) we plot in a logarithmic scale the T-Riccati relative residual norm for the case $n = 22,500$ and $p = 1$, $q = 5$. For this example, the residual norm decreases (non monotonically) until Algorithm 2 is no longer able to solve the eighth T-Sylvester equation

$$(D - X_7^T B_1 B_2^T)\widetilde{X}_8 + \widetilde{X}_8^T (A - B_1 B_2^T X_7)^T = -X_7^T B_1 B_2^T X_7 - C_1^T C_2. \tag{4.1}$$

In particular, in Figure 4.3 (right), the relative residual norm (solid line) produced by Algorithm 2 when applied to equation (4.1) is reported. We can appreciate how the residual norm smoothly decreases in the first 18 iterations, and, after an erratic phase, it starts increasing until the 35th iteration when we stop the procedure. In Figure 4.3 (right) we also plot the threshold (dashed line) passed to Algorithm 2, i.e., $\eta_7 \cdot \|\mathcal{R}_T(X_7)\|$, and we can realize how the relative residual norm gets very close to the desired accuracy without reaching it. A similar behavior has been observed also for the other tests where convergence has not been achieved. We think it may be interesting to further study the convergence property of Algorithm 2 as also the solution of the T-Riccati equation (1.1) can benefit from this.

When the desired accuracy is achieved, the rank of the computed numerical solution X is rather small compared to the problem size n for all the tested values of p and q. This suggests that it may be reasonable to investigate in depth the trend of the singular values of the exact solution to (1.1) in order to justify the search for low-rank approximate solutions and the development of low-rank numerical schemes.

EXAMPLE 4.3. The last example we consider consists in a slight modification of [15, Example 6.1]. In the small-scale setting we generate a random matrix $R = \texttt{rand}(2n, 2n) \in \mathbb{R}^{2n\times 2n}$ and define $W = \text{diag}(R\mathbf{1}) - R$, where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^{2n}$. Then, $A, D \in \mathbb{R}^{n\times n}$ are chosen according to the partition

$$W = \begin{bmatrix} D & M \\ N & A \end{bmatrix},$$

and $B = -N/\|N\|$. We also define an $n \times n$ matrix X_* with random entries and unit norm, and we compute $C = DX_* + X_*^T A - X_*^T B X_*$.

[4]Some examples where Algorithm 2 does not converge are reported also in [12].

ETNA
Kent State University and
Johann Radon Institute (RICAM)

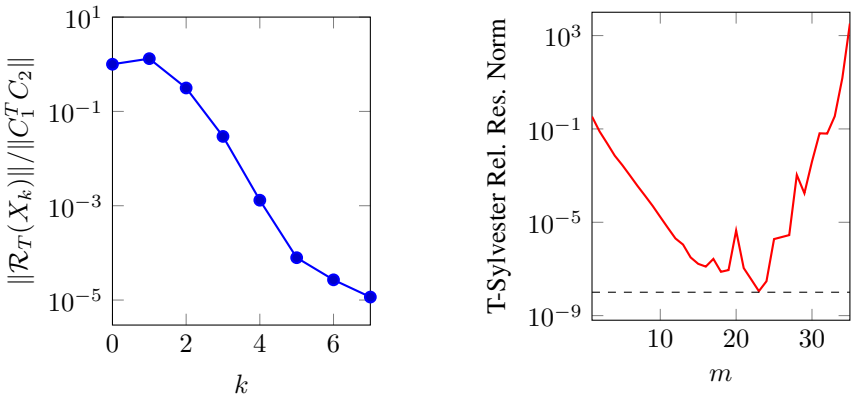

FIG. 4.3. *Example 4.2, $n = 22,500$, $p = 1$, and $q = 5$. Left: $\|\mathcal{R}_T(X_k)\|/\|C_1^T C_2\|$ for $k = 0, \ldots, 7$. Right: relative residual norm produced by Algorithm 2 when applied to equation (4.1) (solid line) and $\eta_7 \cdot \|\mathcal{R}_T(X_7)\|$ (dashed line).*

The results for different n are collected in Table 4.3, where we also report the relative error between the computed solution and X_*.

TABLE 4.3
Example 4.3. Results for different values of (moderate) n.

	n	Its	Rel. Res.	Err. Rel.	Time (secs)
w/o line search	500	3	1.06e-14	7.78e-11	10.80
w/ line search		3	3.48e-13	6.01e-10	10.84
w/o line search	1,000	3	1.49e-14	9.33e-10	78.59
w/ line search		3	1.78e-13	1.45e-9	78.99

The Newton-Kleinman method with line search performs in a very similar manner with respect to the case where no line search is used. Indeed, in this example, the computed step size λ_k is always close to one, for every k.

For the large-scale setting we have to construct the coefficient matrices in a different way to be able to allocate them. To this end we compute two sparse matrices $F, G \in \mathbb{R}^{n \times n}$ with random entries via the MATLAB function `sprand`[5], and we shift them to ensure their nonsingularity. We thus define $D = F + (\rho(F) + 1)I$ and $A = G + (\rho(G) + 20)I$. As in Example 4.2, we consider low-rank matrices $B = B_1 B_2^T$, $B_1, B_2 \in \mathbb{R}^{n \times p}$, and $C = C_1 C_2^T$, $C_1, C_2 \in \mathbb{R}^{n \times q}$ such that B_i, C_i have unit norm and random entries for $i = 1, 2$.

In Table 4.4 we report the results for different values of p, q, and n.

In this example, we manage to reach the desired accuracy for every value of p, q, and n that we tested. Moreover, the numerical solution turns out to be low-rank in all the experiments we ran.

We notice that the computational timings in Table 4.4 are several orders of magnitude smaller than the ones reported in Table 4.2 even when the problem dimension, the rank of B and C, and the number of outer iterations are very similar. This is mainly due to the following factors. The average numbers of inner iterations in Table 4.2 is larger than the ones reported in Table 4.4. Therefore, even if we solve a similar number of T-Sylvester equations to converge, the ones in Example 4.2 require a larger space to be solved leading to an increment in both the

[5]The density of the nonzero entries is set to be equal to $1/n$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 4.4
Example 4.3. Results for different values of p, q, and n.

n	p	q	Its (inner)	Mem.	Rank(X)	Rel. Res.	Time (secs)
	1	1	4 (1.5)	32	4	6.19e-7	0.16
10,000	1	5	5 (1.8)	144	29	1.18e-8	1.11
	5	10	5 (1.8)	360	60	2.35e-9	3.27
	1	1	4 (1.5)	32	4	6.48e-7	0.79
50,000	1	5	5 (1.8)	144	29	1.18e-8	5.44
	5	10	5 (1.8)	360	60	1.19e-9	14.88
	1	1	4 (1.5)	32	4	6.30e-9	1.48
100,000	1	5	5 (1.8)	144	28	1.80e-8	11.33
	5	10	5 (1.8)	360	60	4.71e-10	24.49

memory allocation and the computational efforts. As presented in [12] and outlined in Section 3.3, the effectiveness of Algorithm 2 is related to the location of the spectrum of the matrix $(A - B_1B_2^TX_k)^{-T}(D - X_k^TB_1B_2^T)$, where $D - X_k^TB_1B_2^T$ and $A - B_1B_2^TX_k$ are the coefficient matrices defining the k-th T-Sylvester equation (3.4) in the Newton sequence. In particular, Algorithm 2 performs better whenever this spectrum is well inside/outside the unit circle; see [12]. When we consider the data from Example 4.2 for $n = 10,000$, $p = 1$, and $q = 5$, it turns out that the largest spectral interval of the matrices $(A - B_1B_2^TX_k)^{-T}(D - X_k^TB_1B_2^T)$, $k = 1, \ldots, 6$, is $[1.12, 507.65]$. For the data in Example 4.3 with the same values of n, p, and q, such a spectral interval is given by $[0.06, 0.12]$. Therefore, when we deal with Example 4.3, the eigenvalues of $(A - B_1B_2^TX_k)^{-T}(D - X_k^TB_1B_2^T)$ are highly clustered inside the unit circle, while the ones related to Example 4.2 belong to a larger spectral interval that is rather close to the boundary of the unit circle.

Moreover, the different level of fill in of the coefficient matrices leads to more computationally intensive operations when A and D from Example 4.2 are manipulated. Indeed, for $n = 10,000$, the number of nonzero entries of A and D in Example 4.2 is approximately 50,000 while in Example 4.3 it is 20,000. Similar results are obtained for the other values of n that we tested.

5. Conclusions. By taking inspiration from the rich literature about the algebraic Riccati equation, in this paper we investigated some theoretical and computational aspects of the nonsymmetric T-Riccati equation. Sufficient conditions for the existence and uniqueness of a minimal nonnegative solution $X_{\min}$ have been provided. We have thoroughly explored the numerical computation of $X_{\min}$, and effective procedures for both small and large problem dimensions have been proposed. The reliability of the derived schemes has been established by showing their convergence to $X_{\min}$, whereas several numerical experiments illustrate their efficiency in terms of both memory requirements and computational time.

In the large-scale setting, low-rank approximate solutions turned out to be accurate in terms of the relative residual norm. This suggests that it may be possible to show that the exact solution $X_{\min}$ presents a fast decay in its singular values, and this will be the topic of future works. The projection scheme adopted to solve the T-Sylvester equations arising from the Newton-Kleinman iteration failed to converge in some cases so that the solution to the T-Riccati equation could not be computed. A robust convergence theory for large-scale T-Sylvester equations solvers is still lacking in the literature, and we think it may be a very interesting research topic as also the numerical procedure for T-Riccati equations presented in this paper can benefit from it.

The promising results encourage us to tackle more difficult problems with data coming from real-life applications as the ones discussed in Section 1.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Acknowledgments. This work was inspired by Don Harding (Victoria University, Melbourne, Australia) in the context of the ARC Discovery Grant No DP1801038707 "New methods for solving large models with rational expectations", in which the first author serves as a consultant. Moreover, we are thankful to Froilán Dopico and Valeria Simoncini for providing us with the MATLAB implementations of Algorithm 3.1 in [11] and Algorithm 2 in [12], respectively.

We also thank Bruno Iannazzo for some remarks about the impact that Assumption 1.1 has on the coefficient matrices A and D, and the anonymous reviewers for their constructive comments. The second author is member of the Italian INdAM Research group GNCS.

REFERENCES

[1] P. BENNER AND Z. BUJANOVIĆ, *On the solution of large-scale algebraic Riccati equations by using low-dimensional invariant subspaces*, Linear Algebra Appl., 488 (2016), pp. 430–459.
[2] 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 (2018), pp. 301–330.
[3] ———, *A numerical comparison of different solvers for large-scale, continuous-time algebraic Riccati equations and LQR problems*, SIAM J. Sci. Comput., 42 (2020), pp. A957–A996.
[4] P. BENNER AND R. BYERS, *An exact line search method for solving generalized continuous-time algebraic Riccati equations*, IEEE Trans. Automat. Control, 43 (1998), pp. 101–107.
[5] P. BENNER, M. HEINKENSCHLOSS, J. SAAK, AND H. K. WEICHELT, *An inexact low-rank Newton-ADI method for large-scale algebraic Riccati equations*, Appl. Numer. Math., 108 (2016), pp. 125–142.
[6] P. BENNER, P. KÜRSCHNER, AND J. SAAK, *Low-rank Newton-ADI methods for large nonsymmetric algebraic Riccati equations*, J. Franklin Inst., 353 (2016), pp. 1147–1167.
[7] 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-Mitt., 36 (2013), pp. 32–52.
[8] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.
[9] M. BINDER AND M. H. PESARAN, *Multivariate linear rational expectations models: characterization of the nature of the solutions and their fully recursive computation*, Econometric Theory, 13 (1997), pp. 877–888.
[10] D. A. BINI, B. IANNAZZO, AND B. MEINI, *Numerical Solution of Algebraic Riccati Equations*, SIAM, Philadelphia, 2012.
[11] F. DE TERÁN AND F. M. DOPICO, *Consistency and efficient solution of the Sylvester equation for $\star$-congruence*, Electron. J. Linear Algebra, 22 (2011), pp. 849–863.
[12] F. M. DOPICO, J. GONZÁLEZ, D. KRESSNER, AND V. SIMONCINI, *Projection methods for large-scale T-Sylvester equations*, Math. Comp., 85 (2016), pp. 2427–2455.
[13] F. FEITZINGER, T. HYLLA, AND E. W. SACHS, *Inexact Kleinman-Newton method for Riccati equations*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 272–288.
[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, 2013.
[15] C.-H. GUO, *Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 225–242.
[16] M. H. GUTKNECHT, *Krylov subspace algorithms for systems with multiple right hand sides: an introduction*, Preprint, ETH-Zurich, Zurich, 2006. http://www.sam.math.ethz.ch/~mhg/pub/delhipap.pdf
[17] M. HEYOUNI AND K. JBILOU, *An extended block Arnoldi algorithm for large-scale solutions of the continuous-time algebraic Riccati equation*, Electron. Trans. Numer. Anal., 33 (2008/09), pp. 53–62. http://etna.ricam.oeaw.ac.at/vol.33.2008-2009/pp53-62.dir/pp53-62.pdf
[18] K. JBILOU, *Block Krylov subspace methods for large algebraic Riccati equations*, Numer. Algorithms, 34 (2003), pp. 339–353.
[19] Y. LIN AND V. SIMONCINI, *A new subspace iteration method for the algebraic Riccati equation*, Numer. Linear Algebra Appl., 22 (2015), pp. 26–47.
[20] THE MATHWORKS, *MATLAB version 9.3.0 (R2017b)*, 2017.
[21] D. PALITTA, *The projected Newton-Kleinman method for the algebraic Riccati equation*, Preprint on arXiv, 2019. https://arxiv.org/abs/1901.10199
[22] S. SCHMITT-GROHÉ AND M. URIBE, *Solving dynamic general equilibrium models using a second-order approximation to the policy function*, J. Econom. Dynam. Control, 28 (2004), pp. 755–775.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[23] V. SIMONCINI, *Analysis of the rational Krylov subspace projection method for large-scale algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 1655–1674.
[24] V. SIMONCINI, D. B. SZYLD, AND M. MONSALVE, *On two numerical methods for the solution of large-scale algebraic Riccati equations*, IMA J. Numer. Anal., 34 (2014), pp. 904–920.
[25] C. SIMS, *Solving linear rational expectations models*, Comput. Economics, 20 (2001), pp. 1–20.