THE DESCRIPTOR DISCRETE–TIME RICCATI EQUATION:
NUMERICAL SOLUTION AND APPLICATIONS*

CRISTIAN OARĂ† AND RALUCA ANDREI‡

Abstract. We investigate the numerical solution of a descriptor type discrete–time Riccati equation and give its main applications in several key problems in robust control formulated under very general hypotheses: rank compression (squaring down) of improper or polynomial systems with \(L^\infty\)–norm preservation, \((J, J')\)–spectral and \((J, J')\)–lossless factorizations of a completely general discrete–time system having any type of singularity, including arbitrary normal rank, poles and zeroes at infinity, at zero, or on the unit circle. Necessary and sufficient existence conditions together with computable formulas are given for the stabilizing and antistabilizing solutions of the descriptor Riccati equation in terms of an associated matrix pencil, without imposing any unnecessary restrictive assumptions on the matrix coefficients.

Key words. algebraic Riccati equation, matrix pencils, stabilizing and antistabilizing solutions

AMS subject classifications. 26C15, 65F30, 93B40, 93C35, 93C45, 93C55, 93C60

1. Introduction. Let \(C\) be the complex plane, \(\overline{C} := C \cup \{\infty\}\), \(D\) be the open unit disk, and \(\overline{D}\) its closure. By \(A^*\) we denote the conjugate transpose of the matrix \(A\) with coefficients in \(C\). The matrix pencil \(zE - A\) is called regular if it is square, \(E, A \in \mathbb{C}^{n \times n}\), and \(\det(zE - A) \not\equiv 0\). Given the matrix pencil \(zE - A\), we denote by \(\Lambda(zE - A)\) the union of its generalized eigenvalues (finite and infinite, multiplicities counting) [4].

In this paper we study the descriptor discrete–time algebraic Riccati equation (DDTARE)

\[
(1.1) \quad E^*XE - A^*XA - ((E - A)^*XB + L)R^{-1}(L^* + B^*X(E - A)) + Q = 0
\]

where \(E \in \mathbb{C}^{n \times n}\), \(A \in \mathbb{C}^{n \times n}\), \(B \in \mathbb{C}^{n \times m}\), \(L \in \mathbb{C}^{n \times m}\), \(Q = Q^* \in \mathbb{C}^{n \times n}\), \(R = R^* \in \mathbb{C}^{m \times m}\), and \(R\) is invertible. We do not make any further restrictive assumptions on the positivity of the matrix coefficients which will allow us to use the results for solving several problems in robust control under very general assumptions.

The DDTARE has in general many solutions. In particular, we are interested in two types of solutions introduced next.

Definition 1. We call \(X = X^* \in \mathbb{C}^{n \times n}\) a stabilizing solution for the DDTARE if \(X\) satisfies (1.1) and \(\Lambda(z(E + BF) - (A + BF)) \subset \overline{D}\), where

\[
(1.2) \quad F := -R^{-1}(L^* + B^*X(E - A))
\]

is called the stabilizing Riccati feedback. Correspondingly, \(X = X^* \in \mathbb{C}^{n \times n}\) is called an antistabilizing solution for the DDTARE if \(X\) satisfies (1.1) and \(\Lambda(z(E + BF) - (A + BF)) \subset \overline{C} - \overline{D}\), where \(F\) given by (1.2) is now called the antistabilizing Riccati feedback.

The Theorem bellow states the uniqueness of the stabilizing (antistabilizing) solution, if one exists.

*This work was supported by a grant of the Romanian National Authority for Scientific Research, CNCS - UEFISCDI, project number PN-II-ID-PCE-2011-3-0235.
†Department of Automatic Control and Systems Engineering, University “Politehnica” Bucharest, Splaiul Independenței 313, sector 6, RO 060042, Bucharest, Romania, cristian.oara@acse.pub.ro
‡Department of Automatic Control and Systems Engineering, University “Politehnica” Bucharest, Splaiul Independenței 313, sector 6, RO 060042, Bucharest, Romania, raluca@riccati.pub.ro.
Theorem 2. If the DDTARE (1.1) has a stabilizing (antistabilizing) solution then it is unique.

Our main goal is to give easy checkable necessary and sufficient existence conditions together with a numerical prototype algorithm for the stabilizing and antistabilizing solutions of the DDTARE in the spirit of [5, 6]. To this end, we need some specific characterizations of deflating subspaces of regular matrix pencils given next.

2. Deflating Subspaces and Descriptor Symplectic Pencils. Throughout this section \( zM - N \) denotes a regular matrix pencil. Deflating subspaces are the natural extension of invariant subspaces of a matrix to that of a deflating subspace of a regular pencil.

Definition 3. The linear space \( \mathcal{V} \subset \mathbb{C}^n \) is called a deflating subspace of the regular pencil \( zM - N \) if

\[
\dim(M\mathcal{V} + N\mathcal{V}) = \dim(\mathcal{V}).
\]

For a deflating subspace \( \mathcal{V} \) of the pencil \( zM - N \) denote by \( (zM - N)|_{\mathcal{V}} \) and \( \Lambda(zM - N)|_{\mathcal{V}} \) the map and the spectrum of the pencil restricted to \( \mathcal{V} \) (see for example [6]). The next result is instrumental for our developments and gives a useful characterization of deflating subspaces in terms of associated basis matrices.

Theorem 4. Let \( zM - N \) be a regular \( n \times n \) matrix pencil.

1. If \( \mathcal{V} = \langle V \rangle \) is an \( \ell \)-dimensional deflating subspace of \( zM - N \), where \( V \) is a basis matrix for \( \mathcal{V} \), then there exists a regular \( \ell \times \ell \) pencil \( zT - S \), which is strictly equivalent to \( (zM - N)|_{\mathcal{V}} \), such that

\[
M\mathcal{V} S = NVT.
\]

2. Conversely, if \( M\mathcal{V} S = NVT \) holds for a certain \( n \times \ell \) basis matrix \( V \) and a regular \( \ell \times \ell \) pencil \( zT - S \), then \( \mathcal{V} = \langle V \rangle \) is a deflating subspace of \( zM - N \) and \( zT - S \) is strictly equivalent to \( (zM - N)|_{\mathcal{V}} \).

Remark 5. Relation (2.2) generalizes the well-known characterization of an invariant subspace of a matrix to that of a deflating subspace of a regular pencil. Indeed, if \( M = I \) we get \( NVT = VS \) and since \( zT - S \) is regular it follows that \( T \) is invertible and \( NV = VS \), where \( S := ST^{-1} \).

The following particular matrix pencils will be used to characterize and compute the solutions to the DDTARE.

Definition 6. The \( (2n + m) \times (2n + m) \) matrix pencil

\[
z \begin{bmatrix}
E & 0 & B \\
Q & A^* & L \\
0 & 0 & 0
\end{bmatrix} - \begin{bmatrix}
A & 0 & B \\
Q & E^* & L \\
L^* & B^* & R
\end{bmatrix}
\]

is called the descriptor symplectic pencil (DSP) associated with the DDTARE (1.1) and the \( 2n \times 2n \) matrix pencil

\[
z \begin{bmatrix}
E - BR^{-1}L^* & -BR^{-1}B^* \\
Q - LR^{-1}L^* & A^* - LR^{-1}B^*
\end{bmatrix} - \begin{bmatrix}
A - BR^{-1}L^* & -BR^{-1}B^* \\
Q - LR^{-1}L^* & E^* - LR^{-1}B^*
\end{bmatrix}
\]

is called the reduced descriptor symplectic pencil (RDSP) associated with the DDTARE (1.1), where \( E \in \mathbb{C}^{n \times n} \), \( A \in \mathbb{C}^{n \times n} \), \( B \in \mathbb{C}^{n \times m} \), \( L \in \mathbb{C}^{m \times m} \), \( Q = Q^* \in \mathbb{C}^{n \times n} \), and \( R = R^* \in \mathbb{C}^{m \times m} \) invertible.

We give further a characterization of the spectrum of the DSP and RDSP. Let \( n^e, n^m, n^f, n^0 \) be the number of finite generalized eigenvalues with modulus strictly
greater than one, strictly less than one, equal to 1 and equal to 0, respectively, and let \( n^\infty \) be the number of infinite generalized eigenvalues (everywhere multiplicities counting).

**Proposition 7.** Assume \( E - A \) and \( R \) are invertible. Then we have:
1. The DSP and RDSP are regular pencils.
2. The DSP has \( n^\infty = n_0 + m \) and \( n^{<1} = n^{>1} + n^0 \).
3. The RDSP has \( n^\infty = n^0 \) and \( n^{<1} = n^{>1} + n^\infty \).
4. If \( n^1 = 0 \) then the DSP has \( n^{<1} = n \) and \( n^{>1} + n^\infty = n + m \) while the RDSP has \( n^{<1} = n \) and \( n^{>1} + n^\infty = n \).

From the above result and Theorem 4 it follows that a DSP has an \( n \)-dimensional deflating subspace with stable spectrum (included in \( D \)) if and only if \( n^1 = 0 \). Furthermore, the RDSP has an \( n \)-dimensional deflating subspace with stable spectrum if and only if it has an \( n \)-dimensional deflating subspace with antistable spectrum (included in \( C - D \)) if and only if \( n^1 = 0 \).

The following result gives a useful characterization of deflating subspaces of a regular DSP.

**Lemma 8.** Assume the DSP has an \( n \)-dimensional stable deflating subspace with basis matrix
\[
V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} \quad n \times m.
\]

Then
\[
V_1^*(E - A)^*V_2 = V_2^*(E - A)V_1.
\]

Furthermore, if the RDSP has an \( n \)-dimensional stable (or antistable) deflating subspace with basis matrix
\[
V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \quad n \times m
\]
then (2.6) holds true as well.

Interestingly, (2.6) does not necessary hold true for a basis matrix (2.5) of an \( n \)-dimensional antistable deflating subspace of the DSP (2.3).

An \( n \)-dimensional stable deflating subspace of the DSP is called *disconjugate* provided the matrix \( V_1 \) in (2.5) is invertible. An \( n \)-dimensional stable (antistable) deflating subspace of the RDSP is called *disconjugate* provided the matrix \( V_1 \) in (2.7) is invertible.

**3. Main Result.** The following next two theorems contain the main results, giving numerically checkable existence conditions and computable formulas for the stabilizing and antistabilizing solutions of the DDTARE.

**Theorem 9.** Assume \( E - A \) is invertible. The following two assertions are equivalent.
1. \( R \) is invertible and the DDTARE (1.1) has a stabilizing solution.
2. The DSP (2.3) is regular and has an \( n \)-dimensional stable deflating subspace which is disconjugate.

Moreover, if \( V \) is a basis matrix partitioned as in (2.5) for the \( n \)-dimensional stable deflating subspace of the DSP then the stabilizing solution can be computed from
\[
X = V_2V_1^{-1}(E - A)^{-1}
\]
and the stabilizing Riccati feedback can be computed from
\begin{equation}
F = V_3 V_1^{-1}.
\end{equation}

**Theorem 10.** Assume $E - A$ is invertible. The following two assertions are equivalent.

1. $R$ is invertible and the DDTARE (1.1) has a stabilizing (antistabilizing) solution.
2. The RDSP (2.3) is regular and has an $n$–dimensional stable (antistable) deflating subspace which is disconjugate.

Moreover, if $V$ is a basis matrix partitioned as in (2.7) for the $n$–dimensional stable (antistable) deflating subspace of the RDSP then the stabilizing (antistabilizing) solution can be computed from
\begin{equation}
X = V_2 V_1^{-1}(E - A)^{-1}
\end{equation}
and the stabilizing (antistabilizing) Riccati feedback can be computed from (1.2).

The results above show that the computational burden for checking the existence of and computing the solutions to the DDTARE lies in the computation of a maximal stable or antistable deflating subspace of a regular pencil for which efficient and stable numerical algorithms based solely on unitary transformations are available [13, 6].

Though both the stabilizing and antistabilizing solutions may be computed by employing the RDSP; it is advisable to compute the stabilizing solution by means of the DSP as its coefficients do not imply any possibly ill–conditioned matrix inversion. Unfortunately, the antistabilizing solution can be computed solely from the RDSP since the corresponding Riccati feedback $F$ may place some spectrum of the pencil $z(E+BF)-(A+BF)$ at infinity, which can be mixed up in the DSP with the $m$ simple infinite generalized eigenvalues which actually do not correspond to Smith–McMillan zeros.

**4. Applications.** We exemplify the key role of the DDTARE in the solution of two robust control problems, rank compression (or squaring down) with $L^\infty$–norm preservation [7, 9, 11, 1] and ($J$, $J'$)–spectral factorization [2, 3], both formulated for linear time-invariant discrete-time systems given by generalized state–space representation
\begin{equation}
G(z) = C(zE - A)^{-1}B + D =: \begin{bmatrix} A - zE & B \\ C & D \end{bmatrix}.
\end{equation}

Generalized state-space systems [14] are used to represent the most general discrete–time systems, even improper or polynomial [12]. Other applications of the DDTARE include the ($J$, $J'$)–lossless factorization of a general discrete–time system [10] and zero cancellation with norm preservation [8].

The first result provides in particular a solution to the general squaring–down problem with $L^\infty$ norm preservation and relies on a DDTARE with positive quadratic term. Squaring–down the $p \times m$ system (4.1) with $L^\infty$ norm preservation means to design a postcompensator $G_{\text{post}}(z)$ and a precompensator $G_{\text{pre}}(z)$ such that
\[ \tilde{G}_{\text{sqd}}(z) := G_{\text{post}}(z)G(z)G_{\text{pre}}(z) = \begin{bmatrix} G_{\text{sqd}}(z) & 0 \\ 0 & 0 \end{bmatrix}, \]
where \( G_{sqd}(z) \) is square, invertible, and has the same \( L^\infty \) norm as the original system \( G(z) \). We consider the general case in which \( p, m \) and the normal rank \( r \) of \( G(z) \) are all arbitrary. Further on we deal with the postcompensation problem (full row rank compression) only since the solution for the precompensation follows by duality. Alternatively, the problem may be seen as a rank revealing compression of a rational matrix by an isometric matrix [7].

The second result gives the solution to the general \((J, J')\)-spectral factorization problem of a general discrete–time system having any type of singularity, including arbitrary normal rank, poles and zeroes at infinity, at zero, or on the unit circle [10]. In particular, the result may be applied to spectral factorization of a general polynomial system. The spectral factorization problem for (4.1) consists in finding a system \( \Pi(z) \) which has full row normal rank and only marginally stable zeros such that

\[
G^\#(z)JG(z) = \Pi^\#(z)J' \Pi(z),
\]

where \( G(z)\Pi^{(+)}(z) \) has no poles on the unit circle. Here \( G^\#(z) := G^*(\bar{z}) \), \( \Pi^{(+)}(z) \) stands for the Moore–Penrose pseudoinverse of \( \Pi(z) \), and \( J \) and \( J' \) are two constant signature matrices.

Both results use some special decomposition of the generalized state–space realization of the system to start with, but with specific properties for each case. Let the \( p \times m \) generalized state–space system given by the irreducible realization (4.1) of order \( n \). Then there exist three orthogonal matrices \( U, Q \) and \( Z \) such that

\[
\begin{bmatrix}
U & 0 & 0 & Q
\end{bmatrix}
\begin{bmatrix}
A - zE & B \\
C & D
\end{bmatrix}
\begin{bmatrix}
X \\
0
\end{bmatrix} =
\begin{bmatrix}
A_{rz} - zErz & * & * & * \\
0 & A_t - zEt & B_t(1-z) & * \\
0 & 0 & 0 & B_o
\end{bmatrix}
\begin{bmatrix}
Z
\end{bmatrix}
\]

where the matrices \( A_t - Et \) and \( B_o \) are invertible, \( D_t \) has full column rank, the pair \( [A_t - zEt, B_t] \) is controllable [14] and either of the following two properties hold:

(A) The pencil \( A_{rz} - zErz \) has full row normal rank and the pencil

\[
\begin{bmatrix}
A_t - zEt \\
C_t
\end{bmatrix}
\begin{bmatrix}
B_t(1-z) \\
D_t
\end{bmatrix}
\]

has full column rank \( \forall z \in C \).

(B) The pencil \( A_{rz} - zErz \) has full row rank in \( D, \) \( E_{rz} \) has full row rank, and the pencil (4.4) has full column rank in \( D \).

**THEOREM 11** (\( L^\infty \)-norm preserving rank compression). Let (4.1) be an irreducible generalized state–space system and \( U, Q \) and \( Z \) the three orthogonal matrices for which (4.3) holds with property (A). The DDTARE

\[
E_t^*XEt - A_t^*XAt - ((Et - Aet)^*XBt + C_t^*Dt) \\
\times(D_t^*Dt)^{-1}(D_t^*Ct + B_t^*X(Et - Aet)) + C_t^*C_t = 0
\]

has an invertible stabilizing symmetric solution \( X \). An \( L^\infty \) norm–preserving full row rank compression and the squared–down system are given by

\[
G_{post}(z) = I - (C_t + D_tF)(zE_{post} - A_{post})^{-1}B_{post}(1-z),
\]

\[
G_{sqd}(z) = D_{sqd} + C_{sqd}(zE - A)^{-1}B
\]
where

\[
B_{\text{post}} := -X_{s}^{-1}(E_{\ell} - A_{\ell})^{-1}(C_{\ell} + D_{\ell}F_{s}),
\]

\[
zE_{\text{post}} - A_{\text{post}} := z(E_{\ell} - B_{\text{post}}C_{\ell}) - (A_{\ell} - B_{\text{post}}C_{\ell}),
\]

\[
[\begin{array}{cc}
C_{\text{sqd}} & D_{\text{sqd}}
\end{array}] := [\begin{array}{ccc}
0 & -D_{\ell}F_{s} & D_{\ell} \\
\end{array}] Z^{*},
\]

and \(F_{s}\) is Riccati stabilizing feedback \((1.2)\) for the DDTARE \((4.5)\).

**Theorem 12** \(((J, J')-\text{spectral factorization})\). Let \(G(z)\) be a general \(\text{rmf}\) given by an irreducible realization \((4.1)\), and let \(U, Q\) and \(Z\) be three constant unitary matrices such that \((4.3)\) holds with property \((B)\). The \((J, J')-\text{spectral factorization problem} \((4.2)\) has a solution if and only if there is a constant invertible matrix \(V\) such that

\[
D_{\ell}^{*}JD_{\ell} = V^{*}J'V \quad (4.8)
\]

and the DDTARE

\[
E_{\ell}^{*}XE_{\ell} - A_{\ell}^{*}XA_{\ell} - (E_{\ell} - A_{\ell})^{*}XB_{\ell} + C_{\ell}^{*}JC_{\ell} = 0 \quad (4.9)
\]

has an invertible stabilizing solution \(X_{s}\). The spectral factor in \((4.2)\) is given by

\[
\Pi(z) := D_{\text{spec}}^{*} + C_{\text{spec}}^{*}(zE - A)^{-1}B,
\]

where

\[
[\begin{array}{cc}
C_{\text{spec}} & D_{\text{spec}}
\end{array}] := [\begin{array}{ccc}
0 & -VF_{s} & V \\
\end{array}] Z^{*} \text{ and } F_{s} \text{ is the corresponding Riccati stabilizing feedback \((1.2)\) for the DDTARE \((4.9)\).}
\]

**REFERENCES**


