Numerical Solution of Differential Algebraic Riccati Equations

P. Kunkel [‡] and V. Mehrmann [†]

ABSTRACT We consider Riccati matrix differential algebraic equations arising from singular or descriptor control problems. We discuss solvability of such equations under different conditions. In order to apply numerical methods for differential algebraic systems one has to transform the equation. Unfortunately then these equations have a linear part, which is described by a singular pencil and thus, the usual integration methods do not apply. Under some conditions, which we discuss, these singularities can be removed by a preprocessing algorithm and the equation can then be solved by well–known methods for differential algebraic systems like DASSL of L. Petzold or LIMEX of Deuflhard, Hairer and Zugck. We discuss the numerical procedures and give some numerical examples.

§1 Introduction.

We study the differential algebraic Riccati equation (DARE)

$$(1.1)\quad -E^T(t)\dot{X}(t)E(t) = E^T(t)X(t)A(t)+A^T(t)X(t)E(t)+Q(t)-E^T(t)X(t)W(t)X(t)E(t)$$

with "terminal" condition

$$(1.2)\quad E^T(t_f)X(t_f)E(t_f) = M.$$

Here $X(t), E(t), A(t), Q(t), W(t), M \in \mathbb{R}^{n,n}$ are sufficiently smooth and $X(t), Q(t), W(t), M$ are symmetric in the interval $[t_0, t_f] \subset \mathbb{R}$. The coefficient matrices have to satisfy further conditions which will be discussed later on.
Such equations arise for example from optimal regulator problems with differential algebraic equations (DAE's) or from optimal filters with differential algebraic equations. Consider for example the optimal control problem:

$$(1.3)\ \text{Minimize}\quad J(x,u) = \frac{1}{2}\int_{t_0}^{t_f}[x(t)^T Q(t)x(t) + u(t)^T R(t)u(t)]dt$$

subject to

$$(1.4)\quad E(t)\dot{x}(t) = A(t)x(t) + B(t)u(t)\ ,\quad x(t_0) = x^0\ ,$$

where $E(t), A(t), Q(t) \in \mathbb{R}^{n,n}$, $B(t) \in \mathbb{R}^{n,m}$, $R(t) \in \mathbb{R}^{m,m}$, $Q(t) = Q(t)^T$, $R(t) = R(t)^T, R(t)$ positive definite, $Q(t)$ positive semidefinite.

In the next section we show how under some further assumptions the optimal solution of (1.3), (1.4) can be obtained via the solution of (1.1), (1.2).

For the case that the matrices E, A, Q, R, B are independent of t this is shown in Mehrmann [M 1], [M 2]. In the nonautonomous case where $E(t)$ is nonsingular, this follows from standard linear control theory, e.g. Athans/Falb [A 1] or Knobloch/Kwakernaak [K 1]. For the infinite horizon problem , i.e. $t_f = \infty$, previous work is by Arnold [A 2], Bender and Laub [B 1], see also Mehrmann [M 1] and the list of references therein.

In this paper we discuss the numerical solution of (1.1), (1.2) regardless whether they come from an optimal control problem or not. The obvious approach is to vectorize this equation and, leaving out the extraneous components of the symmetric matrix, to transform it into a standard differential algebraic equation in $\mathbb{R}^{\frac{n(n+1)}{2}}$. If $E(t)$ is singular then this approach unfortunately produces an equation, where the linear part of the equation is described by a singular pencil and thus the standard solvers for differential algebraic equations like LIMEX [D 1] or DASSL [P 1] do not apply. We will discuss under which conditions these unfortunate singularities can be removed and how this can be done numerically, so that the solution can then be obtained from standard solvers.
We demonstrate the proposed method with some numerical examples.
For descriptor control problems also numerical methods for solving the variational equations, which form a linear two point boundary value problem, could be used. It is also possible to embed the Riccati differential equation in such a boundary value problem even if it does not come from a control application. For linear differential algebraic boundary value problems, numerical methods have been examined recently by several authors. Reviews of the known methods are given in [P 3] or [G 6].

§2 Preliminaries.

In this section we introduce our notation, give some preliminary results and show how equation (1.1), (1.2) can be obtained from the control problem (1.3), (1.4). By $\mathbb{R}^{n,n}(\mathbb{C}^{n,n})$ we denote the real (complex) $n \times n$ matrices.

2.1 Definition. *Let* $A \in \mathbb{R}^{n,n}$. *Then we denote by*

$\mathcal{N}(A)$ *the nullspace of A;*

$\mathcal{N}(A)^{\perp}$ *the orthogonal complement of* $\mathcal{N}(A)$*;*

$\mathrm{vec}(A)$ *the vector in* $\mathbb{R}^{n^2}$ *obtained by concatenating the columns of A one after the other to one big vector.*

In the following we make frequent use of the Kronecker canonical form of a matrix pencil $\alpha E - \beta A$, which is as follows:

2.2 Theorem. *Let* $E, A \in \mathbb{R}^{n,n}$, *then there exist* $P, Z \in \mathbb{C}^{n,n}$ *nonsingular such that*

$$P(\alpha E - \beta A)Z = \mathrm{diag}\{L_{\epsilon_1}, \cdots, L_{\epsilon_p}, L^p_{\eta_1}, \cdots, L^p_{\eta_q}, J_{\rho_1}, \cdots, J_{\rho_r}, N_{\sigma_1}, \cdots, N_{\sigma_s}\} \tag{2.3}$$

where

a) L_{ϵ_j} *is an* $\epsilon_j \times (\epsilon_j + 1)$ *bidiagonal pencil of the form*

$$\alpha \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} - \beta \begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}; \tag{2.4}$$

b) $L^p_{\eta_j}$ *is an* $(\eta_j + 1) \times \eta_j$ *bidiagonal pencil of the form*

$$\alpha \begin{bmatrix} 1 & & \\ 0 & \ddots & \\ & \ddots & 1 \\ & & 0 \end{bmatrix} - \beta \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 0 \\ & & 1 \end{bmatrix}; \tag{2.5}$$

c) $J_{\rho_j}(\lambda_j)$ *is a* $\rho_j \times \rho_j$ *Jordan block*

$$\alpha \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} - \beta \begin{bmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{bmatrix}; \tag{2.6}$$

d) N_{σ_j} *is a* $\sigma_j \times \sigma_j$ *nilpotent pencil of the form*

$$\alpha \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} - \beta \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}. \tag{2.7}$$

PROOF: E.g. Gantmacher [G 2]. ■

We then have the following definition:

2.8 DEFINITION. *A pencil* $\alpha E - \beta A$ *is called regular if no blocks of type (2.4), (2.5) occur in its Kronecker canonical form (2.3).*

By $\operatorname{ind}_\infty(E, A)$ *we denote the size of the largest block* N_{σ_j}.

Two pencils $\alpha E - \beta A$, $\alpha \widetilde{E} - \beta \widetilde{A}$ *are called equivalent if there exist* $P, Z \in \mathbb{C}^{n,n}$ *non-singular, such that*

$$\alpha \widetilde{E} - \beta \widetilde{A} = P(\alpha E - \beta A)Z. \tag{2.9}$$

We now consider control problem (1.3), (1.4). In the following for simplicity we often leave out the dependency on t in the formulas. We give some sufficient conditions here such that control problem (1.3), (1.4) leads to the DARE (1.1), (1.2). These conditions do not give a complete characterization of all the possible cases but this is not the topic of this paper. A discussion of this point and a list of references can be found in [M 1]. Assume that

$$\begin{aligned} &\alpha E(t) - \beta A(t) \text{ is regular, rank (E(t)) fixed, independent of t} \\ &\text{and } \operatorname{ind}_\infty(E(t), A(t)) = 1 \text{ for all } t \in [t_0, t_f], \end{aligned} \tag{2.10}$$

This means in the notation of Petzold [P 2] that the DAE (1.4) is solvable and has index one.
Assume further that in the singular value decomposition

$$E(t) = U(t) \begin{bmatrix} \Sigma(t) & 0 \\ 0 & 0 \end{bmatrix} V^T(t) \tag{2.11}$$

the function $V(t)$ is differentiable.
Let $y(t) := V^T(t)x(t),\ y^0 = V^T(t_0)x^0$, then

$$\dot{y}(t) = \dot{V}^T(t)x(t) + V^T(t)\dot{x}(t) = \dot{V}^T(t)V(t)y(t) + V^T(t)\dot{x}(t). \tag{2.12}$$

Then system (1.4) is equivalent to the system

$$\begin{bmatrix} \Sigma(t) & 0 \\ 0 & 0 \end{bmatrix} \dot{y}(t) = \left[U^T(t)A(t)V(t) + \begin{bmatrix} \Sigma(t) & 0 \\ 0 & 0 \end{bmatrix} \dot{V}^T(t)V(t) \right] y(t) + U^T(t)B(t)u(t) \tag{2.13}$$

and (1.3) transforms to

$$\widetilde{J}(y,u) := \frac{1}{2}\int_{t_0}^{t_f} [y(t)^T V(t)^T Q(t)V(t)y(t) + u(t)^T R(t)u(t)]dt. \tag{2.14}$$

Set

$$\widetilde{Q}(t) := \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} := V(t)^T Q(t)V(t),$$

$$\widetilde{A}(t) := \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} := U^T(t)A(t)V(t) + \begin{bmatrix} \Sigma(t) & 0 \\ 0 & 0 \end{bmatrix} \dot{V}(t)V(t),$$

$$\widetilde{B}(t) := \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix} := U^T(t)B(t), \quad y := \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

where the partitioning is analogous to $\begin{bmatrix} \Sigma(t) & 0 \\ 0 & 0 \end{bmatrix}$. Assumption (2.10) guarantees that A_{22} is nonsingular, thus we can solve for y_2 and insert this solution.

$$y_2(t) = -A_{22}^{-1}(t)(A_{21}(t)y_1(t) + B_{21}(t)u(t)) \tag{2.15}$$

$$y_2(t_0) = -A_{22}^{-1}(t_0)(A_{21}(t_0)y_1^0 + B_{21}(t_0)u(t_0)). \tag{2.16}$$

The latter condition is an initial condition for $u(t)$, which restricts the set of admissible controls or the set of initial conditions x^0 ! Inserting (2.15) in (2.13), (2.14) and reordering gives

$$\Sigma(t)\dot{y}_1(t) = [A_{11}(t) - A_{12}(t)A_{22}^{-1}(t)A_{21}(t)]y_1(t) + \tag{2.17}$$

$$[B_{11}(t) - A_{12}(t)A_{22}^{-1}(t)B_{21}(t)]u(t) =: A_1(t)y_1(t) + B_1(t)u(t), \quad y_1(t_0) = y_1^0.$$

$$\widehat{J}(y_1,u) := \frac{1}{2}\int_{t_0}^{t_f} [y_1^T[Q_{11} - A_{21}^T A_{22}^{-T} Q_{21} - Q_{12}A_{22}^{-1}A_{21} + A_{21}^T Q_{22} A_{22}^{-1} A_{21}]y_1 \tag{2.18}$$

$$+2y_1^T[A_{21}^T A_{22}^{-T} Q_{22} A_{22}^{-1} B_{21} - Q_{12}A_{22}^{-1}B_{21}]u + u^T[R + B_{21}^T A_{22}^{-T} Q_{22} A_{22}^{-1} B_{21}]u]dt$$

$$=: \frac{1}{2}\int_{t_0}^{t_f}[y_1^T Q_1 y_1 + 2y_1^T S_1 u + u^T R_1 u]dt.$$

Except for the restriction on $u(t)$, this is now a standard control problem for y_1 , which under the typical controllability conditions (e.g. Athans/Falb [A 1]) has the unique solution

$$u(t) = -R_1^{-1}(t)[S_1^T(t) + B_1^T(t)X_1(t)\Sigma(t)]y_1(t) \tag{2.19}$$

where $X_1(t)$ satisfies the Riccati differential equation

$$-\Sigma^T(t)\dot{X}_1(t)\Sigma(t) = \Sigma^T(t)X_1(t)A_1(t) + A_1^T(t)X_1(t)\Sigma(t) + Q_1(t)-$$

$$\Sigma^T(t)[B_1(t)^T X_1(t) + S_1^T(t)]^T R_1(t)^{-1}[B_1^T(t)X_1(t) + S_1^T(t)]\Sigma(t), \tag{2.20}$$

$$\Sigma(t_f)^T X_1(t_f)\Sigma(t_f) = 0.$$

Thus, if we assume that $u(t)$ satisfies condition (2.16), we have a solution which is only given in terms of y_1 . In many applications the differential algebraic part of the DAE (1.4) is given by a linear equation or linear differential equation, not subject to control. Then (2.16) is a consistency condition for the initial vector x^0 and does not lead to a restricyion of $u(t)$.
We are in this paper mainly interested in the DARE (1.1), (1.2), thus we assume that condition (2.16) is satisfied. Then it follows that the optimal solution to the original problem is given only in terms of y_1 and the corresponding trajectory is

$$x(t) = V(t)\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = V\begin{bmatrix} y_1 \\ -A_{22}^{-1}(A_{21} - B_{21}R_1^{-1}[S_1^T + B_1^T X_1 \Sigma])y_1 \end{bmatrix} \tag{2.21}$$

where $y_1(t)$ is the solution of

$$\Sigma(t)\dot{y}_1(t) = [A_1(t) - B_1(t)R_1^{-1}(t)[S_1^T(t) + B_1^T(t)X_1(t)\Sigma(t)]y_1(t), \quad y_1(t_0) = y_1^0. \tag{2.22}$$

The analogue to (2.20) without a pre-reduction is

$$\begin{aligned} -\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}^T \overset{\bullet}{\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} &= \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}^T \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \\ &+ \begin{bmatrix} A_{11}^T & A_{21}^T \\ A_{12}^T & A_{22}^T \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \\ &- \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}^T \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix} R^{-1}[B_{11}^T \; B_{21}^T] \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \\ \Sigma(t_f)^T X_{11}(t_f)\Sigma(t_f) &= 0. \end{aligned} \tag{2.23}$$

This system decouples into the four equations

$$0 = Q_{22}, \tag{2.24}$$

which, by assuming that Q is positive semidefinite, implies that also Q_{21}, Q_{12} vanish.

$$0 = (A_{12}^T X_{11} + A_{22}^T X_{21})\Sigma + Q_{21} \tag{2.25}$$

$$0 = \Sigma(X_{12}A_{22} + X_{11}A_{12}) + Q_{12} \tag{2.26}$$

$$\begin{aligned} -\Sigma\dot{X}_{11}\Sigma &= \Sigma(X_{11}A_{11} + X_{12}A_{21}) + (A_{11}^T X_{11} + A_{21}^T X_{21})\Sigma \\ &+ Q_{11} - \Sigma(B_{11}^T X_{11} + B_{21}^T X_{21})^T R^{-1}(B_{11}^T X_{11} + B_{21}^T X_{21})\Sigma, \end{aligned} \tag{2.27}$$

where we have used the symmetry of X in (2.27). Solving (2.25), (2.26) for X_{12}, X_{21} and inserting in (2.27) we obtain differential equation (2.20) with X_1 replaced by X_{11}. And if we set

$$u = -R^{-1}[B_{11}^T B_{21}^T]\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{2.28}$$

we obtain exactly the solution in (2.19). Thus we can work directly with the differential algebraic Riccati equation (2.23). The components in X_{22} are not occuring in the optimal control, so they can be arbitrary and do not affect the solution.

§3 Reduction to standard Riccati differential equations.

For standard matrix Riccati differential equations, i.e. $E(t) = I$ in (1.1), (1.2) the theory is well established, e.g. Reid [R 1]. We discuss now several cases, where we can transform the DARE (1.1), (1.2) to a form, where the standard theory can be applied.

The first case we consider is

$$E(t) \text{ nonsingular for all } t \in [t_0, t_f] \tag{3.1}$$

In this case (1.1), (1.2) is equivalent to the standard Riccati differential equation

$$\begin{aligned} -\dot{X}(t) &= X(t)A(t)E^{-1}(t) + (E^{-1}(t))^T A(t)^T X(t) \\ &+ (E^{-1}(t))^T Q(t)E(t) - X(t)W(t)X(t), \end{aligned} \tag{3.2}$$

$$X(t_f) = (E(t_f)^{-1})^T M \; E(t_f)^{-1}. \tag{3.3}$$

Although this is a viable approach from the theoretical point of view, it is clearly not in general advisable for the numerical treatment, since $E(t)^{-1}$ may be very ill conditioned.

The second case we consider is

$$\begin{aligned} &E(t) \text{ continuous, } \alpha E(t) - \beta A(t) \text{ regular, } \operatorname{rank}(E(t)) = p < n \\ &\operatorname{ind}_\infty(E(t), A(t)) = q \leq 1 \text{ for all } t \in [t_0, t_f], \\ &p, q \text{ independent of } t, \end{aligned} \tag{3.4}$$

and if $U(t_f)\Sigma(t_f)V(t_f)^T = E(t_f)$ is the singular value decomposition of $E(t_f)$, then

$$U(t_f)^T E(t) = \begin{bmatrix} E_1(t) \\ 0 \end{bmatrix} \begin{matrix} \} \, p \\ \end{matrix} \text{ for all } t \in [t_0, t_f]. \tag{3.5}$$

Observe that this case includes the case

$$E, A \text{ constant in } t, \; \alpha E - \beta A \text{ regular }, \operatorname{ind}_\infty(E, A) \leq 1.$$

If neither of the conditions (3.1), (3.4) and (3.5) holds then the solution of these equations faces the well–known difficulties as discussed for example in Petzold [P 2]. One may have no solution or nonunique solutions or one needs further differentiability conditions. For these reasons we discuss here only the cases, where conditions (3.1) or (3.4) and (3.5) hold.

In order to perform the reduction under condition (3.4) and (3.5) we need a continous decomposition of $E(t)$ of the form

$$U(t_f)^T E(t) V(t) = \begin{bmatrix} \underbrace{E_{11}(t)}_{p} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} \} \, p \\ \end{matrix}, \tag{3.7}$$

where $E_{11}(t)$ is nonsingular in $[t_0, t_f]$ by (3.4).

To get this let

$$E_1(t)^T = \widetilde{V}(t) \begin{bmatrix} R(t) \\ 0 \end{bmatrix} =: \left[\widetilde{V}_1(t), \widetilde{V}_2(t) \right] \begin{bmatrix} R(t) \\ 0 \end{bmatrix} \tag{3.8}$$

be the QR–decomposition of $E_1(t)^T$ in (3.5). Then the columns of $\widetilde{V}_2$ belong to $\mathcal{N}(E_1)$ and the columns of $\widetilde{V}_1$ belong to $\mathcal{N}(E_1)^\perp$. Continuity of V can be obtained via the solution of an optimization problem for the computation of orthogonal matrices $Q_1(t), Q_2(t)$ such that

$$V(t) = [\widetilde{V}_1(t) Q_1(t), \widetilde{V}_2(t) Q_2(t)], \tag{3.9}$$

as it is discussed in Rheinboldt [R 2]. One obtains

$$U(t_f)^T E(t) V(t) = \begin{bmatrix} E_1(t) \\ 0 \end{bmatrix} [\widetilde{V}_1(t) Q_1(t), \widetilde{V}_2(t) Q_2(t)] = \begin{bmatrix} \underbrace{E_{11}(t)}_{p} & 0 \\ 0 & 0 \end{bmatrix} \begin{matrix} \} \, p \\ \end{matrix} =: \widetilde{E}(t), \tag{3.10}$$

since $E_1(t)\widetilde{V}_2(t) = 0$ for all $t \in [t_0, t_f]$ and hence $E_{11}(t)$ is continuous and nonsingular. Now transform (1.1), (1.2) with $U(t_f), V(t)$. Set

$$\left\{ \begin{array}{lll} y(t) & := \begin{bmatrix} Y_{11}(t) & Y_{12}(t) \\ Y_{21}(t) & Y_{22}(t) \end{bmatrix} := U(t_f)^T X(t) U(t_f), \\ \widetilde{A}(t) & := \begin{bmatrix} A_{11}(t) & A_{12}(t) \\ A_{21}(t) & A_{22}(t) \end{bmatrix} := U(t_f)^T A(t) V(t), \\ \widetilde{Q}(t) & := \begin{bmatrix} Q_{11}(t) & Q_{12}(t) \\ Q_{21}(t) & Q_{22}(t) \end{bmatrix} := V(t)^T Q(t) V(t), \\ \widetilde{W}(t) & := \begin{bmatrix} W_{11}(t) & W_{12}(t) \\ W_{21}(t) & W_{22}(t) \end{bmatrix} := U(t_f)^T W(t) U(t_f), \\ \widetilde{M} & := \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} := V(t_f)^T M V(t_f). \end{array} \right. \tag{3.11}$$

Then (1.1), (1.2) decouples into the four equations:

$$\begin{aligned} -E_{11}^T \dot{Y}_{11} E_{11} = {} & (A_{11}^T Y_{11} + A_{21}^T Y_{21}) E_{11} + E_{11}^T (Y_{11} A_{11} + Y_{12} A_{21}) + Q_{11} \\ & - E_{11}^T ([Y_{11} W_{11} + Y_{12} W_{21}] Y_{11} + [Y_{11} W_{12} + Y_{12} W_{22}] Y_{21}) E_{11} \end{aligned} \tag{3.12}$$

$$0 = E_{11}^T [Y_{11} A_{12} + Y_{12} A_{22}] + Q_{12} \tag{3.13}$$

$$0 = [A_{22}^T Y_{21} + A_{12}^T Y_{11}]E_{11} + Q_{21} \tag{3.14}$$

$$0 = Q_{22}\,. \tag{3.15}$$

$$E_{11}^T(t_f)Y_{11}(t_f)E_{11}(t_f) = M_{11}, \quad 0 = M_{12}, \quad 0 = M_{22}, \tag{3.16}$$

where we have left out the argument t in (3.12),...,(3.15) for simplification.

Equations (3.13), (3.14) are equivalent, equation (3.16) contains two consistency conditions and condition (3.15) is a solvability condition.

Theoretically we could now invert $E_{11}(t)$, solve for $X_{12}(t)$, vectorize $X_{11}(t)$, throw away the extraneous equations due to symmetry in $X_{11}(t)$ and use an initial value solver to solve for $X_{11}(t)$. For the obvious numerical reasons we would like to use a DAE solver instead, as we will do in section 5.

§4 Vectorizing the differential algebraic Riccati equation.

The numerical solution of the differential algebraic Riccati equation can be approached by vectorizing the equation, leaving out extraneous variables and then using a software package for systems of nonlinear differential algebraic equations. In this section we discuss the properties of the vectorized system in terms of the spectral properties of the pencil $\alpha E - \beta A$. Equation (1.1), (1.2) in the vectorized form reads as follows

$$\begin{aligned} -\underbrace{(E(t)\otimes E(t))}_{\widehat{E}}{}^T \operatorname{vec}(\dot{X}(t)) &= (\underbrace{(E(t)\otimes A(t))^T + (A(t)\otimes E(t))^T}_{\widehat{A}})\operatorname{vec}\,(X(t)) \\ + \underbrace{\operatorname{vec}(Q(t))}_{\widehat{Q}} - (E(t)\otimes E(t))^T &\underbrace{\operatorname{vec}(X(t)W(t)X(t))}_{\Phi(x)} \end{aligned} \tag{4.1}$$

$$(E(t_f)\otimes E(t_f))^T \operatorname{vec}(X(t_f)) = \operatorname{vec}(M) \tag{4.2}$$

Set $\widehat{E} = (E(t)\otimes E(t))^T, \widehat{A} = (E(t)\otimes A(t))^T + (A(t)\otimes E(t))^T$. To analyze the properties of systems (4.1), (4.2) we discuss the autonomous case, i.e. E, A, Q, W are independent of t . In the case that E, A are dependent of t, the difficulties that we observe clearly occur also in the nonautonomous case.

In the following we treat only the case that $\alpha E - \beta A$ is a regular pencil. If $\alpha E - \beta A$ is singular, then we will obtain a singular system (4.1), (4.2) which in general has index higher than one and we face the difficulties already mentioned in section 3. We now discuss how $\operatorname{ind}_\infty(E, A)$ and $\operatorname{ind}_\infty(\widehat{E}, \widehat{A})$ are related.

Since we have assumed constant E, A we may assume w.l.o.g. that

$$\alpha E - \beta A = \alpha \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} - \beta \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} \begin{matrix} \} & r \\ \} & n-r \end{matrix} \tag{4.3}$$

is in Kronecker canonical form.

Let

$$N = \begin{bmatrix} 0 & \alpha_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_{n-r-1} \\ & & & 0 \end{bmatrix}, \quad J = \begin{bmatrix} \lambda_1 & \beta_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \beta_{r-1} \\ & & & \lambda_r \end{bmatrix} \tag{4.4}$$

where $\alpha_i, \beta_j \in \{0,1\}$ $\quad i = 1, \ldots, n-r-1,\ j = 1, \ldots, r-1.$ Let $\ell = n-r-1,$ then it follows that

(4.5)

$$\widehat{E} = (E \otimes E)^T = \left[\begin{array}{cccccc|cccccccccc} I & 0 & & & & & & & & & & & & & \\ 0 & N^T & & & & & & & & & & & & & \\ & & \ddots & & & & & & & & & & & & \\ & & & I & 0 & & & & & & & & & & \\ & & & 0 & N^T & & & & & & & & & & \\ \hline & & & & & & 0 & 0 & & & & & & & \\ & & & & & & 0 & 0 & & & & & & & \\ & & & & & & \alpha_1 I & 0 & 0 & 0 & & & & & \\ & & & & & & 0 & \alpha_1 N^T & 0 & 0 & & & & & \\ & & & & & & & & \alpha_2 I & 0 & 0 & 0 & & & \\ & & & & & & & & 0 & \alpha_2 N^T & 0 & 0 & & & \\ & & & & & & & & & & \ddots & & \ddots & & \\ & & & & & & & & & & & \alpha_\ell I & 0 & 0 & 0 \\ & & & & & & & & & & & 0 & \alpha_\ell N^T & 0 & 0 \end{array}\right]$$

(4.6)

$$\widehat{A} = (E \otimes A)^T + (A \otimes E)^T = \left[\begin{array}{cccccc|cccccccccc} J^T & 0 & & & & & & & & & & & & & \\ 0 & I & & & & & & & & & & & & & \\ & & \ddots & & & & & & & & & & & & & \\ & & & J^T & 0 & & & & & & & & & & \\ & & & 0 & I & & & & & & & & & & \\ \hline & & & & & & 0 & 0 & & & & & & & \\ & & & & & & 0 & 0 & & & & & & & \\ & & & & & & \alpha_1 J^T & 0 & 0 & 0 & & & & & \\ & & & & & & 0 & \alpha_1 I & 0 & 0 & & & & & \\ & & & & & & & & \ddots & & \ddots & & & & \\ & & & & & & & & & \alpha_\ell J^T & 0 & 0 & 0 & & \\ & & & & & & & & & 0 & \alpha_\ell I & 0 & 0 & & \end{array}\right]$$

$$+\left[\begin{array}{cccccccc|cc}
\lambda_1 I & 0 & & & & & & & & \\
0 & \lambda_1 N^T & & & & & & & & \\
\beta_1 I & 0 & \lambda_2 I & 0 & & & & & & \\
0 & \beta_1 N^T & 0 & \lambda_2 N^T & & & & & & \\
 & & \ddots & & \ddots & & & & & \\
 & & & \beta_{r-1} I & 0 & \lambda_r I & 0 & & & \\
 & & & 0 & \beta_{r-1} N^T & 0 & \lambda_r N^T & & & \\
\hline
 & & & & & & & & I & 0 \\
 & & & & & & & & 0 & N^T \\
 & & & & & & & & & \ddots \\
 & & & & & & & & & I \quad 0 \\
 & & & & & & & & & 0 \quad N^T
\end{array}\right]$$

$$=\left[\begin{array}{cccccccc|cc}
J^T+\lambda_1 I & 0 & & & & & & & & \\
0 & I+\lambda_1 N^T & & & & & & & & \\
\beta_1 I & 0 & J^T+\lambda_2 I & 0 & & & & & & \\
0 & \beta_1 N^T & 0 & I+\lambda_2 N^T & & & & & & \\
 & & \ddots & & \ddots & & & & & \\
 & & & \beta_{r-1} I & 0 & J^T+\lambda_r I & 0 & & & \\
 & & & 0 & \beta_{r-1} N^T & 0 & I+\lambda_r N^T & & & \\
\hline
 & & & & & & & & I & 0 \\
 & & & & & & & & 0 & N^T \\
\alpha_1 J^T & 0 & I & 0 & & & & & & \\
0 & \alpha_1 I & 0 & N^T & & & & & & \\
 & & \ddots & & \ddots & & & & & \\
 & & & \alpha_\ell J^T & 0 & I & 0 & & & \\
 & & & 0 & \alpha_\ell I & 0 & N^T & & &
\end{array}\right]$$

Thus

$$(4.7)\quad \alpha\widehat{E}-\beta\widehat{A}=\alpha\left[\begin{array}{cccc|cccc|cccc}
I & & & & & & & & & & & \\
 & I & & & & & & & & & & \\
 & & \ddots & & & & & & & & & \\
 & & & I & & & & & & & & \\
\hline
 & & & & N^T & & & & & & & \\
 & & & & & N^T & & & & & & \\
 & & & & & & \ddots & & & & & \\
 & & & & & & & N^T & & & & \\
\hline
 & & & & & & & & 0 & & & \\
 & & & & \alpha_1 I & 0 & & & & & & \\
 & & & & & \alpha_2 I & 0 & & & & & \\
 & & & & & & \ddots & \ddots & & & & \\
 & & & & & & & \alpha_\ell I & 0 & & & \\
\hline
 & & & & & & & & 0 & & & \\
 & & & & & & & & \alpha_1 N^T & 0 & & \\
 & & & & & & & & & \ddots & \ddots & \\
 & & & & & & & & & & \alpha_\ell N^T & 0
\end{array}\right]-\beta$$

$$
\left[\begin{array}{c|c|c|c}
\begin{matrix} J^T+\lambda_1 I & & & \\ \beta_1 I & J^T+\lambda_2 I & & \\ & \ddots & \ddots & \\ & & \beta_{r-1} I & J^T+\lambda_r I \end{matrix} & & & \\ \hline
 & \begin{matrix} I+\lambda_1 N^T & & & \\ \beta_1 N^T & I+\lambda_2 N^T & & \\ & \ddots & \ddots & \\ & & \beta_{r-1} N^T & I+\lambda_r N^T \end{matrix} & & \\ \hline
 & & \begin{matrix} I & & & & \\ \alpha_1 J^T & I & & & \\ & \alpha_2 J^T & I & & \\ & & \ddots & \ddots & \\ & & & \alpha_\ell J^T & I \end{matrix} & \\ \hline
 & & & \begin{matrix} N^T & & \\ \alpha_1 I & N^T & \\ & \ddots & \ddots \\ & & \alpha_\ell I & N^T \end{matrix}
\end{array}\right]
$$

$$
= \alpha \begin{bmatrix} \hat{E}_1 & & & \\ & \hat{E}_2 & & \\ & & \hat{E}_3 & \\ & & & \hat{E}_4 \end{bmatrix} - \beta \begin{bmatrix} \hat{A}_1 & & & \\ & \hat{A}_2 & & \\ & & \hat{A}_3 & \\ & & & \hat{A}_4 \end{bmatrix} .
$$

The pencil $\alpha\widehat{E}_1 - \beta\widehat{A}_1 = \alpha I - \beta\widehat{A}_1$ is regular and $\operatorname{ind}_\infty(\widehat{E}_1, \widehat{A}_1) = 0$.

If $\operatorname{ind}_\infty(E, A) = 1$ then it follows that all $\alpha_i = 0,\ i = 1, \ldots, n-p-1,$ hence $N = 0$ and it follows that

$$\alpha\widehat{E}_2 - \beta\widehat{A}_2 = \alpha 0 - \beta I \text{ is regular, } \operatorname{ind}_\infty(\widehat{E}_2, \widehat{A}_2) = 1,$$

$$\alpha\widehat{E}_3 - \beta\widehat{A}_3 = \alpha 0 - \beta I \text{ is regular, } \operatorname{ind}_\infty(\widehat{E}_3, \widehat{A}_3) = 1,$$

$$\alpha\widehat{E}_4 - \beta\widehat{A}_4 = \alpha 0 - \beta 0 \text{ is singular,}$$

and hence in system (4.1) we obtain a consistency condition for the corresponding part of $\widehat{Q}$ and arbitrary components, which fortunately in the control application do not occur in the solution, as we have shown in section 2.

If $\operatorname{ind}_\infty(E, A) = q > 1,$ then in $\widehat{E}_2 - \beta\widehat{A}_2$ we obtain Kronecker blocks of size $q,$ and singular blocks $\alpha 0 - \beta 0,$ but we may also have other singular blocks in $\alpha\widehat{E} - \beta\widehat{A}$ as is demonstrated by the following example.

4.8 Example. *Let* $E = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},\ A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$ *then* $\operatorname{ind}_\infty(E, A) = 3.$ *In*

the vectorized form we obtain for $\alpha\widehat{E}-\beta\widehat{A}$ *the pencil*

$$(4.9)\qquad \alpha\begin{bmatrix} 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 1\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,1\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 1\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,1\,0 & 0\,0\,0 \end{bmatrix} -\beta\begin{bmatrix} 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 1\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,1\,0 & 0\,0\,0 & 0\,0\,0\\ 1\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,1\,0 & 1\,0\,0 & 0\,0\,0\\ 0\,0\,1 & 0\,1\,0 & 0\,0\,0\\ 0\,0\,0 & 1\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,1\,0 & 1\,0\,0\\ 0\,0\,0 & 0\,0\,1 & 0\,1\,0 \end{bmatrix},$$

equivalent to

$$(4.10)\qquad \alpha\begin{bmatrix} 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 1\,0\,0 & 0\,0\,0\\ 0\,0\,1 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 1\,0\,0\\ 0\,0\,0 & 0\,0\,1 & 0\,0\,0 \end{bmatrix} -\beta\begin{bmatrix} 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,1 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 1\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,1\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 1\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,1\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,1 \end{bmatrix},$$

equivalent to

$$(4.11)\qquad \alpha\begin{bmatrix} 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,1 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,1\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 1\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,1\,0 \end{bmatrix} -\beta\begin{bmatrix} 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,1 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 1\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,1\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,1 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 1\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,0\\ 0\,0\,0 & 0\,0\,0 & 0\,0\,1 \end{bmatrix},$$

which has two singular blocks $\alpha 0-\beta 0$ *, two index two blocks* $\alpha\begin{bmatrix}0&0\\1&0\end{bmatrix}-\beta\begin{bmatrix}1&0\\0&1\end{bmatrix}$ *,*
one singular block $\alpha\begin{bmatrix}0\\1\end{bmatrix}-\beta\begin{bmatrix}1\\0\end{bmatrix}$ *and one singular block* $\alpha[1\ 0]-\beta[0\ 1]$.

We can summarize the previous considerations in the following theorem

4.12 Theorem. *Consider the DARE (1.1), (1.2) and vectorize it in the form (4.1), (4.2).*

i) *If* $\alpha E-\beta A$ *is regular and* $\operatorname{ind}_\infty(E,A)=1$ *then* $\alpha\widehat{E}-\beta\widehat{A}$ *is a singular pencil with singular blocks of the form* $\alpha 0-\beta 0$ *and* $\operatorname{ind}_\infty(\widehat{E},\widehat{A})=1$.

ii) *If* $\alpha E - \beta A$ *is regular and* $\operatorname{ind}_\infty(E,A) \geq 2$, *then* $\alpha\widehat{E} - \beta\widehat{A}$ *is a singular pencil, which may have different kinds of singular blocks and also different* $\operatorname{ind}_\infty(\widehat{E},\widehat{A})$.

Note that due to the special form of the nonlinear term in (1.1), (1.2) it follows that if $\operatorname{ind}_\infty(E(t),A(t)) \leq 1$, then the quadratic part only occurs in the differential part of the vectorized system. (Compare (3.12), ... ,(3.16)!)

We have seen in section 2 that in the problems arising in optimal control, the components, corresponding to singular blocks $\alpha 0 - \beta 0$, which are arbitrary, do not occur in the final solution, thus we may simply omit them from the system. We may not do this in general with singular blocks arising from higher index blocks in $\alpha E - \beta A$. This is another reason why we restrict ourselves to index one problems. If $\operatorname{ind}_\infty(\widetilde{E},\widetilde{A}) \geq 2$ but $\alpha\widetilde{E} - \beta\widetilde{A}$ regular, then we may in principle perform an index reduction as described in Gear [G 3]. This would require further differentiability conditions and would have to be done analytically. For these reasons we consider only problems where $\operatorname{ind}_\infty(E,A) \leq 1$.

§5 Numerical methods and results.

The approach discussed in section 3 is now transformed into a numerical method. We first describe the algorithm for the autonomous case.

5.1 Algorithm DAREC.

Given $A,E,Q,W,M \in \mathbb{R}^{n,n}$ constant in t with E,A satisfying (3.4), $Q = Q^T$, $W = W^T$, $M = M^T$, t_0, $t_f \in \mathbb{R}$ and a rank tolerance $\epsilon \in \mathbb{R}$, this algorithm solves the Riccati differential equation (1.1), (1.2) in the interval $[t_0, t_f]$.

Step 1 Perform a singular value decomposition of E using the LINPACK DSVDC routine, i.e. determine $U,V \in \mathbb{R}^{n,n}$ orthogonal, such that $U^TEV = \Sigma = \operatorname{diag}\{\sigma_1,\ldots,\sigma_n\}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$.

Step 2 Determine the rank p of Σ as the largest index $i \in \{1,\ldots,n\}$ such that $\sigma_i \geq \epsilon\sigma_1$ and set $\sigma_{p+1},\ldots,\sigma_n$ to zero.

Step 3 Perform the transformation with U,V to the whole Riccati equation, i.e. set $A := U^TAV$, $Q := V^TQV$, $W := U^TWU$, $Y(t) := U^TX(t)U$, $M := V^TMV, E := \begin{bmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{bmatrix}$, where $\Sigma_0 = \operatorname{diag}\{\sigma_1,\ldots,\sigma_p\}$. Partition the matrices analogous to the partitioning in E, for example as in (3.11).

Step 4 Check the consistency conditions in (3.15), (3.16)

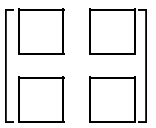
$$Q_{22} = 0,\ M_{12} = 0,\ M_{22} = 0.$$

If these conditions do not hold, the system is not solvable,

STOP

Step 5 Solve the algebraic equation (3.13) for $Y_{12}(t_f)$, i.e. for the "terminal" value t_f using the QR–decomposition with column pivoting for A_{22}, e.g. Golub/Van Loan [G 5].

$$Y_{12}(t_f) = -(Y_{11}(t_f)A_{12} + \Sigma_0^{-1}\ Q_{12})A_{22}^{-1} = -\Sigma_0^{-1}(M_{11}\Sigma_0^{-1}A_{12} + Q_{12})A_{22}^{-1}.$$

Step 6 Rewrite the relevant part of $Y(t) = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$ which is given by the shaded region in :  i.e. the upper triangle of Y_{11} including the diagonal and Y_{12}, in vector form, by going through $[Y_{11}, Y_{12}]$ rowwise.

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{22} \\ \vdots \\ y_{2n} \\ \vdots \\ y_{pp} \\ \vdots \\ y_{pn} \end{bmatrix} =: z =: \begin{bmatrix} z_1 \\ \vdots \\ \vdots \\ \vdots \\ z_\ell \end{bmatrix}, \quad \text{where} \quad \ell = \tfrac{p(2n-p+1)}{2}, \quad \text{and rewrite system (3.12), (3.13),}$$

(3.14) in terms of z. Call the new system

$$\widetilde{E}\dot{z}(t) = \widetilde{f}(t, z(t)) \tag{5.2}$$

Step 7 Solve system (5.2) by calling LIMEX (or DASSL or any other solver for differential algebraic equations).

Step 8 Rewrite the solution z as a matrix by completing the matrix $Y(t_0) = \begin{bmatrix} Y_{11}(t_0) & Y_{12}(t_0) \\ Y_{21}(t_0) & Y_{22}(t_0) \end{bmatrix}$ with 0 components in Y_{22}, and in a symmetric way in Y_{21}.

Step 9 Form $X(t_0) = U\ Y(t_0)U^T$.

In the time dependent case the algorithm is more complicated.

5.3 Algorithm DAREV.

Given $A(t), E(t), M(t), Q(t), W(t) \in \mathbb{R}^{n,n}, Q = Q^T, W = W^T, M = M^T, t_0, t_f \in \mathbb{R}$ and a rank tolerance $\epsilon \in \mathbb{R}$, with $E(t), A(t)$ satisfying (3.4), (3.5), this algorithm solves the Riccati differential algebraic equation (1.1), (1.2) in the interval $[t_0, t_f]$.

Step 1 Perform a singular value decomposition of $E(t_f)$ using the LINPACK DSVDC routine, i.e. determine $U = U(t_f),\ V(t_f) \in \mathbb{R}^{n,n}$ orthogonal, such that $U^T E(t_f) V(t_f) = \Sigma(t_f) = \mathrm{diag}\{\sigma_1(t_f), \cdots, \sigma_n(t_f)\},\ \sigma_1(t_f) \geq \cdots \geq \sigma_n(t_f) \geq 0.$

Step 2 Determine the rank p of $\Sigma(t_f)$ as the largest index $i \in \{1, \cdots, n\}$ such that $\sigma_i(t_f) \geq \epsilon\sigma_1(t_f)$ and set $\sigma_{p+1}(t_f), \cdots, \sigma_n(t_f) = 0.$

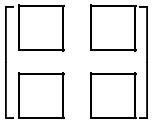
Step 3 Perform the transformations with $U,\ V(t_f)$ to $A(t_f), Q(t_f), M(t_f)$, i.e. set $A(t_f) := U^T A(t_f) V(t_f),\ Q(t_f) := V(t_f)^T Q(t_f) V(t_f),\quad M := V(t_f)^T M\ V(t_f),\ E(t_f) := \begin{bmatrix} \Sigma_0(t_f) & 0 \\ 0 & 0 \end{bmatrix}$, where $\Sigma_0(t_f) = \mathrm{diag}\{\sigma_1(t_f), \cdots, \sigma_p(t_f)\}.$

Step 4 Check the consistency of system (3.12), (3.13), (3.14), (3.15), (3.16), i.e. whether $Q_{22}(t_f) = 0, M_{12} = 0, M_{22} = 0$.

If these conditions do not hold, the system is not solvable,

STOP

Step 5 Solve the algebraic equations (3.13) for $Y_{12}(t_f)$ using the QR-decomposition with column pivoting for $A_{22}(t_f)$: $Y_{12}(t_f) = -\Sigma_0^{-1}(M_{11}\Sigma_0^{-1}(t_f)A_{12}(t_f)+Q_{12}(t_f))A_{22}(t_f)^{-1}$.

Step 6 Rewrite the relevant part of $Y(t_f) = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$, given by the shaded region in  , in vector form, by going again through $[Y_{11}\ Y_{12}]$ rowwise as in step 6 of Algorithm 5.1. Rewrite the system in terms of z, with the transformed system

$$\widetilde{E}(t)\dot{z}(t) = \widetilde{f}(t, z(t)). \tag{5.4}$$

Step 7 Solve system (5.4) with LIMEX, where

a) in the beginning of each time step, a new V_{ref} is determined as the Q factor in the QR factorization of $E^T(t_i)U$, and

b) for the evaluation of $V(t_i)$ continuous, use the algorithm described in Rheinboldt [R 2], then transform $A(t_i), Q(t_i), W(t_i), E(t_i)$ with U and this continuous $V(t_i)$ according to the formulas

$$A(t_i) := U^T A(t_i)V(t_i), \quad Q(t_i) := V(t_i)^T Q(t_i)V(t_i),$$
$$W(t_i) = U^T W(t_i)U, \quad Y(t_i) = U^T X(t_i)U, \ E(t_i) := U^T E(t_i)V(t_i)$$

and evaluate the vectorized functions $\widetilde{E}(t_i), \widetilde{f}(z(t_i))$ with these transformed matrices.

Step 8 Rewrite the solution in matrix form by completing it with 0 in position Y_{22} and in a symmetric way in Y_{21}.

Step 9 Form $X(t_0) = U\ Y(t_0)U^T$.

The above algorithms have been implemented as FORTRAN subroutines on an IBM 3090VF in double precision. In the subsequent numerical tests, the common required accuracy for LIMEX was 10^{-5} and the integration was carried out from $t_f = 1$ to $t_0 = 0$.

5.5 Example [T 1].

The following differential algebraic Riccati equation arises in optimal control of an electrical circuit with cost functional $J(x, u) = ||x||_2 + ||u||_2$.

$$n = 6, \ E = \begin{bmatrix} 0 & 0 \\ 0 & I_3 \end{bmatrix}, \ A = \begin{bmatrix} A_1 & 0 \\ A_2 & 0 \end{bmatrix}$$

where

$$A_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \text{ and } A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

$$
\begin{aligned}
Q &= \mathrm{diag}(0,0,0,1,1,0),\\
W &= \mathrm{diag}(1,0,1,0,0,0),\\
M &= 0.
\end{aligned}
$$

After 4 steps in 0.135 sec CPU-time, DAREC obtained the solution

$$
X(0) = \begin{bmatrix}
0 & 0 & 0 & x_{14} & x_{15} & 0\\
0 & 0 & 0 & x_{24} & x_{25} & 0\\
0 & 0 & 0 & x_{34} & x_{35} & 0\\
x_{14} & x_{24} & x_{34} & x_{44} & x_{45} & 0\\
x_{15} & x_{25} & x_{35} & x_{45} & x_{55} & 0\\
0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

with

$$
\begin{aligned}
-x_{14} = -x_{15} = x_{34} = x_{35} &= 0.38080\\
x_{24} = -x_{25} &= 0.50000\\
x_{44} = x_{55} &= 0.88080\\
x_{45} &= -0.11920.
\end{aligned}
$$

5.6 Example.

The following differential algebraic Riccati equation with variable coefficients was obtained by adjusting the inhomogeneity $Q(t)$.

$$
n = 5,\ E(t) = \begin{bmatrix}
2-t & 0 & 0 & t & 0\\
0 & 1 & 0 & 0 & 0\\
0 & 0 & 1 & 0 & t/2\\
0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0
\end{bmatrix},
$$

$$
A(t) = \mathrm{diag}(0,t,0,1,1),
$$

$$
Q(t) = \begin{bmatrix}
0 & 0 & q_{13} & q_{14} & q_{15}\\
0 & q_{22} & 0 & 0 & q_{25}\\
q_{13} & 0 & 0 & q_{34} & 0\\
q_{14} & 0 & q_{34} & q_{44} & q_{45}\\
q_{15} & q_{25} & 0 & q_{45} & 0
\end{bmatrix}
$$

with

$$
\begin{aligned}
q_{13} &= -(2-t)\\
q_{14} &= -(2-t)e^{-t}\\
q_{15} &= -(2-t)\frac{t}{2}\\
q_{22} &= e^{2t} + (e^t - 1)^2 - (2t+1)e^t\\
q_{25} &= 1 - e^t\\
q_{34} &= -t\\
q_{44} &= -2te^{-t}\\
q_{45} &= -\frac{t^2}{2},
\end{aligned}
$$

$$X(t) = \begin{bmatrix} 1 & 0 & t & e^{-t} & 0 \\ 0 & e^t & 0 & 0 & e^t - 1 \\ t & 0 & 0 & 0 & 0 \\ e^{-t} & 0 & 0 & 0 & 0 \\ 0 & e^t - 1 & 0 & 0 & 0 \end{bmatrix},$$

$$M = E(1)^T \; X(1) \; E(1).$$

After 6 steps in 0.315 sec CPU-time, DAREV obtained $X(0)$ with a maximum error of $1.3 \cdot 10^{-5}$.

§6 Conclusion.

We have discussed the numerical solution of matrix Riccati differential algebraic equations as they arise for example from descriptor control problems.
Under some typical conditions to the underlying pencil, we can preprocess the system, such that we can use standard packages for differential algebraic equations for the solution.

It is obvious how this approach can be carried over to the case of complex matrices $A(t), E(t), R(t), W(t), Q(t)$ thus we refrain here from presenting it.

Another problem which can usually be treated analogously, is the solution of discrete Riccati difference equations, which in the analogous case come from autonomous discrete optimal control problems. It takes the form

$$E^T X_k E = \tag{6.1}$$

$$Q + A^T X_{k+1} A - A^T X_{k+1} B (R + B^T X_{k+1} B)^{-1} B^T X_{k+1} A, \quad k = K-1, \cdots, 0.$$

$$E^T X_K E = M. \tag{6.2}$$

For each time step the solution for X_k is a purely algebraic problem of a different nature than in the continuous case. This is clearly another interesting problem to analyze. Other generalizations can also be considered in the control problem, investigating more general cost functionals or more complicated differential equations. All these are interesting topics for future research.

§7 Acknowledgement.

We thank IBM Deutschland GmbH and in particular Dr. R. Janßen for giving us the opportunity to carry out this work.

§8 References.

[A 1] M. Athans, P.L. Falb, Optimal Control, MacGraw Hill, New York, 1966.

[A 2] F.W. Arnold III, Numerical Solution of Algebraic Riccati Equations, Technical Rep. Naval Weapons Research Center, China Lake, NWCTP 652 , 1984.

[B 1] D.J. Bender, A.J. Laub, The linear–quadratic optimal regulator for descriptor systems, IEEE Trans. Autom. Control, 32, 1987, 672–68.

[C 1] S.L. Campbell, Singular Systems of Differential Equations I, Pitman, San Francisco, 1980.

[C 2] S.L. Campbell, Singular Systems of Differential Equations II, Pitman, San Francisco, 1982.

[D 1] P. Deuflhard, E. Hairer, J. Zugck, One–step and Extrapolation Methods for Differential–Algebraic Systems, Numer. Math. 51, 1987, 501–516.

[G 1] F.R. Gantmacher, Matrix Theory, Vol. I, Chelsea, New York, 1959.

[G 2] F.R. Gantmacher, Matrix Theory, Vol. II, Chelsea, New York, 1959.

[G 3] C.W. Gear, Differential–Algebraic Index Transformations , SIAM J. Sci. Stat. Comp. 9, 1988, 39–47.

[G 4] C.W. Gear, L. Petzold, Differential–Algebraic Systems and Matrix Pencils, in Matrix Pencils, B. Kågström, A. Ruhe, Eds., Springer, Berlin, 1983, 75–90.

[G 5] G.H. Golub, C.R. Van Loan, Matrix Computation, North Oxford Academic, Oxford , 1983.

[G 6] E. Griepentrog and R. März, Differential-Algebraic Equations and their Numerical Treatment, Teubner-Texte, Leipzig, 1986 .

[K 1] H.W. Knobloch, H. Kwakernaack , Lineare Kontrolltheorie , Springer, Berlin , 1985.

[L 1] P. Lancaster, M. Tismenetsky, The Theory of Matrices , Academic Press, Orlando, 1985.

[M 1] V. Mehrmann, The Linear Quadratic Control Problem: Theory and Numerical Algorithms, Habilitationsschrift, Universität Bielefeld, 1987.

[M 2] V. Mehrmann, Existence, Uniqueness and Stability of Solutions to Singular Linear Quadratic Optimal Control Problems, to appear in Lin. Alg. Appl..

[P 1] L. Petzold, A description of DASSL: A Differential–Algebraic System Solver, IMACS Trans. on Scient. Comp. Vol. 1, R.S. Stepleman, Ed., 1982.

[P 2] L. Petzold, Differential/Algebraic Equations are not ODE's, SIAM J. Sci. Stat. Comput. 3, 1982, 367–384.

[P 3] L. Petzold, Recent Development in the Numerical Solution of Differential/Algebraic Systems, Proceedings of the summer seminar on computational aspects of VLSI–design, Minneapolis, Minn. USA, 1988, .

[R 1] W.T. Reid, Riccati Differential Equations, Academic Press, New York, 1972.

[R 2] W.C. Rheinboldt, On the Computation of Multi–Dimensional Solution Manifolds of Parametrized Equations, Numer. Math. 53, 1988, 165–181.

[T 1] S. Tan, A Unified Approach for Analysis and Control of Multivariable Non–Causal Systems, Ph. D. thesis, Catholic University of Leuven, Belgium, 1987.

Keywords and phrases. Riccati differential algebraic equations, Kronecker canonical form, linear quadratic optimal control problems, singular value decompositions

AMS classification: 65L05, 93C15, 93E25, 34A34, 34A10

† Fakult"at f"ur Mathematik, Universit"at Bielefeld, Postfach 8640, D–4800 Bielefeld 1, FRG, email: UMATF108 @DBIUNI11.BITNET

‡ Fachbereich Mathematik, Universit"at Oldenburg, Postfach 2503, D–2900 Oldenburg, FRG, email: 015040 @DOLUNI1.BITNET