Numerische Mathematik Manuscript-Nr.
(will be inserted by hand later)

Preserving monotonicity in the numerical solution of Riccati differential equations ⋆

Luca Dieci and Timo Eirola

[1] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.
[2] Institute of Mathematics, Helsinki University of Technology, FIN-02150 Espoo, Finland

January 30, 1996

Subject Classifications: 65L

Key words: Riccati equations, positive definite matrices, monotonicity, symplectic Runge-Kutta schemes

Summary. Solutions of symmetric Riccati differential equations (RDEs for short) are in the usual applications positive semidefinite matrices. Moreover, in the class of semidefinite matrices, solutions of different RDEs are also monotone, with respect to properly ordered data. Positivity and monotonicity are essential properties of RDEs. In Dieci and Eirola (1994), we showed that, generally, a direct discretization of the RDE cannot maintain positivity, and be of order greater than one. To get higher order, and to maintain positivity, we are thus forced to look into indirect solution procedures. Here, we consider the problem of how to maintain monotonicity in the numerical solutions of RDEs. Naturally, to obtain order greater than one, we are again forced to look into indirect solution procedures. Still, the restrictions imposed by monotonicity are more stringent that those of positivity, and not all of the successful indirect solution procedures of Dieci and Eirola (1994) maintain monotonicity. We prove that by using symplectic Runge-Kutta (RK) schemes with positive weights (e.g.,

⋆ This work was supported in part under NSF Grant #DMS-9306412.

Gauss schemes) on the underlying Hamiltonian matrix, we eventually maintain monotonicity in the computed solutions of RDEs.

Notation. We say that a matrix A is *positive* if it is symmetric and positive definite, and *nonnegative* if it is symmetric positive semidefinite, and we write $A > 0$, and $A \geq 0$, respectively. A matrix S is *symplectic* if $S^T J S = J$, where $J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$, or – equivalently – if $S^{-1} = -JS^TJ$.

1. Introduction

In this paper we consider the problem of solving numerically the symmetric Riccati differential equation (RDE):

$$\dot{X}(t) = A(t)X(t) + X(t)A^T(t) - X(t)B(t)X(t) + C(t) \ , \ X(0) = X_0 \ , \qquad (1)$$

where all matrices are in $\mathbb{R}^{n\times n}$, bounded, piecewise continuous, and moreover, $B(t)$, $C(t)$ and X_0 are nonnegative.

This equation arises naturally in many engineering applications, e.g. in optimal control. Since the specific engineering problem provides the motivation for our study, let us briefly recall the so-called "finite time regulator problem" (Kalman (1960), Anderson and Moore (1971), Kwakernaak and Sivan (1972)). We have a linear time-varying system

$$\dot{x}(t) = F(t)x(t) + G(t)u(t) \ , \quad x(0) = x_0 \ , \qquad (2)$$

where $F(t) \in \mathbb{R}^{n\times n}$, $G(t) \in \mathbb{R}^{n\times p}$, and vectors $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^p$. We want to find the "optimal control" $u^*(t)$, which minimizes the quadratic criterion

$$\int_0^{t_f} \big(x^T(t)C(t)x(t) + u^T(t)R(t)u(t)\big)dt + x^T(t_f)X_0x(t_f), \qquad (3)$$

where $X_0 \geq 0$, $C \geq 0$, and $R > 0$. It is well known that the minimum value of the criterion is $x_0^T X(0)x_0$ (see: e.g. Anderson and Moore (1971)), where $X(t)$ solves the RDE

$$-\dot{X} = F^TX + XF - X(GR^{-1}G^T)X + C \ , t < t_f \text{ and } X(t_f) = X_0 \ . \qquad (4)$$

The optimal linear feedback control law is then $u^*(t) = -R^{-1}(t)G^T(t)X(t)x(t)$. An equivalent way to get this result (see Kwakernaak and Sivan (1972)) is to express $u^*(t)$ uniquely as $u^*(t) = -R^{-1}(t)G^T(t)p(t)$, where now $x(t)$ and $p(t)$ solve the linear Hamiltonian two-point boundary value problem (TPBVP)

$$\begin{bmatrix} \dot{p} \\ \dot{x} \end{bmatrix} = \begin{bmatrix} -F^T & -C \\ -GR^{-1}G^T & F \end{bmatrix} \begin{bmatrix} p \\ x \end{bmatrix} , \; x(0) = x_0 , \; p(t_f) - X_0 x(t_f) = 0 \; . \tag{5}$$

Notice that (1) is (4) after the time reversal $t \leftarrow t_f - t$, in which case we would have the Hamiltonian TPBVP

$$\begin{bmatrix} \dot{p} \\ \dot{x} \end{bmatrix} = H(t) \begin{bmatrix} p \\ x \end{bmatrix} , \; H(t) = \begin{bmatrix} A(t) & C(t) \\ B(t) & -A^T(t) \end{bmatrix} , \; p(0) = X_0 x(0) \, , \; x(t_f) = x_0 \, . \tag{6}$$

The matrix $H(t) \in \mathbb{R}^{2n\times 2n}(t)$ is a *Hamiltonian matrix*: $JH(t) = (JH(t))^T$, $\forall t$.

Remark 1. In theory, to get the optimal control we might bypass solving the RDE, and solve –in some way– the TPBVP (6). However, this class of TPBVPs is known to be dichotomic (see Johnson and Nerurkar (1992), Ikeda et al. (1972)), and a successful solution strategy needs to go through some form of decoupling of the solution space (Ascher et al. (1988)). The Riccati equation is an expression of this decoupling (see Lemma 1.3 below), and thus it suggests itself also as the tool for solving the TPBVP. In other words, the RDE is not only an elegant tool, but also a computationally convenient one.

The following result is well known (e.g., see Dieci and Eirola (1994))

Theorem 1. *The unique solution of (1), with $X(0) = X_0 \geq 0$, is nonnegative and exists for all $t \geq 0$. Further, if $X(s) > 0$ or $C(s) > 0$ for some $s \geq 0$, then $X(t) > 0$ for all $t > s$.*

We also have the following Lemma, easy to verify by direct substitution.

Lemma 1. *The Riccati equation (1) is obtained from the system (6), upon requiring that the change of variables*

$$T^{-1}(t) \begin{bmatrix} p \\ x \end{bmatrix} , \; T(t) = \begin{bmatrix} I & X(t) \\ 0 & I \end{bmatrix} , \tag{7}$$

induces a block lower triangular system. Moreover, let Y and Z be the solutions of

$$\begin{bmatrix} \dot{Y}(t) \\ \dot{Z}(t) \end{bmatrix} = H(t) \begin{bmatrix} Y(t) \\ Z(t) \end{bmatrix} , \qquad \begin{bmatrix} Y(0) \\ Z(0) \end{bmatrix} = \begin{bmatrix} X_0 \\ I \end{bmatrix} .$$

Then the solution of (1) is given by

$$X(t) = Y(t) Z^{-1}(t) \, . \tag{8}$$

Solutions of (1), enjoy an important monotonicity property with respect to the initial data. More precisely, we have the following theorem (first given in Reid (1970)):

Theorem 2. *Consider the Hamiltonian matrices*

$$H(t) = \begin{bmatrix} A(t) & C(t) \\ B(t) & -A^T(t) \end{bmatrix} \quad \text{and} \quad \widetilde{H}(t) = \begin{bmatrix} \widetilde{A}(t) & \widetilde{C}(t) \\ \widetilde{B}(t) & -\widetilde{A}^T(t) \end{bmatrix} \tag{9}$$

and the RDEs associated to these Hamiltonian matrices

$$\begin{aligned} \dot{X} &= A(t)X + XA^T(t) - XB(t)X + C(t)\ ,\ X(0) = X_0\ , \\ \dot{\widetilde{X}} &= \widetilde{A}(t)\widetilde{X} + \widetilde{X}\widetilde{A}^T(t) - \widetilde{X}\widetilde{B}(t)\widetilde{X} + \widetilde{C}(t)\ ,\ \widetilde{X}(0) = \widetilde{X}_0\ . \end{aligned} \tag{10}$$

Assume

$$\widetilde{H}J \le HJ \quad , \text{ i.e.,} \quad \begin{bmatrix} \widetilde{C} - C & A - \widetilde{A} \\ A^T - \widetilde{A}^T & B - \widetilde{B} \end{bmatrix} \ge 0 \tag{11}$$

together with $0 \le X_0 \le \widetilde{X}_0$ *. Then, for every* $t \ge 0$ *we have* $X(t) \le \widetilde{X}(t)$ *.*

Proof. Let $U(t) = \widetilde{X}(t) - X(t)$. Then, it suffices to realize that $U(t)$ satisfies the RDE

$$\dot{U} = (\widetilde{A} - X\widetilde{B})U + U(\widetilde{A} - X\widetilde{B})^T - U\widetilde{B}U + [\,I \quad -X\,]\,(HJ - \widetilde{H}J) \begin{bmatrix} I \\ -X \end{bmatrix}$$

with a nonnegative definite initial condidion so that the result follows at once from Theorem 1. □

Our concern in this paper is to study conditions under which the monotonicity property of Theorem 2 is maintained under discretization. More precisely: when are two numerically computed solutions of RDEs, with coefficient matrices verifying the assumptions of Theorem 2, ordered within the class of nonnegative matrices?

In the next Sections we will answer the above question. But, first, we should make clear why the properties of positivity and monotonicity expressed by Theorem 1 and Theorem 2 are important. Positivity is important chiefly for stability considerations: in the context of the regulator problem, $x^T(t)X(t)x(t)$ is a Lyapunov function (strict, if $X(t) > 0$) for the closed loop system. Moreover, in terms of cost criterion, failure to maintain positivity can lead to totally erroneous physical interpretations. Monotonicity is of utmost importance from the designer's point of view. It is exactly the ability to modify the costs and the optimal control, through modifications in the input data, which makes it so important. Moreover, our own interest in maintaining monotonicity arose from the case of a (possibly) singular weight-matrix $R(t)$ in (3), a fact which might preclude forming the RDE. In fact, in case we only have a nonnegative $R(t)$, we might regularize the problem by replacing such $R(t)$ with, say, $R(t)+\lambda I$, $\lambda > 0$. Of course, it is of interest to study the limiting case of regularization parameter going to 0. If we let X_λ be the solution of the RDE associated to a given λ, then we need to maintain

$$0 \leq X_{\lambda_1}(t) \leq X_{\lambda_2}(t)\,, \quad 0 < \lambda_1 < \lambda_2\,,$$

if there is any hope of obtaining meaningful results.

In a similar spirit to our previous work, see Dieci and Eirola (1994), in the next two Sections we look at direct and indirect discretization strategies for (1). Like in Dieci and Eirola (1994), no direct discretization can have order greater than one, and maintain monotonicity. Unlike in Dieci and Eirola (1994), amongst the indirect discretizations, we have to narrow even further the list of appropriate choices. As it turns out, to get higher order schemes, a good choice is given by symplectic RK schemes with positive weights on the underlying Hamiltonian, and recovering the solution of the RDE by (8). The main result of this paper (Thm. 6) states that for such schemes the monotonicity of the numerical solution is preserved.

2. Direct methods

In this Section, we consider *integration formulas* for the RDE: these are the one-step or multi-step methods resulting from a direct discretization of the matrix equations.

Definition 1. *We say that an integration formula* **preserves monotonicity** *for the RDE, if for any pair of systems (10) satisfying (11) there exists $h_0 > 0$, such that the method applied with stepsizes in $(0, h_0)$ produces trajectories satisfying $X_k \leq \widetilde{X}_k,\ k = 0, 1, \ldots$.*

The first result is positive, showing that there are formulas preserving monotonicity, under mild assumptions on the coefficients.

Theorem 3. *Suppose that $\widetilde{B}(t) > 0$ and $X_0 > 0$. Then, the backward Euler method preserves monotonicity for RDEs.*

Proof. By assumption, $X_0 \leq \widetilde{X}_0$. Now, suppose we have computed solutions X_k and $\widetilde{X}_k$ at the point t_k such that $X_k \leq \widetilde{X}_k$. Then, the solutions at t_{k+1} satisfy the following two algebraic Riccati equations (AREs), respectively:

$$(hA - \tfrac{1}{2}I)X_{k+1} + X_{k+1}(hA - \tfrac{1}{2}I)^T - X_{k+1}hBX_{k+1} + (hC + X_k) = 0\,,$$
$$(h\widetilde{A} - \tfrac{1}{2}I)\widetilde{X}_{k+1} + \widetilde{X}_{k+1}(h\widetilde{A} - \tfrac{1}{2}I)^T - \widetilde{X}_{k+1}h\widetilde{B}\widetilde{X}_{k+1} + (h\widetilde{C} + \widetilde{X}_k) = 0\,,$$

where $h = t_{k+1} - t_k$ and the matrices $A, B, C, \widetilde{B}, \widetilde{C}$ are evaluated at t_{k+1}. These AREs are associated to the two following constant Hamiltonian matrices

$$M_k := \begin{bmatrix} hA - \frac{1}{2}I & hC + X_k \\ hB & -(hA - \frac{1}{2}I)^T \end{bmatrix}\,, \quad \widetilde{M}_k := \begin{bmatrix} h\widetilde{A} - \frac{1}{2}I & h\widetilde{C} + \widetilde{X}_k \\ h\widetilde{B} & -(h\widetilde{A} - \frac{1}{2}I)^T \end{bmatrix}\,.$$

Under the stated assumptions, then, it is known (e.g., see Theorem 2.2 of Ran and Vreughdenil (1988)) that we can uniquely obtain $X_{k+1} > 0$ and $\widetilde{X}_{k+1} > 0$, and that moreover $X_{k+1} \leq \widetilde{X}_{k+1}$. □

Remark 2. In general, solutions of AREs are not unique. Thus, the above theorem really says that, if we always select the positive solution of the AREs arising during backward Euler discretization, we do eventually get monotonicity.

Remark 3. We can weaken the positivity assumptions on $\widetilde{B}$ and X_0 in case the RDE arises –as it is the case for us– from the control setting. In this case, from (5)-(6), we have that $B = GR^{-1}G^T$, and similarly for $\widetilde{B}$. Then, the above Theorem still holds if we require $X_0 \geq 0$, and $(\widetilde{A}(t), \widetilde{G}(t))$ stabilizable (see Ran and Vreughdenil (1988)).

Although Theorem 3 does not impose restrictions on the stepsize, it only gives us a first order method. In general, this is optimal.

Theorem 4. *Any one-step method or strictly stable multistep method that preserves monotonicity in the numerical solution of RDEs has order at most one.*

Proof. This is a direct consequence of Theorem 2.3 in Dieci and Eirola (1994). There, we proved that, under the stated assumptions, solutions of the RDE (1) cannot be guaranteed to be positive. Thus, compared to the case $C = 0$, $X(t) = X_0 = 0$ they cannot be monotone. □

Remark 4. It should be noticed (see Dieci and Eirola (1994)) that the negative result of Theorem 4 cannot be improved even in the case of Lyapunov equations, that is when the quadratic term is missing:

$$\dot{X} = A(t)X + XA^T(t) + C(t)\,, \quad X(0) = X_0\,. \tag{12}$$

3. Indirect methods

In our previous work on positivity preservation, we chiefly focused on two indirect solution techniques for RDEs. One of them is based on Lemma 1, the other is based on a linearization approach. Let us first recall the latter one.

3.1. Linearization approach

The starting point, here, is an explicit solution formula for the Lyapunov Equation (12). In fact, it is easy to see that in this case the solution $X(t)$ satisfies for $t \geq s \geq 0$

$$X(t) = \Phi(t,s)X(s)\Phi(t,s)^T + \int_s^t \Phi(t,\tau)C(\tau)\Phi(t,\tau)^T d\tau \ , \tag{13}$$

where Φ solves

$$\partial_t \Phi(t,\tau) = A(t)\Phi(t,\tau) \ , \qquad \Phi(\tau,\tau) = I \ . \tag{14}$$

It is trivial to maintain positivity, here: one can use any quadrature rule with positive weights on (13), by supplying the values for Φ upon integrating (14) with any other rule.

If we consider another Lyapunov equation with $C(t) \leq \widetilde{C}(t)$, and $X_0 \leq \widetilde{X}_0$, it is again a trivial matter to maintain monotonicity, and order higher than one. Many choices, along the lines of the following example, are possible.

Example 1 (Second order rule). Use the implicit midpoint rule to perform the integration in (14), and the trapezoidal rule for the quadrature on (13). If we let Φ_{i+1} to be our approximation to Φ at t_{i+1}, we get

$$\begin{aligned} X_{i+1} &= \Phi_{i+1} X_i \Phi_{i+1}^T + \tfrac{h}{2}(\Phi_{i+1} C_i \Phi_{i+1}^T + C_{i+1}) \,, \\ \widetilde{X}_{i+1} &= \Phi_{i+1} \widetilde{X}_i \Phi_{i+1}^T + \tfrac{h}{2}(\Phi_{i+1} \widetilde{C}_i \Phi_{i+1}^T + \widetilde{C}_{i+1}) \,, \end{aligned}$$

from which monotonicity is obvious.

In Dieci and Eirola (1994), we noticed that there were several possibilities for RDEs, of algorithms which maintained positivity. The starting point was the following representation of solutions of (1):

$$\begin{aligned} &X(t) = \Phi(t,s)X(s)\Phi(t,s)^T + \int_s^t \Phi(t,\tau)C(\tau)\Phi(t,\tau)^T d\tau \ , \\ &\dot{\Phi}(t,\tau) = [A(t) - \tfrac{1}{2}X(t)B(t)]\Phi(t,\tau) \ , \ \Phi(\tau,\tau) = I \ . \end{aligned} \tag{15}$$

From this, we realized that many rules – also explicit in X – could be devised in order to maintain positivity. Eventually, we focused on the following formula (it is a misprint-free version of Dieci and Eirola (1994), (3.13)):

$$\begin{aligned}
\Phi_{i+\frac{1}{2}} &= [I - \tfrac{h}{4}(A_i - \tfrac{1}{2}X_iB_i)]^{-1}[I + \tfrac{h}{4}(A_i - \tfrac{1}{2}X_iB_i)]\,, \\
X_{i+\frac{1}{2}} &= \Phi_{i+\frac{1}{2}}[X_i + \tfrac{h}{4}C_i]\Phi^T_{i+\frac{1}{2}} + \tfrac{h}{4}C_{i+\frac{1}{2}}\,, \\
\Phi_{i+1} &= [I - \tfrac{h}{2}(A_{i+\frac{1}{2}} - \tfrac{1}{2}X_{i+\frac{1}{2}}B_{i+\frac{1}{2}})]^{-1}[I + \tfrac{h}{2}(A_{i+\frac{1}{2}} - \tfrac{1}{2}X_{i+\frac{1}{2}}B_{i+\frac{1}{2}})]\,, \\
X_{i+1} &= \Phi_{i+1}[X_i + \tfrac{h}{2}C_i]\Phi^T_{i+1} + \tfrac{h}{2}C_{i+1}\,.
\end{aligned}$$

Unfortunately, this same rule generally does not maintain monotonicity. To see it, consider the scalar RDE with $A = C = 0$, $X_0 = 1$, and B a (positive) scalar; then, use this above discretization with $h = 1$, and observe that the function $X_1(B)$ ceases being decreasing around $B = 1.9$.

In fairness, it is possible that there are schemes resulting from an appropriate discretization of (15), which do maintain monotonicity. However, the basic appeal of being able to easily construct simple ones, of arbitrarily high order, is certainly lost.

3.2. Fundamental solution approach

The idea, here, is to use Lemma 1. That is, to integrate

$$\begin{bmatrix} \dot{Y} \\ \dot{Z} \end{bmatrix} = \begin{bmatrix} A & C \\ B & -A^T \end{bmatrix} \begin{bmatrix} Y \\ Z \end{bmatrix}\,, \quad \begin{bmatrix} Y(0) \\ Z(0) \end{bmatrix} = \begin{bmatrix} X_0 \\ I \end{bmatrix} \tag{16}$$

and then form $X(t)$ via (8): $X(t) = Y(t)Z(t)^{-1}$.

Of course, this raises the question of how to integrate (16). Consider a k-stage RK scheme, compactly represented by the tableau

$$\begin{array}{c|ccc}
c_1 & a_{11} & \dots & a_{1k} \\
\vdots & \vdots & \dots & \vdots \\
c_k & a_{k1} & \dots & a_{kk} \\
\hline
 & b_1 & \dots & b_k
\end{array}\,,$$

where the c_j are the abscissas, and the b_j are the weights. A RK scheme is called *symplectic* if

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0\,, \quad i, j = 1, \dots, k\,,$$

and we will consider symplectic schemes with positive weights b_i. The Gauss schemes are such.

In Dieci and Eirola (1994), we showed that if we use these RK schemes to integrate (16), then eventually the end result is a nonnegative approximation for X. This result, as well as many other results on integrating certain systems with such RK schemes, can be actually obtained as consequences of a general monotonicity preserving property of these methods (see Eirola (1995), Theorem 2.1):

A map $q\,:\,\mathbb{R}^d \to E$ is *quadratic* if it is of the form $q(x)=\beta(x,x)\,,\ x\in\mathbb{R}^d$, where β is bilinear. Let K be a closed convex cone in E and write $u\geq v$ if $u-v\in K$.

We say that q is *nondecreasing along trajectories* of an ODE

$$\dot{x}(t)=f(t,x(t))\,,\qquad x(t)\in\mathbb{R}^d$$

if for a solution on any interval $[\tau,\tau']$ holds: $q(x(\tau))\leq q(x(\tau'))$.

Theorem 5. *If a quadratic map is nondecreasing along trajectories of an ordinary differential equation, then it is also nondecreasing, at grid-points, along the numerical trajectories — if defined — obtained from a symplectic RK scheme with positive weights.*

Remark 5. To apply the above theorem, one needs to find an appropriate quadratic form, and show that this form is nondecreasing along exact trajectories. Of course, this might be hard, but the importance of the theorem is that it allows to work with the continuous problem, which is much easier (we have differentiability!). Conclusions are then reached on the mesh where we have discretized the problem.

The following is our main result.

Theorem 6. *Consider the following two Hamiltonian systems*

$$\begin{aligned}
\frac{d}{dt}\begin{bmatrix} Y(t) \\ Z(t)\end{bmatrix} &= \begin{bmatrix} A(t) & C(t) \\ B(t) & -A^T(t)\end{bmatrix}\begin{bmatrix} Y(t) \\ Z(t)\end{bmatrix}, \quad \begin{bmatrix} Y(0) \\ Z(0)\end{bmatrix}=\begin{bmatrix} X_0 \\ I\end{bmatrix}, \\
\frac{d}{dt}\begin{bmatrix} \widetilde{Y}(t) \\ \widetilde{Z}(t)\end{bmatrix} &= \begin{bmatrix} \widetilde{A}(t) & \widetilde{C}(t) \\ \widetilde{B}(t) & -\widetilde{A}^T(t)\end{bmatrix}\begin{bmatrix} \widetilde{Y}(t) \\ \widetilde{Z}(t)\end{bmatrix}, \quad \begin{bmatrix} \widetilde{Y}(0) \\ \widetilde{Z}(0)\end{bmatrix}=\begin{bmatrix} \widetilde{X}_0 \\ I\end{bmatrix}, \qquad (17)
\end{aligned}$$

and assume that the coefficients and initial conditions satisfy

$$\widetilde{H}(t)J\leq H(t)J \qquad \text{and} \qquad 0\leq X_0\leq \widetilde{X}_0\,.$$

Let Y_k, Z_k, $\widetilde{Y}_k$, and $\widetilde{Z}_k$ be the approximations obtained — for the same stepsize sequence — by using a symplectic RK scheme with positive weights on (17), and *let* $X_k = Y_k Z_k^{-1}$, $\widetilde{X}_k = \widetilde{Y}_k \widetilde{Z}_k^{-1}$. *Then, as long as both are defined, we have* $X_k \le \widetilde{X}_k$.

Idea of Proof. By Dieci and Eirola (1994) we know that X_k and $\widetilde{X}_k$ are symmetric and positive semidefinite. First, we will show the result when we have different ICs $X_0 \le \widetilde{X}_0$, but the same coefficient matrices, and then in the case of same ICs, but coefficient matrices satisfying the inequalities. Combining these gives then the general case.

The main tool we will use is Theorem 5 in our context: any inequality (or equality) that is obtained from a monotonicity of a quadratic map along trajectories of (17) will automatically be satisfied also by numerical solutions obtained from the RK–scheme. □

Consider systems

$$\dot{S}(t) = H(t)S(t) , \quad S(0) = I , \qquad \dot{\widetilde{S}}(t) = \widetilde{H}(t)\widetilde{S}(t) , \quad \widetilde{S}(0) = I , \tag{18}$$

together with their RK–discretizations, where H and $\widetilde{H}$ are defined in (9) (i.e., the coefficient matrices of (17)). Then, by linearity, the true solutions and corresponding numerical ones of (17) satisfy

$$\begin{bmatrix} Y \\ Z \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} X_0 \\ I \end{bmatrix} , \quad \begin{bmatrix} \widetilde{Y} \\ \widetilde{Z} \end{bmatrix} = \begin{bmatrix} \widetilde{S}_{11} & \widetilde{S}_{12} \\ \widetilde{S}_{21} & \widetilde{S}_{22} \end{bmatrix} \begin{bmatrix} \widetilde{X}_0 \\ I \end{bmatrix} .$$

Recall that $S(t)$ and $\widetilde{S}(t)$ are symplectic matrices, and that also their numerical approximations (obtained with symplectic RK schemes) are symplectic matrices (see Sanz-Serna (1992)).

Lemma 2. *The matrix S of (18) satisfies: $S_{21}(t)S_{22}^T(t) \ge 0$ for all $t \ge 0$. Moreover, if we discretize (18) with a symplectic RK scheme with positive weights, then the approximate matrix, if defined, satisfies this at the grid-points.*

Proof. Set $Q = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$. We have

$$\frac{d}{dt}(S^T QS) = S^T(H^T Q + QH)S = 2S^T \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix} S \ge 0 ,$$

and since $S(0)^T QS(0) = Q$, we have $S^T QS \ge Q$ for all $t \ge 0$. Symplecticity of S means $S^{-1}J = JS^T$, so that from $Q \ge S^{-T}QS^{-1}$, we get

$$-Q = J^T QJ \ge J^T S^{-T} QS^{-1} J = SJ^T QJS^T = -SQS^T ,$$

which further implies

$$2S_{21}S_{22}^T = [\,0 \quad I\,]\,SQS^T \begin{bmatrix} 0 \\ I \end{bmatrix} \geq [\,0 \quad I\,]\,Q \begin{bmatrix} 0 \\ I \end{bmatrix} = 0\ .$$

The statement for the numerical solutions follows by using first the quadratic map $q(S) = S^TQS$ in Theorem 5. to get $q(S_k) \geq q(I)$ and then proceeding as above (we know that also S_k is symplectic). □

Remark 6. Also the matrices $S_{11}^T S_{21}$, $S_{22}^T S_{12}$, and $S_{11}S_{12}^T$ can similarly be shown to be nonnegative, but below we will need only $S_{21}S_{22}^T \geq 0$.

We are ready to consider monotonicity with respect to the ICs only.

Proposition 1. *In the situation of Theorem 6, let the coefficient matrices satisfy* $\widetilde{H} = H$, *but* $X_0 \leq \widetilde{X}_0$. *Then, as long as both are defined, we have* $X_k \leq \widetilde{X}_k$.

Proof. We have

$$\begin{aligned} \tfrac{d}{dt}(Z^T\widetilde{Y} - Y^T\widetilde{Z}) = \quad & (Y^TB - Z^TA)\widetilde{Y} + Z^T(A\widetilde{Y} + C\widetilde{Z}) \\ & -(Y^TA^T + Z^TC)\widetilde{Z} - Y^T(B\widetilde{Y} - A^T\widetilde{Z}) = 0\,. \end{aligned}$$

Hence the quadratic form $Z^T\widetilde{Y} - Y^T\widetilde{Z}$ is constant on the trajectories of the system (17). That is,

$$Z^T\widetilde{Y} - Y^T\widetilde{Z} = \Delta := \widetilde{X}_0 - X_0\,. \tag{19}$$

Since the RK-method applied to the pair of equations (17) is equivalent to the application to each of them separately, we have (19) also for the numerical solutions by Theorem 5.

On the other hand we have now $\widetilde{S} = S$, so that for the true and numerical solutions we have:

$$\widetilde{Z} - Z = S_{21}(\widetilde{X}_0 - X_0) = S_{21}Z^TZ^{-T}\Delta = (S_{21}X_0S_{21}^T + S_{21}S_{22}^T)Z^{-T}\Delta\,.$$

Using this, (19), and symmetry of X we get:

$$\begin{aligned} \widetilde{Z}^T(\widetilde{X} - X)\widetilde{Z} \quad &= \quad \widetilde{Z}^TZ^{-T}(Z^T\widetilde{Y} - Y^T\widetilde{Z}) = \widetilde{Z}^TZ^{-T}\Delta = \\ &= \quad \Delta + (\widetilde{Z} - Z)^TZ^{-T}\Delta = \\ &= \quad \Delta + \Delta^TZ^{-1}(S_{21}X_0S_{21}^T + S_{22}S_{21}^T)Z^{-T}\Delta\,, \end{aligned}$$

which is nonnegative by Lemma 2. □

It remains to prove monotonicity with respect to the coefficients only, while keeping the same initial conditions. We will use the following result.

Lemma 3. *Let S and $\widetilde{S}$ be the symplectic fundamental solution matrices in (18), $W \in \mathbb{R}^{2n\times 2n}$ any constant matrix, and Q and J be as before. Then, we have*

$$(\widetilde{S}-SW)^TQ(\widetilde{S}-SW)+\widetilde{S}^TJSW-W^TS^TJ\widetilde{S} \geq (I-W^T)Q(I-W)+JW-W^TJ\,. \tag{20}$$

The same inequalities hold also for numerical solutions obtained from a symplectic RK scheme with positive weights.

Proof. Set $\widehat{Q} = \begin{bmatrix} Q & J-Q \\ J^T-Q & Q \end{bmatrix}$. The claims follow (again via Theorem 5) from the monotonicity:

$$\frac{d}{dt}\Big(\begin{bmatrix} \widetilde{S} \\ SW \end{bmatrix}^T \widehat{Q} \begin{bmatrix} \widetilde{S} \\ SW \end{bmatrix}\Big) = \begin{bmatrix} \widetilde{S} \\ SW \end{bmatrix}^T \Big(\begin{bmatrix} \widetilde{H}^T & 0 \\ 0 & H^T \end{bmatrix}\widehat{Q}+\widehat{Q}\begin{bmatrix} \widetilde{H} & 0 \\ 0 & H \end{bmatrix}\Big)\begin{bmatrix} \widetilde{S} \\ SW \end{bmatrix} \geq 0 \tag{21}$$

because

$$\begin{bmatrix} \widetilde{H}^T & 0 \\ 0 & H^T \end{bmatrix}\widehat{Q}+\widehat{Q}\begin{bmatrix} \widetilde{H} & 0 \\ 0 & H \end{bmatrix} =$$
$$2\begin{bmatrix} \widetilde{B} & 0 & -\widetilde{B} & 0 \\ 0 & C & 0 & -C \\ -\widetilde{B} & 0 & \widetilde{B} & 0 \\ 0 & -C & 0 & C \end{bmatrix} + 2\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \widetilde{C}-C & \widetilde{A}-A & 0 \\ 0 & \widetilde{A}^T-A^T & B-\widetilde{B} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

is positive semidefinite. Hence

$$\begin{bmatrix} \widetilde{S} \\ SW \end{bmatrix}^T \widehat{Q} \begin{bmatrix} \widetilde{S} \\ SW \end{bmatrix} \geq \begin{bmatrix} I \\ W \end{bmatrix}^T \widehat{Q} \begin{bmatrix} I \\ W \end{bmatrix},$$

which is equivalent to (20). □

Remark 7. We have (20) also for any time dependent W : just consider (21) with $S(t)$, $\widetilde{S}(t)$ and $W(\tau)$ and then apply the result with $\tau = t$.

The following result then deals with monotonicity with respect to the coefficients only.

Proposition 2. *In the situation of Theorem 6, let the coefficient matrices $H(t)$, $\widetilde{H}(t)$ satisfy the inequalities there, but now $\widetilde{X}_0 = X_0$. Then, as long as both are defined, we have $X_k \leq \widetilde{X}_k$.*

Proof. Omitting the subscripts "k" take

$$W = \begin{bmatrix} X_0 \\ I \end{bmatrix} Z^{-1}\widetilde{Z}Z^{-1}\,[\, S_{21} \quad S_{22}\,]\,.$$

Then we have:

$$W\begin{bmatrix} X_0 \\ I \end{bmatrix} = \begin{bmatrix} X_0 \\ I \end{bmatrix} Z^{-1}\widetilde{Z} \qquad \text{and} \qquad SW\begin{bmatrix} X_0 \\ I \end{bmatrix} = \begin{bmatrix} YZ^{-1}\widetilde{Z} \\ \widetilde{Z} \end{bmatrix},$$

so that

$$(\widetilde{S} - SW)\begin{bmatrix} X_0 \\ I \end{bmatrix} = \begin{bmatrix} \widetilde{Y} - YZ^{-1}\widetilde{Z} \\ 0 \end{bmatrix},$$

which further implies

$$[\, X_0 \quad I\,]\,(\widetilde{S} - SW)^T Q(\widetilde{S} - SW)\begin{bmatrix} X_0 \\ I \end{bmatrix} = 0\,,$$

where Q is defined as before. Thus, from (20) we get:

$$\begin{aligned}
&\begin{bmatrix} \widetilde{Y} \\ \widetilde{Z} \end{bmatrix}^T J \begin{bmatrix} YZ^{-1}\widetilde{Z} \\ \widetilde{Z} \end{bmatrix} - \begin{bmatrix} YZ^{-1}\widetilde{Z} \\ \widetilde{Z} \end{bmatrix}^T J \begin{bmatrix} \widetilde{Y} \\ \widetilde{Z} \end{bmatrix} \geq \\
&\geq (I - Z^{-1}\widetilde{Z})^T\,[\, X_0 \quad I\,]\,Q\begin{bmatrix} X_0 \\ I \end{bmatrix}(I - Z^{-1}\widetilde{Z}) + \\
&+ [\, X_0 \quad I\,]\,J\begin{bmatrix} X_0 \\ I \end{bmatrix} Z^{-1}\widetilde{Z} - \widetilde{Z}^T Z^{-T}\,[\, X_0 \quad I\,]\,J\begin{bmatrix} X_0 \\ I \end{bmatrix},
\end{aligned}$$

i.e.,

$$\widetilde{Y}^T\widetilde{Z} - \widetilde{Z}^T YZ^{-1}\widetilde{Z} - \widetilde{Z}^T Z^{-T} Y^T\widetilde{Z} + \widetilde{Z}^T\widetilde{Y} \geq 2(I - Z^{-1}\widetilde{Z})^T X_0 (I - Z^{-1}\widetilde{Z}) \geq 0\,.$$

By symmetry of X and $\widetilde{X}$ we finally get

$$X = YZ^{-1} \leq \widetilde{Y}\widetilde{Z}^{-1} = \widetilde{X}\,. \qquad \square$$

Completion of the proof of Theorem 6. Let $V_0 = \begin{bmatrix} X_0 \\ I \end{bmatrix}$, and $\widetilde{V}_0 = \begin{bmatrix} \widetilde{X}_0 \\ I \end{bmatrix}$. Let V, $\widehat{V}$, $\widetilde{V}$ be defined as follows

$$\dot{V} = H(t)V,\ V(0) = V_0, \quad \dot{\widehat{V}} = \widetilde{H}(t)\widehat{V},\ \widehat{V}(0) = V_0, \quad \dot{\widetilde{V}} = \widetilde{H}(t)\widetilde{V},\ \widetilde{V}(0) = \widetilde{V}_0\,.$$

Let X_k, $\widehat{X}_k$, and $\widetilde{X}_k$ be as usual; e.g., $\widehat{X}_k = \widehat{Y}_k\widehat{Z}_k^{-1}$, where $\widehat{V}_k = \begin{pmatrix} \widehat{Y}_k \\ \widehat{Z}_k \end{pmatrix}$ is the approximation to $\widehat{V}(t_k)$. Then, upon applying Propositions 2 and 1, in the order, we have

$$X_k \leq \widehat{X}_k \leq \widetilde{X}_k\,. \qquad \square$$

4. Conclusions

In this work, we have investigated how to maintain monotonicity in the numerical solution of Riccati equations. After realizing that no direct discretization can maintain monotonicity and have order greater than one, we paid attention to indirect solution procedures. Our main result shows that monotonicity is maintained if we integrate the underlying linear Hamiltonian system with a symplectic RK scheme with positive weights, and then recover the solution of the RE. To obtain our result we use the general property of these schemes: they maintain the monotonicity of quadratic forms (see Eirola (1995)).

In our mind, a major outcome of this work, and of Dieci and Eirola (1994), is that it once more shows that the Gauss schemes have very desirable mathematical properties. However, it remains a challenging task to implement them (or their approximation) in an efficient way.

In future work, we plan to address some of these aspects, as well as to complete the analysis for the case of singular weight matrix $R(t)$ in (3), a fact which provided our initial motivation.

References

Anderson, B. D. O., and Moore, J. B. (1971): *Linear Optimal Control.* Prentice-Hall, Englewood Cliffs.

Ascher, U., Mattheij, R. M., and Russell, R. D.(1988): *Solution of Boundary Value Problems for ODEs.* Prentice-Hall, Englewood Cliffs, N.J.

Dieci, L., and Eirola, T. (1994) Positive definiteness in the numerical solution of Riccati differential equations. *Numer. Math.*, **67**, 303–313.

Eirola, T. (1995): Monotonicity of quadratic forms with symplectic Runge–Kutta methods. *Applied Numerical Mathematics*, **17**, 293-298.

Ikeda, M., Maeda, H., and Kodama, S. (1972): Stabilization of linear systems. *SIAM J. Control*, **10**, 716–729.

Johnson, R., and Nerurkar, M. (1992): Stabilization and linear regulator problem for linear non-autonomous control processes. *preprint.*

Kalman, R. E. (1960): Contributions to the theory of optimal control. *Bol. Soc. Mat. Mexicana*, **5**, 102–119.

Kwakernaak, U., and Sivan, R. (1972): *Linear Optimal Control Systems.* Wiley-Interscience.

Ran, A. C., and Vreughdenil, R. (1988): Existence and comparison theorems for ares for continuous and discrete-time systems. *Linear Alg. & Its Applics.*, **99**, 63–83.

Reid, W. T. (1970): Monotoneity properties of solutions of Hermitian Riccati differential equations. *SIAM J. Math. Analysis*, **1**, 195–213.

Sanz-Serna, J. M. (1992): Symplectic integrators for Hamiltonian problems: An overview. *Acta Numerica*, **1**, 243–286.

This article was processed by the author using the LaTeX style file *cljour1* from Springer-Verlag.