Numer. Math. **67**, 303-313 (1994)

POSITIVE DEFINITENESS IN THE NUMERICAL SOLUTION OF RICCATI DIFFERENTIAL EQUATIONS

LUCA DIECI AND TIMO EIROLA

ABSTRACT. In this work we address the issue of integrating symmetric Riccati and Lyapunov matrix differential equations. In many cases – typical in applications – the solutions are positive definite matrices. Our goal is to study when and how this property is maintained for a numerically computed solution.

There are two classes of solution methods: *direct* and *indirect* algorithms. The first class consists of the schemes resulting from direct discretization of the equations. The second class consists of algorithms which recover the solution by exploiting some special formulae that these solutions are known to satisfy.

We show first that using a direct algorithm – a one-step scheme or a strictly stable multistep scheme (explicit or implicit) – limits the order of the numerical method to one if we want to guarantee that the computed solution stays positive definite. Then we show two ways to obtain positive definite higher order approximations by using indirect algorithms. The first is to apply a symplectic integrator to an associated Hamiltonian system. The other uses stepwise linearization.

1. INTRODUCTION

Consider the task to solve numerically the symmetric matrix Riccati equation

$$\dot{X}(t) = A(t)X(t) + X(t)A(t)^T - X(t)B(t)X(t) + C(t) \ , \tag{1.1}$$

in $\mathbf{R}^{d\times d}$ with symmetric positive semidefinite initial condition $X(0) = X_0$. We assume that the coefficients are bounded, real, piecewise continuous, and that the matrices B and C are symmetric and positive semidefinite.

Here we say that a matrix is *positive* if it is positive definite and *nonnegative* if it is positive semidefinite.

Together with (1.1) we will also consider a special case of it, namely the Lyapunov equation

$$\dot{X}(t) = A(t)X(t) + X(t)A(t)^T + C(t) \ , \tag{1.2}$$

i.e., (1.1) with $B = 0$. The following proposition states a central property of equations (1.1) and (1.2). Often it is proved via an optimization problem (e.g. [1]). As an introduction we prove it here using ideas that are applied also in section 3.

Proposition 1.1. *The solution of* (1.1) *exists and is symmetric and nonnegative for all* $t \geq 0$. *Furher, if* $X(s)$ *or* $C(s)$ *is positive for some* $s \geq 0$, *then* $X(t)$ *is positive for all* $t > s$.

This research was supported by NSF Grant DMS-9104564 and by the Academy of Finland.

Proof. Since the right-hand side of (1.1) is symmetric for symmetric $X(t)$ the solution will stay in the set of symmetric matrices.

Direct substitution shows that the solution of the Lyapunov equation (1.2) satisfies for $t \geq s \geq 0$

$$X(t) = \Phi(t,s)X(s)\Phi(t,s)^T + \int_s^t \Phi(t,\tau)C(\tau)\Phi(t,\tau)^T d\tau \ , \tag{1.3}$$

where Φ is the solution of

$$\partial_t \Phi(t,\tau) = A(t)\Phi(t,\tau)\ , \qquad \Phi(\tau,\tau) = I\ . \tag{1.4}$$

The claims for the Lyapunov equation follow now from the fact that $\Phi(t,\tau)$ is nonsingular for all t, τ.

Consider, then, the Riccati equation (1.1). Set $Y(t) := \frac{1}{2}X(t)$. Since $Y(t)$ is symmetric we get

$$\dot{X}(t) = [A(t) - Y(t)B(t)]X(t) + X(t)[A(t) - Y(t)B(t)]^T + C(t)\ , \tag{1.5}$$

i.e. X is the solution of a Lyapunov equation. So, the solution is nonnegative as long as it exists. Since for symmetric nonnegative matrices holds $||X|| = \max_{||\xi||=1} \xi^T X \xi$ we get from

$$X(t) = X(0) + \int_0^t [A(\tau)X(\tau) + X(\tau)A(\tau)^T - X(\tau)B(\tau)X(\tau) + C(\tau)]d\tau\ ,$$

the inequality

$$||X(t)|| \leq ||X(0)|| + \int_0^t [2||A(\tau)||||X(\tau)|| + ||C(\tau)||]d\tau\ ,$$

from which it follows by the Gronwall's inequality that $||X(t)||$ is finite and hence $X(t)$ exists for all $t > 0$. $\square$

The question addressed in this paper is whether also a discrete version of Proposition 1.1 is true. More precisely: does the numerical solution method preserve positivity of the solution?

The plan of this paper is as follows. In Section 2 we consider *direct* solution algorithms for solving (1.1) and (1.2). These are the methods which result from discretizing the differential equations directly, with a one-step or a multistep method. We show that any one of these discretizations, if it must preserve positivity, has order at most one. This is clearly a severe restriction, which prompted us to consider *indirect* solution algorithms. These are considered in Section 3. In particular, we look at two approaches. The *fundamental solution* method relates the solution of the Riccati equation to the solution of a linear Hamiltonian system. We show that if one uses a symplectic Runge-Kutta integrator for the latter, then positivity of the solution of the Riccati equation will be preserved. The *linearization* method results from using Proposition 1.1 and formula (1.3) for solutions of Lyapunov equations, along with a linearization procedure for Riccati equations. Some other approaches are also shortly discussed.

2. Direct methods

In this Section, we consider those methods resulting from a direct discretization of the matrix equations using one-step or multistep formulas. We call *integration formulas* the formulas obtained in these cases.

It is straightforward to use direct algorithms for solving these equations. Preserving symmetry is equally straightforward. However, even a possibly very accurate solution does not necessarily stay positive, since positivity is not a quantitative measure embedded into errors' control. To check whether a computed solution stays positive, we make use of the following definition.

Definition 2.1. We say that an integration formula *preserves positivity* for the Riccati (resp. Lyapunov) equation if for any equation of type (1.1) (resp. (1.2)) and positive definite X_0 there exists $h_0 > 0$ such that the method applied with any $h \in (0, h_0)$ produces a trajectory of positive matrices. For multistep (k-step) methods the extra starting values $X(h), \ldots, X((k-1)h)$ may be selected arbitrarily to the order of the method.

The following fact, first proved in [4], tells us that there are formulas which do preserve positivity for Riccati equations.

Proposition 2.1. *Assume that $B(t)$ in (1.1) and that X_0 are positive. Then, the backward Euler method preserves positive definiteness for Riccati equations.*

Proof. The integration formula in this case is

$$(2.1)\ \frac{1}{h}(X_{n+1} - X_n) = AX_{n+1} + X_{n+1}A^T - X_{n+1}BX_{n+1} + C, \quad n = 0, 1, \ldots \ .$$

We can interpret (2.1) as an algebraic Riccati equation (ARE) associated with the following constant coefficients Hamiltonian matrix (where all matrix blocks of (1.1) are evaluated at t_{n+1})

$$\hat{H}_{n+1} := \begin{bmatrix} -(hA - \frac{1}{2}I)^T & hB \\ hC + X_n & hA - \frac{1}{2}I \end{bmatrix} .$$

That is, (2.1) can be interpreted as resulting from the requirement that the transformation

$$\begin{bmatrix} I & 0 \\ -X_{n+1} & I \end{bmatrix} \hat{H}_{n+1} \begin{bmatrix} I & 0 \\ X_{n+1} & I \end{bmatrix}$$

produce a block upper triangular matrix. But then, it is known that we can uniquely obtain $X_{n+1} > 0$ (e.g., see [9]). $\square$

It is interesting that Proposition 2.1 poses no restriction on the stepsize h. The main drawback, of course, is that the order of the integration formula is only one. Unfortunately, this is just the best one can have. In fact, next we show that no integration formula can be expected to preserve positivity and also have order greater than one. This result is reminiscent of that in [2], where componentwise positivity of solutions of linear systems is considered. There is a big difference, however, between positivity of matrices and positivity of vector components. In fact, we could not use the tools of [2] to obtain our result.

Since we will use Hadamard (elementwise) products of matrices the following notation for elementwise functions will become practical

Notation. For a scalar function f and a matrix A denote by $H_f(A)$ the matrix of the size of A consisting of elements $f(A_{ij})$.

We will denote by J the $d \times d$–matrix consisting of ones, i.e., the "corner" of positive definite matrices, and by J_k the corresponding $dk \times d$–matrix.

The following lemma is the crucial point of Theorem 2.3 below and may also be of independent interest.

Lemma 2.2. *Suppose ϕ is real analytic in a neighbourhood of 0 such that for any $a_1, a_2, a_3 \in \mathbf{R}$ we have*

$$\begin{aligned}&(i) \quad \phi(z) = 1 + O(z^3)\\ &(ii) \quad Det\left(\left\{\phi(\nu(a_i+a_j))\right\}_{i,j=1}^{3}\right) \geq 0 \quad \text{for small enough } \nu \geq 0.\end{aligned}$$

Then $\phi(z) \equiv 1$.

Proof. Assume the contrary: $\phi(z) = 1 + z^m f(z)$, where $m \geq 3$ and $f(0) \neq 0$. Take $a_i = \varepsilon^{i-1}$. Then collecting terms of order $\leq \nu^{2m}\varepsilon^2$ for $\nu \ll \varepsilon \ll 1$:

$$\begin{aligned}
0 &\leq \mathrm{Det}\left(\left\{\phi(\nu(a_i+a_j))\right\}_{i,j=1}^{3}\right) = \\
&= \begin{vmatrix} 1+2^m\nu^m f(2\nu) & 1+(1+\varepsilon)^m\nu^m f((1+\varepsilon)\nu) & 1+(1+\varepsilon^2)^m\nu^m f((1+\varepsilon^2)\nu) \\ 1+(1+\varepsilon)^m\nu^m f((1+\varepsilon)\nu) & 1+2^m\varepsilon^m\nu^m f(2\varepsilon\nu) & 1+(\varepsilon+\varepsilon^2)^m\nu^m f((\varepsilon+\varepsilon^2)\nu) \\ 1+(1+\varepsilon^2)^m\nu^m f((1+\varepsilon^2)\nu) & 1+(\varepsilon+\varepsilon^2)^m\nu^m f((\varepsilon+\varepsilon^2)\nu) & 1+2^m\varepsilon^{2m}\nu^m f(2\varepsilon^2\nu) \end{vmatrix} = \\
&= -\nu^{2m}\Big[\big[(1+\varepsilon)^{2m} - 2(1+\varepsilon)^m(1+\varepsilon^2)^m + (1+\varepsilon^2)^{2m}\big]f(0)^2 + O(\varepsilon^3)\Big] = \\
&= -\nu^{2m}[m^2\varepsilon^2 f(0)^2 + O(\varepsilon^3)].
\end{aligned}$$

Thus $f(0) = 0$, a contradiction. $\square$

We consider only such methods, which when applied to a system of independent equations, give the same result as the application to each of the independent equations separately. Also we assume that the application to the test equation $\dot{x} = \lambda x$ with stepsize h gives x_{n+1} as a function which is rational in $h\lambda$ and linear in the previous x-values. All the standard methods (e.g. Runge-Kutta methods or linear multistep methods, explicit or implicit) satisfy this.

Here is the anticipated negative result:

Theorem 2.3. *Any one-step method or strictly stable multistep method that preserves positive definiteness in the numerical solution of the Lyapunov equation has order at most one.*

In more detail: for any given k–step method of order $p \geq 2$ there exists a diagonal constant 3×3 matrix A and an open set $\mathbf{X}$ of 3×3 positive matrices such that if we apply the method to (1.2) (with $C = 0$) with any small enough h and any initial values $\tilde{X}(0), \ldots, \tilde{X}((k-1)h) \in \mathbf{X}$, then there exists a $t_n = O(h^{-p})$ such that $\det(\tilde{X}(t_n)) < 0$.

Proof. Consider first a one-step method of order $p \geq 2$. Take a homogeneous Lyapunov equation, i.e., $C = 0$, with $A = \mathrm{diag}[a_1, a_2, a_3]$. Then $\dot{X}_{ij} = (a_i + a_j)X_{ij}$, $i, j \in \{1, 2, 3\}$, i.e.,

$$\dot{X}(t) = \tilde{A} \bullet X(t)\,, \tag{2.2}$$

where $\tilde{A}_{ij} := a_i + a_j$ and $\bullet$ denotes the Hadamard (elementwise) product. The numerical method applied to (2.2) produces

$$\tilde{x}_{ij}((n+1)h) = r(h(a_i + a_j))\tilde{x}_{ij}(nh)\,,$$

where r is the rational function corresponding to the method: for $\dot{y} = \lambda y$ we have $y(h) \approx \tilde{y}(h) = r(h\lambda)y(0)$. Thus

$$\tilde{X}(nh) = H_r(h\tilde{A})^{\bullet n} \bullet X_0\,.$$

We will show that for suitable A there exists an open set $\mathbf{X}$ of positive initial values X_0 such that for any h small enough and for any $X_0 \in \mathbf{X}$ the determinant of $\tilde{X}(nh)$ will be negative for some $n = O(h^{p+1})$.

Let f be such that $f(0) \neq 0$ and

$$e^{-z}r(z) = 1 + z^{p+1}f(z)\,.$$

Since

$$\mathrm{Det}(\tilde{X}(nh)) = \mathrm{Det}(H_r(h\tilde{A})^{\bullet n} \bullet X_0) = e^{2hn(a_1+a_2+a_3)}\mathrm{Det}\left(H_{r/\exp}(h\tilde{A})^{\bullet n} \bullet X_0\right)\,,$$

it suffices to study the sign of the last determinant. For $\nu > 0$ set

$$\phi(\nu z) := \lim_{h\to 0}(1 + (hz)^{p+1}f(hz))^{(\nu/h)^{p+1}} = e^{(\nu z)^{p+1}f(0)}\,. \tag{2.3}$$

Then

$$\lim_{h\to 0} H_{r/\exp}(h\tilde{A})^{\bullet\lfloor(\nu/h)^{p+1}\rfloor} = H_\phi(\nu\tilde{A})\,.$$

Using Lemma 2.2 take $a_1, a_2, a_3 \in \mathbf{R}$ and $\nu > 0$ such that

$$\mathrm{Det}(H_\phi(\nu\tilde{A}) \bullet J) = \mathrm{Det}(\{\phi(\nu(a_i + a_j))\}_{i,j=1}^3) < 0\,. \tag{2.4}$$

By continuity of the determinant and the Hadamard product there exists an $h_0 > 0$ and a neighbourhood V of J such that

$$\mathrm{Det}(\tilde{X}(h\lfloor(\nu/h)^{p+1}\rfloor)) < 0$$

for all $h \in (0, h_0)$ and X_0 in V.

Turn now to the multistep case. The strictly stable k–step method of order $p \geq 2$ applied to $\dot{y} = \lambda y$ with stepsize h runs like

$$\begin{bmatrix} \tilde{y}((j+1)h) \\ \cdot \\ \cdot \\ \cdot \\ \tilde{y}((j+k)h) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & 1 \\ d_1(h\lambda) & d_2(h\lambda) & d_3(h\lambda) & \dots & d_k(h\lambda) \end{bmatrix} \begin{bmatrix} \tilde{y}((j)h) \\ \cdot \\ \cdot \\ \cdot \\ \tilde{y}((j+k-1)h) \end{bmatrix},$$

where d_i's are rational functions. Denote the matrix above by $D(h\lambda)$. By strict stability there exists an f analytic and nonzero in a neighbourhood of 0 such that the principal eigenvalue $r(z)$ of $D(z)$ satisfies

$$r(z) = e^z(1 + z^{p+1} f(z))$$

and has maximum modulus (see [6]). For this f define ϕ as in (2.3).

Let $\xi(z) = [1, \ldots, 1]^T + O(z)$ and $\eta(z)$ be the right and left eigenvectors of $D(z)$ – corresponding to $r(z)$ – such that $\eta(z)^T \xi(z) = 1$. Then uniformly for small z

$$\lim_{n\to\infty} \left[r(z)^{-1} D(z)\right]^n = \xi(z)\eta(z)^T \tag{2.5}$$

(see [8] Lemma 8.2.7, the uniformity follows easily from (j) there).

Consider as before a homogeneous Lyapunov equation with $A = \mathrm{diag}[a_1, a_2, a_3]$. Applying the method with stepsize h to (2.2) produces

(2.6)

$$\begin{bmatrix} \tilde{X}(nh) \\ \cdot \\ \cdot \\ \cdot \\ \tilde{X}((n+k-1)h) \end{bmatrix} = \begin{bmatrix} 0 & J & & 0 \\ 0 & 0 & & 0 \\ & & \ddots & \\ 0 & 0 & & J \\ \tilde{d}_1(h\tilde{A}) & \tilde{d}_2(h\tilde{A}) & \ldots & \tilde{d}_k(h\tilde{A}) \end{bmatrix}^{\circ n} \circ \begin{bmatrix} \tilde{X}(0) \\ \cdot \\ \cdot \\ \cdot \\ \tilde{X}((k-1)h) \end{bmatrix}$$

using (2.5) to each of the mutually independent equations for X_{ij} gives

$$\lim_{n\to\infty} \left[\tilde{r}(h\tilde{A})^{\bullet -1} \circ \tilde{D}(h\tilde{A})\right]^{\circ n} = \tilde{\xi}(h\tilde{A}) \circ \tilde{\eta}(h\tilde{A})^T , \tag{2.7}$$

uniformly for small h . Here $\tilde{r}$ and the entries of $\tilde{D}$, $\tilde{\xi}$, and $\tilde{\eta}$ are 3×3 matrices in the natural way, e.g. $\tilde{d}_i(h\tilde{A}) = H_{d_i}(h\tilde{A})$. Notation $\circ$ means that the products are the usual matrix products, but elements, which are 3×3 matrices are multiplied Hadamard-wise.

For studying the signs of the determinants of $\tilde{X}(nh)$ note that

$$e^{-2hn\sum_i a_i} \det(\tilde{X}(nh)) = \det(H_{\exp}(-h\tilde{A}) \bullet \tilde{X}(nh)) .$$

So, Hadamard-multiply (2.6) by $H_{\exp}(-h\tilde{A})$. Then (2.7) gives

$$\begin{aligned} &\lim_{h\to 0} \left[H_{\exp}(-h\tilde{A}) \circ \tilde{D}(h\tilde{A})\right]^{\circ\lfloor(\nu/h)^{p+1}\rfloor} \circ J_k = \\ &= \lim_{h\to 0} \left[H_{r/\exp}(h\tilde{A})^{\bullet\lfloor(\nu/h)^{p+1}\rfloor}\right] \circ \lim_{h\to 0} \left[\tilde{r}(h\tilde{A})^{\bullet -1} \circ \tilde{D}(h\tilde{A})\right]^{\circ\lfloor(\nu/h)^{p+1}\rfloor} \circ J_k = \\ &= H_\phi(\nu\tilde{A}) \circ J_k . \end{aligned}$$

As before take $a_1, a_2, a_3 \in \mathbf{R}$ and $\nu > 0$ such that (2.4) holds. Then there exists a neighbourhood V of J and $h_0 > 0$ such that for any $h \in (0, h_0)$ and $[\tilde{X}(0), \tilde{X}(h), \ldots, \tilde{X}((k-1)h)]$ in $V \times V \times \cdots \times V$ we have

$$\mathrm{Det}(\tilde{X}(h\lfloor(\nu/h)^{p+1}\rfloor)) < 0 .$$

Finally, let $\mathbf{X}$ be the intersection of V with the set of positive matrices. $\square$

Remark 2.1. Since it is a negative result, Theorem 2.3 applies to Riccati equations as well.

Remark 2.2. It is seen from the proof that the time it takes to loose positivity is $O(h^{-p})$. This is the worst we can expect in general.

Remark 2.3. The approach here for the multistep methods does not assume anything about the initialization method. *Any* consistent starting routine can be taken. This is in the spirit of the approach in [5].

3. Indirect methods

The negative result of Theorem 2.3 motivates the attempt to recover solutions of Riccati equations through indirect procedures. Here below we look at two of them.

3.1. The fundamental solution method. This is probably the best known alternative to a direct integration and is based on the fact that the solution of the Riccati equation can be obtained from the solution of the Hamiltonian system

$$\begin{bmatrix} Y(t) \\ Z(t) \end{bmatrix}^{\cdot} = \begin{bmatrix} A(t) & C(t) \\ B(t) & -A(t)^T \end{bmatrix} \begin{bmatrix} Y(t) \\ Z(t) \end{bmatrix} \quad , \quad \begin{bmatrix} Y(0) \\ Z(0) \end{bmatrix} = \begin{bmatrix} X_0 \\ I \end{bmatrix} \tag{3.1}$$

by $X(t) = Y(t)Z(t)^{-1}$.

One possible approach to solve the Riccati equation numerically is to apply a discretization to (3.1) and then form $X_n = Y_n Z_n^{-1}$. Then the question of interest here is what discretization methods will finally produce nonnegative sequence of X_n's. The following is the first positive result in this direction. It says that the Gauss (Gauss-Legendre) methods really might be a choice.

The Gauss methods (diagonal Padé approximants) belong to the class of *symplectic* Runge-Kutta methods, since their coefficients satisfy (see: [11]):

$$M_{ij} := b_i a_{ij} + b_j a_{ji} - b_i b_j = 0 \quad \forall\, i,j = 1,\dots,k \; . \tag{3.2}$$

Further, the b_j's of Gauss methods are nonnegative. Note that the matrix M above is the one that is required to be nonnegative (together with b_j's) in the definition of *algebraic stability* by Burrage&Butcher and Crouzeix (see: e.g. [7]).

Theorem 3.1. *Application of a Runge-Kutta method which satisfies* (3.2) *with nonnegative* b_i*'s to equation* (3.1) *produces (when defined) symmetric nonnegative matrices* X_n.

Proof. Let

$$H = \begin{bmatrix} A & C \\ B & -A^T \end{bmatrix} \; , \quad E = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \; , \quad S = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \; , \quad V = \begin{bmatrix} Y \\ Z \end{bmatrix} \; .$$

We prove that $V_{n+1}^T S V_{n+1}$ is zero and $V_{n+1}^T E V_{n+1}$ is nonnegative, assuming that $V_n^T S V_n$ and $V_n^T E V_n$ are so. Now, the RK step can be written as follows

$$V_{n+1} = V_n + h \sum_{i=1}^{k} b_i H_i P_i \quad , \tag{3.3}$$

$$P_i = V_n + h \sum_{j=1}^{k} a_{ij} H_j P_j \quad , \tag{3.4}$$

where $H_i = H(t_n + c_i h)$ and P_i's are the Runge-Kutta stages. Let Q stand for either E or S. Equation (3.3) gives

$$V_{n+1}^T Q V_{n+1} = (V_n^T + h \sum_{i=1}^{k} b_i P_i^T H_i^T) Q (V_n + h \sum_{j=1}^{k} b_j H_j P_j) = \tag{3.5}$$

$$= V_n^T Q V_n + h \Big(\sum_{i=1}^{k} b_i P_i^T H_i^T Q V_n + \sum_{j=1}^{k} b_j V_n^T Q H_j P_j \Big) + h^2 \sum_{i,j=1}^{k} b_i b_j P_i^T H_i^T Q H_j P_j \,.$$

From equation (3.4) we get

$$P_i^T H_i^T Q V_n = P_i^T H_i^T Q P_i - h \sum_{j=1}^{k} a_{ij} P_i^T H_i^T Q H_j P_j \,,$$

$$V_n^T Q H_j P_j = P_j^T Q H_j P_j - h \sum_{i=1}^{k} a_{ji} P_i^T H_i^T Q H_j P_j \,.$$

Putting these in (3.5) gives

$$\begin{aligned} V_{n+1}^T Q V_{n+1} = V_n^T Q V_n &+ h \sum_{i=1}^{k} b_i P_i^T (H_i^T Q + Q H_i) P_i \\ &- h^2 \sum_{i,j=1}^{k} (b_i a_{ij} + b_j a_{ji} - b_i b_j) P_i^T H_i^T Q H_j P_j \,. \end{aligned} \tag{3.6}$$

Now, by (3.2) the h^2 –terms vanish.

When $Q = S$, we have $V_n^T S V_n = 0$ and $H_i^T S + S H_i = 0$. Then (3.6) gives $Y_{n+1}^T Z_{n+1} = Z_{n+1}^T Y_{n+1}$. This means that $Z_{n+1}^T Y_{n+1}$ and hence also $X_{n+1} = Z_{n+1}^{-T} (Z_{n+1}^T Y_{n+1}) Z_{n+1}^{-1}$ is symmetric.

When $Q = E$, we have $V_n^T E V_n \geq 0$ and

$$H^T E + E H = \begin{bmatrix} B & 0 \\ 0 & C \end{bmatrix} \,,$$

where $B, C \geq 0$. Since $b_i \geq 0$ the result follows from (3.6). $\square$

3.2. Linearization. Here we use the formulae (1.3) to obtain positive higher degree ($p \geq 2$) approximations to Riccati and Lyapunov equations.

Consider first the Lyapunov equation (1.2)

$$\dot{X}(t) = A(t)X(t) + X(t)A(t)^T + C(t) \tag{3.7}$$

and its solution formula (1.3):

$$X(t) = \Phi(t,s)X(s)\Phi(t,s)^T + \int_s^t \Phi(t,\tau)C(\tau)\Phi(t,\tau)^T d\tau \quad , \tag{3.8}$$

$$\partial_t \Phi(t,\tau) = A(t)\Phi(t,\tau) \ , \quad \Phi(\tau,\tau) = I \ . \tag{3.9}$$

There are many ways to use these so that positivity will be preserved.

To proceed from a positive X_n to positive X_{n+1} one can take:
- any method to solve (3.9) on $[t_n, t_{n+1}]$
- any formula with positive weights for the integral in (3.8).

Consider, for example, solving (3.9) using the implicit midpoint rule and integrating (3.8) with the trapezoidal rule. This leads to

$$\begin{aligned} \Phi_n &:= [I - \tfrac{h}{2}A(t_n + \frac{h}{2})]^{-1}[I + \tfrac{h}{2}A(t_n + \frac{h}{2})] \\ X_{n+1} &:= \Phi_n \left[X_n + \tfrac{h}{2}C(t_n)\right] \Phi_n^T + \tfrac{h}{2}C(t_n + h) \end{aligned} \tag{3.10}$$

for $n = 0, 1, 2, \ldots$ Clearly this produces second degree approximations for $X(nh)$. Denote this map $X_n \to X_{n+1}$ by $X_{n+1} = G(X_n, h, A, C)$.

Consider, then, the Riccati equation and write it as:

$$\dot{X}(t) = [A(t) - \tfrac{1}{2}X(t)B(t)]X(t) + X(t)[A(t) - \tfrac{1}{2}X(t)B(t)]^T + C(t)$$

Now we have again:

$$X(t) = \Phi(t,s)X(s)\Phi(t,s)^T + \int_s^t \Phi(t,\tau)C(\tau)\Phi(t,\tau)^T d\tau \ , \text{ where} \tag{3.11}$$

$$\partial_t \Phi(t,\tau) = [A(t) - \tfrac{1}{2}X(t)B(t)]\Phi(t,\tau) \ , \ \Phi(\tau,\tau) = I \ . \tag{3.12}$$

To proceed from a positive X_n to positive X_{n+1} in this case one can take:
- any method to solve (3.12) on $[t_n, t_{n+1}]$
- obtain the necessary X–values for it using any other method
- any formula with positive weights for the integral in (3.11).

Consider, again, solving (3.12) by the implicit midpoint rule and integrating (3.11) with the trapezoidal rule to get a second degree method. The midpoint rule for (3.12) requires an $O(h^2)$ –approximation for $X(t_n + \frac{h}{2})$. This can be obtained e.g. using an explicit–in–X step first. The resulting scheme is

$$
\begin{aligned}
\Phi_{n+\frac{1}{2}} &= [I - \tfrac{h}{4}(A_n - \tfrac{1}{2}B_nX_n)]^{-1}[I + \tfrac{h}{4}(A_n - \tfrac{1}{2}B_nX_n)] \\
X_{n+\frac{1}{2}} &= \Phi_{n+\frac{1}{2}}[X_n + \tfrac{h}{4}C_n]\Phi^T_{n+\frac{1}{2}} + \tfrac{h}{4}C_{n+\frac{1}{2}} \\
\Phi_n &= [I - \tfrac{h}{2}(A_{n+\frac{1}{2}} - \tfrac{1}{2}B_{n+\frac{1}{2}}X_{n+\frac{1}{2}})]^{-1}[I + \tfrac{h}{2}(A_{n+\frac{1}{2}} - \tfrac{1}{2}B_{n+\frac{1}{2}}X_{n+\frac{1}{2}})] \\
X_{n+1} &= \Phi_n[X_n + \tfrac{h}{2}C_n]\Phi^T_n + \tfrac{h}{2}C_{n+1}
\end{aligned}
\tag{3.13}
$$

Remark 3.1. Using the map $G : X_n \to X_{n+1}$ of the Lyapunov solver (3.10), we can write the Riccati solver as

$$
X_{n+\frac{1}{2}} = G(\, X_n \,,\, \tfrac{h}{2} \,,\, A - \tfrac{1}{2}BX_n \,,\, C\,) \,, \quad X_{n+1} = G(\, X_n \,,\, h \,,\, A - \tfrac{1}{2}BX_{n+\frac{1}{2}} \,,\, C\,) \,. \tag{3.14}
$$

Remark 3.2. It is straightforward to build lots of methods and with arbitrarily high degree along these lines. These examples can be thought of as linearly implicit RK's. A fully implicit version can be obtained e.g. by taking only the last two lines of (3.13) and replacing there $X_{n+\frac{1}{2}}$ by $\frac{1}{2}(X_n + X_{n+1})$.

Remark 3.3. Since these methods are not direct discretizations, the equilibria of the systems are usually not preserved exactly. However, the maps $X_n \to X_{n+1}$ are $O(h^{p+1})$ C^1–close to the time–h map of the flow. It follows that nondegenerate equilibria are preserved to the order $O(h^p)$ of the method.

Remark 3.4. A similar technique to that above is involved in the *fractional step method* (see e.g. [10]) for nonlinear PDE's. There it is used mainly to avoid nonlinear systems of equations. Our reasons are in positivity, but of course it is nice to solve only linear systems. Compare (3.13) to the trapezoidal rule applied directly to the Riccati equation:

$$
\begin{aligned}
X_{n+1} - \tfrac{h}{2}[A_{n+1}X_{n+1} + X_{n+1}A^T_{n+1} - X_{n+1}B_{n+1}X_{n+1}] &= \\
= X_n + \tfrac{h}{2}[A_nX_n + X_nA^T_n - X_nB_nX_n + C_n + C_{n+1}]&
\end{aligned}
$$

This one needs a solver for the algebraic Riccati equation.

3.3. Other possibilities. Some other approaches to preserve positivity can be obtained from e.g. the *continuous eigendecomposition*, i.e., integrating for the orthogonal matrix $U(t)$ and the diagonal matrix $\Lambda(t)$ resulting from the continuous Schur decomposition of $X(t)$. That is, one writes

$$
X(t) = U^T(t)\Lambda(t)U(t) \,,
$$

where U is orthogonal: $U^T(t)U(t) = I$, $\forall t$, and $\Lambda(t) = \mathrm{diag}(\lambda_1(t), \ldots, \lambda_d(t))$. Then one sets up equations for $U(t)$ and $\Lambda(t)$. Clearly, $X(t)$ is nonnegative exactly when $\lambda_i(t) \geq 0$.

If one takes a symplectic integrator for U, then the orthogonality is automatically satisfied by the numerical solution, since $U^TU = I$ is a system of quadratic first integrals, hence preserved (see: [3] and [11]). Preserving positivity here is then easy by monitoring the behaviour of λ_i's.

It is difficult for us, however, to see why this approach or other projection type methods could be more beneficial – in computational cost or otherwise – compared to the approaches above.

4. Conclusions

In this paper we addressed the issue of preserving positivity in the numerical solution of Riccati and Lyapunov equations. It was shown that a direct discretization of these equations then limits the order to one. To obtain higher order of accuracy indirect solution methods are needed. Two approaches leading to arbitrarily accurate and positive approximations have been given. The question of how to implement these in an efficient way will be the subject of future study. There are a number of interesting design problems in the choice of schemes and the linear algebra routines involved: e.g. how to "recycle" the factorization of a matrix most economically and how to choose the discretization for the linearization schemes so to minimize the overall expense.

References

1. B. D. O. Anderson and J. B. Moore, *Linear optimal control*, Prentice-Hall, Englewood Cliffs, 1971.
2. C. Bolley and M. Crouzeix, *Conservation de la positivité lors de la discrétization des problèmes d'évolution paraboliques,*, R.A.I.R.O. Analyse Numérique **12** (1978), 237–245.
3. L. Dieci, R. D. Russell, and E. S. van Vleck, *Unitary integrators and applications to continuous orthonormalization techniques*, to appear in SIAM J. Numer. Anal. (1993).
4. Luca Dieci, *Numerical integration of the differential Riccati equation and some related issues*, SIAM J. Numer. Anal. **29** (1992), 781–815.
5. Timo Eirola and Olavi Nevanlinna, *What do multistep methods approximate?*, Numer. Mathematik **53** (1988), 559–569.
6. E. Gekeler, *Discretization methods for stable initial value problems*, Springer-Verlag, Berlin, 1984, Lecture Notes in Math. 1044.
7. E. Hairer and G. Wanner, *Solving ordinary differential equations II*, Springer-Verlag, Berlin-Heidelberg, 1991.
8. R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge University Press, New York, 1985.
9. P. Lancaster and L. Rodman, *Existence and uniqueness theorems for the algebraic Riccati equation*, Int. J. Control **32** (1980), 285–309.
10. J. L. Lions, *Optimal control of systems governed by partial differential equations*, Springer-Verlag, Berlin, 1971.
11. J. M. Sanz-Serna, *Symplectic integrators for Hamiltonian problems: An overview*, Acta Numerica **1** (1992), 243–286.

Department of Mathematics, Georgia Tech, Atlanta, GA, 30332 U.S.A
E-mail address: dieci@math.gatech.edu

Inst. of Mathematics, Helsinki Univ. of Technology, SF-02150 Espoo, Finland
E-mail address: Timo.Eirola@hut.fi