MATHEMATICS OF COMPUTATION
Volume 81, Number 277, January 2012, Pages 233–265
S 0025-5718(2011)02498-1
Article electronically published on May 6, 2011

A FAMILY OF ANADROMIC NUMERICAL METHODS FOR MATRIX RICCATI DIFFERENTIAL EQUATIONS

REN-CANG LI AND WILLIAM KAHAN

ABSTRACT. Matrix Riccati Differential Equations (MRDEs)

$$X' = A_{21} - XA_{11} + A_{22}X - XA_{12}X, \quad X(0) = X_0,$$

where $A_{ij} \equiv A_{ij}(t)$, appear frequently throughout applied mathematics, science, and engineering. Naturally, the existing conventional Runge-Kutta methods and linear multi-step methods can be adapted to solve MRDEs numerically. Indeed, they have been adapted. There are a few unconventional numerical methods, too, but they are suited more for time-invariant MRDEs than time-varying ones. For stiff MRDEs, existing implicit methods which are preferred to explicit ones require solving nonlinear systems of equations (of possibly much higher dimensions than the original problem itself of, for example, implicit Runge-Kutta methods), and thus they can pose implementation difficulties and also be expensive.

In the past, the property of an MRDE which has been most preserved is the symmetry property for a symmetric MRDE; many other crucial properties have been discarded. Besides the symmetry property, our proposed methods also preserve two other important properties — *Bilinear Rational Dependence* on the initial value, and a *Generalized Inverse Relation* between an MRDE and its complementary MRDE. By preserving the generalized inverse relation, our methods are accurately able to integrate an MRDE whose solution has singularities. By preserving the property of bilinear dependence on the initial value, our methods also conserve the rank of change to the initial value and a solution's monotonicity property.

Our methods are *anadromic*,[1]meaning if an MRDE is integrated by one of our methods from $t{=}\tau$ to $\tau{+}\theta$ and then integrated backward from $t{=}\tau{+}\theta$ to τ using the same method, the value at $t{=}\tau$ is recovered in the absence of rounding errors. This implies that our methods are necessarily of even order of convergence. For time-invariant MRDEs, methods of any even order of convergence are established, while for time-varying MRDEs, methods of order as high as 10 are established; but only methods of order up to 6 are stated in detail.

Our methods are semi-implicit, in the sense that there are no nonlinear systems of matrix equations to solve, only linear ones, unlike any pre-existing implicit method. Given the availability of high quality codes for linear matrix equations, our methods can easily be implemented and embedded into any application software package that needs a robust MRDE solver.

Numerical examples are presented to support our claims.

[1] We coined "anadromic" from the Greek roots "$\alpha\nu\alpha$-" (back up) and "$\delta\rho o\mu o\sigma$" (act of running) of "anadromous" (which describes fish that must return from the ocean to spawn in the same streams whence they hatched) after we had tried to describe our numerical methods with words such as "reflexive" [27, 28], "symmetric" (which is still used today in the literature, e.g., [22]), and "reversible" which turned out to be overused.

Received by the editor Received August 25, 2009 and, in revised form, October 30, 2010.

2010 *Mathematics Subject Classification.* Primary 65L05.

Key words and phrases. Matrix Riccati differential equations, solution singularity, anadromic numerical method, group of two-sided bilinear rational functions, generalized inverse relation.

This work was supported in part by the National Science Foundation Grant DMS-0810506.

1. Introduction

Matrix Riccati Differential Equations (MRDEs) arise frequently throughout applied mathematics, science and engineering. In particular they play major roles in optimal control, filtering and estimation [30] and in solving linear two point boundary value problems of ordinary differential equations (ODEs) [3, 4, 18, 19]. A number of algorithms have been proposed in the past for solving MRDEs numerically. These include carefully redesigned conventional Runge-Kutta methods and Linear Multi-step Methods for ODEs by Choi and Laub [10] and by Dieci [15], and unconventional methods for MRDEs arising from optimal control theory, e.g., [12, 29, 32, 31, 33, 34, 37, 40, 42, 44]. It is known that these unconventional methods are either not suited or inefficient for time-varying MRDEs. While the redesigned conventional methods benefit greatly from past development of sophisticated general-purposed computer programs for ODEs, they easily evolve into complicated programs thousands of lines long with complicated interfaces. Implicit conventional methods which are preferred to explicit ones for stiff systems require solving nonlinear systems of equations (of possibly much higher dimensions than the original problem itself for Runge-Kutta methods) which can pose implementation difficulties and also be expensive.

None of these preexisting methods can integrate over solution singularities (poles) that occur occasionally in applications, e.g., in one-dimensional quantum Hamilton-Jacobi equations [11]. In this paper, we will establish a family of unconventional numerical methods that can produce meaningful numerical results even if there are poles in the solution. This capability is a byproduct of our numerical formulas that preserve crucial structural properties previously disregarded. One of our second order methods is not entirely new. It was used in 1987 by Babuška and Majer [4, Section 3.2] for both time-invariant and time-varying MRDEs, and had also been used independently by the second author here in one of his unpublished notes in the 1980s. While this method is only of order 2, it conserves many important properties of MRDEs that are crucial to our investigation here. Such conservations were not mentioned in [4], however.

Generally, an MRDE takes the form

$$\text{(MRDE)} \qquad X' = A_{21} - XA_{11} + A_{22}X - XA_{12}X, \quad X(0) = X_0,$$

where X is an n-by-m (not necessarily square) matrix-valued function of time t, and all A_{ij} are smooth matrix-valued functions of time t, too, with dimensions determined by the following conformal partitioning:

$$A \equiv A(t) = \begin{array}{cc} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \end{array}. \tag{1.1}$$

The general form of our methods for integrating from $t = \tau$ to $t = \tau + \theta$ is

$$\frac{\boldsymbol{Y} - \boldsymbol{X}}{\theta/2} = \sum_{\ell=0}^{k-1} (\tfrac{1}{2}\theta)^{2\ell} c_\ell f_\ell(\boldsymbol{X}, \boldsymbol{Y}), \quad \frac{\boldsymbol{Z} - \boldsymbol{Y}}{\theta/2} = \sum_{\ell=0}^{k-1} (\tfrac{1}{2}\theta)^{2\ell} c_\ell f_\ell(\boldsymbol{Z}, \boldsymbol{Y}), \tag{1.2}$$

where $\boldsymbol{X} \approx X(\tau)$ and $\boldsymbol{Z} \approx X(\tau + \theta)$, and $f_\ell(\,\cdot\,,\,\cdot\,)$ are matrix-valued functions to be determined so that this will give a method of order $2k$ and at the same time keep the equations in (1.2) linear in $\boldsymbol{Y}$ and $\boldsymbol{Z}$, separately, and c_ℓ are the coefficients in the power series of $\tanh t$. For time-invariant MRDEs, we have found all such methods of all even orders. For time-varying MRDEs, we have found methods of order as high as 10; but only methods of order up to 6 are described in detail owing to their exponentially growing complexities as the order increases. *Mathematica* plays a major role in our finding those methods for the time-varying MRDEs.

The method of (1.2) can be cast into the framework of modified integrators in the sense of [8], and is also closely related to the implicit midpoint rule on an associated linear differential equation. The latter fact, generously shared with us by Hairer [23], makes it possible for us to simplify significantly our earlier construction of high order methods in the form of (1.2) in [35]. This is done now by establishing modified implicit midpoint rules for a linear differential equation [36].

Geometrically, an MRDE can be viewed as a flow on the Grassmann manifold [46]. Schiff and Shnider [41] appear to be the first to take advantage of this point of view and proposed so-called *Möbius Schemes* to better simulate the flow. The basic idea is to numerically preserve the (bi)linear rational dependence property, one of the three properties we shall discuss in Section 2, and it is done through approximating the fundamental solution of the associated linear differential equation. It was argued that this preservation would enable the schemes "to deal with numerical instability and pass accurately through the singularities". Our methods (1.2) preserve the (bi)linear rational dependence property, besides two other properties—symmetry and a generalized inverse relation. In this sense, our methods fit into the form of their Möbius Schemes. But we argue that it is the preservation of the generalized inverse relation property that enables us to give a rigorous justification for why our methods can pass through singularities.

Solutions $X(t)$ to an MRDE can also be regarded as sample-values of members of the two-sided bilinear rational matrix function group selected by t and sampled at an indeterminate $X(0)$. Our methods (1.2), as well as *Möbius Schemes* in [46], approximate $X(t)$ by a sequence of sample-values each drawn from the same group of two-sided bilinear rational matrix functions, regardless of $X(0)$. More detail is given in Section 5.

In analyzing our methods, the MRDE *complementary* to (MRDE), namely

$$\text{(cMRDE)} \qquad U' = A_{12} - UA_{22} + A_{11}U - UA_{21}U, \quad U(0) = U_0$$

plays a major role. We say (cMRDE) is the *complement* of (MRDE). The complement of the complement of an MRDE is the MRDE itself. In particular, the complement of (cMRDE) becomes (MRDE). To single out these two equations (MRDE) and (cMRDE) from the rest, we choose to label them differently to make them recognizable instantaneously.

The rest of this paper is organized as follows. Section 2 reviews three important properties of MRDEs that we should prefer our numerical methods to preserve, other things being equal. In Section 3, we first introduce two simple second order anadromic methods, based on the fact there are two pre-existing techniques to compute higher orders approximations—classical extrapolation and composition. Our focus in this article, however, is on a third technique that produces higher order anadromic methods in the general form of (1.2). In Section 4, we prove that our proposed methods (1.2) indeed preserve the three important properties, and as a byproduct a solution's monotonicity property. Because of the preserved properties, we argue in Section 6 that our methods will have the capability to march over solution poles and still render numerical solutions that are as accurate as dictated by the step-size, the order of the method used, and rounding errors in solving encountered linear matrix equations along the way. A linear stability theory is outlined in Section 7 for our methods. The claim that our methods can march over the poles and our linear stability theory are validated by three numerical examples in Section 8. Section 9 presents our conclusions.

Notation. X and $X(t)$ is the solution of (MRDE), and $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are numerical approximates at $t=\tau$, $\tau+\theta/2$, and $\tau+\theta$, respectively. Similarly, symbols U, $U(t)$, $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{W}$ mean the same for the corresponding (cMRDE). I_k is the $k\times k$ identity matrix, or simply I when its dimension is clear from the context. Superscript $(\,\cdot\,)^{\mathrm{T}}$ denotes transpose, while $(\,\cdot\,)^{*}$ denotes conjugate transpose.

2. Important properties of MRDEs

An MRDE is said to be *symmetric* if

$$A_{21}=A_{21}^{\mathrm{T}},\quad A_{11}=-A_{22}^{\mathrm{T}},\quad A_{12}=A_{12}^{\mathrm{T}},\quad\text{and}\quad X(0)=X(0)^{\mathrm{T}}. \tag{2.1}$$

This definition of symmetry is valid regardless of whether A and $X(0)$ are complex, and in fact, in our later development, it is not necessary for A and $X(0)$ to be real when we refer to a symmetric MRDE. We say (MRDE) is *Hermitian* if

$$A_{21}=A_{21}^{*},\quad A_{11}=-A_{22}^{*},\quad A_{12}=A_{12}^{*},\quad\text{and}\quad X(0)=X(0)^{*}. \tag{2.2}$$

An MRDE is said to be *time-invariant* if A in (1.1) does not depend on t, i.e., a constant matrix; otherwise it is *time-varying.*

In what follows we shall explain three properties that are important to us in this study: *Bilinear Rational Dependence* on the initial value, *Generalized Inverse Property*, and *Symmetry.* Among them, the *Symmetry* property (Hermitian MRDEs included) is easiest to preserve, and almost all known numerical schemes do achieve that, e.g., the straightforward applications of the Runge-Kutta and linear multistep methods. But preserving the other two properties cheaply is a nontrivial task achieved by few previous numerical schemes.

All three properties, however, are preserved by our methods in the next section. Because of this, we argue later that our methods preserve a solution's monotonicity for symmetric MRDEs and are capable of marching over solution's poles.

2.1. **Bilinear rational property.** In principle, (MRDE) could be reduced to a (time-varying) linear homogeneous equation by the Bernoulli substitution of the form $X=TS^{-1}$ [39]:

$$\frac{dP}{dt}=AP,\ P(0)=\begin{pmatrix} S_0 \\ T_0 \end{pmatrix},\ \text{wherein } P=\begin{array}{c@{}c} & m \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} S \\ T \end{pmatrix} \end{array}. \tag{2.3}$$

The solution to (MRDE) with $X(0)=T_0S_0^{-1}$ relates to the solution to (2.3) by $X(t)=T(t)S(t)^{-1}$ so long as $S(t)$ remains invertible, and it would if $X(t)$ stayed finite because $S(t)$ satisfies a linear homogeneous differential equation $S'=(A_{11}+A_{12}X)S$. But this reduction of the given MRDE to a linear homogeneous differential equation is vulnerable to numerical instability when, as happens sometimes, all the columns of P approach a subspace of dimension lower than the number of columns, and then S becomes too nearly noninvertible to allow X to be recovered accurately from TS^{-1}. This is why $X(t)$, if it must be computed numerically, is usually computable better from the given MRDE than from either its foregoing linear reduction or a second linear reduction, namely

$$R' = -RA,\ R(0) = \begin{pmatrix} T_0 & -S_0 \end{pmatrix},\ \text{wherein } R = \begin{pmatrix} T & -S \end{pmatrix} \tag{2.4}$$

for which $X(t) = S(t)^{-1}T(t)$, provided $X(0) = S_0^{-1}T_0$. Both reductions can fizzle numerically.

Still the two linear reductions shed light on how the desired solution $X(t)$ depends on its initial value $X(0)$. Let $\Phi \equiv \Phi(t)$ be the *Fundamental Solution (or Matrizant)* of (2.3):

$$\frac{d\Phi}{dt} = A\Phi, \quad \Phi(0) = I_{m+n} \quad \text{(identity matrix)}, \tag{2.5}$$

so $P(t) = \Phi P(0)$. Consequently, after partitioning

$$\Phi = \begin{array}{c c} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \end{array}, \tag{2.6}$$

we find that

$$X(t) = [\Phi_{21} + \Phi_{22}X(0)][\Phi_{11} + \Phi_{12}X(0)]^{-1}. \tag{2.7}$$

This is well defined when $[\Phi_{11} + \Phi_{12}X(0)]^{-1}$ exists, which is guaranteed for small enough t. As t increases, if $\Phi_{11} + \Phi_{12}X(0)$ becomes singular at some point t_0, then t_0 becomes either a singularity or a removable singularity of $X(t)$. In any case, for any fixed t the matrix is a *Bilinear Rational Function* of $X(0)$.

$X(t)$ is actually a *Two-Sided Bilinear Rational Function* of $X(0)$, bilinear rational in two ways simultaneously due to the second linear reduction (2.4). Actually, the two kinds of bilinear rational dependence of $X(t)$ on $X(0)$ coexist partly because of the MRDE, but more because of an obscure matrix identity independent of that differential equation.

Theorem 2.1 (Most bilinear rational functions are two-sided)**.** *Let Φ and $\widetilde{\Phi}$ be two $(m+n)$-by-$(m+n)$ matrices partitioned in the same way as in* (2.6). *Suppose that there is at least one $n \times m$ matrix G such that both $\Phi_{11} + \Phi_{12}G$ and $G\widetilde{\Phi}_{12} - \widetilde{\Phi}_{22}$ are invertible*[2]. *Then*

$$[\Phi_{21} + \Phi_{22}G][\Phi_{11} + \Phi_{12}G]^{-1} \equiv [G\widetilde{\Phi}_{12} - \widetilde{\Phi}_{22}]^{-1}[-G\widetilde{\Phi}_{11} + \widetilde{\Phi}_{21}] \tag{2.8}$$

for every $n \times m$ matrix G such that both matrix inverses exist if and only if

$$\widetilde{\Phi}\,\Phi \equiv \begin{pmatrix} \widetilde{\Phi}_{11} & \widetilde{\Phi}_{12} \\ \widetilde{\Phi}_{21} & \widetilde{\Phi}_{22} \end{pmatrix} \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} = \mu I$$

for some scalar μ.

Proof. The first identity is equivalent to

$$[G\widetilde{\Phi}_{12} - \widetilde{\Phi}_{22}][\Phi_{21} + \Phi_{22}G] = [-G\widetilde{\Phi}_{11} + \widetilde{\Phi}_{21}][\Phi_{11} + \Phi_{12}G],$$

which expands into

$$\begin{aligned}(\widetilde{\Phi}_{21}\Phi_{11} + \widetilde{\Phi}_{22}\Phi_{21}) - G(\widetilde{\Phi}_{11}\Phi_{11} + \widetilde{\Phi}_{12}\Phi_{21})& \\ + (\widetilde{\Phi}_{21}\Phi_{12} + \widetilde{\Phi}_{22}\Phi_{22})G - G(\widetilde{\Phi}_{11}\Phi_{12} + \widetilde{\Phi}_{12}\Phi_{22})G &= 0.\end{aligned} \tag{2.9}$$

[2] This implies that both matrices are invertible for some nonempty open set of matrices G.

This holds for every G (such that $[\Phi_{11} + \Phi_{12}G]^{-1}$ and $[G\widetilde{\Phi}_{12} - \widetilde{\Phi}_{22}]^{-1}$ exist) if $\widetilde{\Phi}\Phi = \mu I$ is satisfied. On the other hand, if (2.8) holds for every matrix G in a nonempty open set, so does (2.9) for every matrix G in the nonempty open set and, consequently, for every $n \times m$ matrix G because (2.9) is a constraint upon quadratic polynomials in the entries of G; if satisfied by all elements in an open set, the identity must be satisfied by every aptly dimensioned G. Now run G through every such matrix each with at most one nonzero element to conclude $\widetilde{\Phi}\Phi = \mu I$ for some scalar μ. □

What use is the two-sided property of this bilinear rational function $X(t)$ of $X(0)$? One application is a far simpler proof than is found in Reid [39, p.12] of the following theorem. Write the bilinear rational function $X(t)$ of $X(0)$ as defined in (2.7) as $X(t) = \mathscr{F}(X(0))$.

Theorem 2.2 ([39])**.** Rank$(\mathscr{F}(X_1) - \mathscr{F}(X_2))$ = rank$(X_1 - X_2)$ *if* $\mathscr{F}(X_1)$ *and* $\mathscr{F}(X_2)$ *are both finite.*

Proof. Take $\mathscr{F}(X_1) = [\Phi_{21} + \Phi_{22}X_1][\Phi_{11} + \Phi_{12}X_1]^{-1}$ but change $\mathscr{F}(X_2) = [\Phi_{21} + \Phi_{22}X_2][\Phi_{11}+\Phi_{12}X_2]^{-1}$ to $\mathscr{F}(X_2) = [X_2\widetilde{\Phi}_{12}-\widetilde{\Phi}_{22}]^{-1}[-X_2\widetilde{\Phi}_{11}+\widetilde{\Phi}_{21}]$ as Theorem 2.1 provides with $\mu = 1$ since the fundamental solution Φ of (2.3) is nonsingular. Consequently,

$$\mathscr{F}(X_1) - \mathscr{F}(X_2) = [X_2\widetilde{\Phi}_{12} - \widetilde{\Phi}_{22}]^{-1}(X_1 - X_2)[\Phi_{11} + \Phi_{12}X_1]^{-1},$$

and the conclusion follows. □

This theorem implies that, when the solution $X(t) = \mathscr{F}(X(0))$ of a matrix Riccati differential equation changes because its initial value $X(0)$ has been changed, the rank of the change is conserved.

Observe that bilinear rational functions are closed under composition. In fact, two-sided bilinear rational matrix functions like (2.7) form a *group*. Solutions $X(t)$ to (MRDE) can be regarded as sample-values of members of the group selected by t and sampled at an indeterminate $X(0)$. More about this is in Section 5. We should prefer

Numerical methods that preserve this bilinear rational property in its computed solution $\boldsymbol{X}$, other things being equal.

2.2. **Generalized inverse property.** The MRDE ensures that all the inverses in Subsection 2.1 exist while t is small enough, and this ensures $X(t)$ is a two-sided bilinear rational function of $X(0)$ unless t gets so big that $X(t)$ has a ***pole***; if such a thing exists it is a finite $t \neq 0$ at which $X(t)$ becomes infinite.

What happens when a square solution $X(t)$ (i.e., $m = n$) passes through a pole? Typically its inverse $U(t)$ passes through a singular matrix; this $U(t)$ satisfies (cMRDE). When $m \neq n$, this complementary MRDE is still well defined. Theorem 2.3 below shows that a so-called *Generalized Inverse Relation* is enforced between two complementary MRDEs.

Theorem 2.3. *If* $X_0U_0 = I$ *(or* $U_0X_0 = I$*), and if the solutions* $X(t)$ *to* (MRDE) *and* $U(t)$ *to* (cMRDE) *have only isolated singularities and share none in common, then* $X(t)U(t) \equiv I$ *(or* $U(t)X(t) \equiv I$*, respectively).*

Proof. If $U_0X_0 = I$, then $(UX) = I$ solves the following initial value problem for (UX):

$$\begin{aligned} \frac{d}{dt}(UX) =&(A_{12} - UA_{22} + A_{11}U - UA_{21}U)X \\ &+ U(A_{21} - XA_{11} + A_{22}X - XA_{12}X) \\ =&A_{12}X - (UX)A_{12}X + A_{11}(UX) - (UX)A_{11} - UA_{21}(UX) + UA_{21} \\ =&[I - (UX)]A_{12}X + A_{11}[(UX) - I] \\ &- [(UX) - I]A_{11} - UA_{21}[(UX) - I], \\ (UX)|_{t=0} =& I. \end{aligned}$$

Therefore, $(UX) = I$ at least initially. Since all singularities in $X(t)$ and $U(t)$ are assumed to be isolated, so are those of $U(t)X(t)$. Thus all the singularities in the right-hand side of this ODE are removable.

The other case when $X_0U_0 = I$ is similar. □

We should prefer

> ***Numerical methods that retain this generalized inverse property in their computed solutions X and U***, other things being equal.

Equation $XU = I_n$ necessarily implies $n \leq m$ and that X's rows are linearly independent. We call such U a *right generalized inverse* of X. Similarly, $UX = I_m$ necessarily implies $n \geq m$ and that X's columns are linearly independent. We call such U a *left generalized inverse* of X. In either case, we call U a *generalized inverse* of X.

2.3. **Symmetry property.** Symmetric MRDEs, i.e., (2.1) holds, appear most commonly in optimal control and filtering problems [10, 29]. For a symmetric MRDE, A's eigenvalues come in pairs with opposite signs because

$$A \text{ is similar to } A^{\mathrm{T}} = -JAJ^{-1} \text{ which is similar to } -A,$$

where

$$J = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}, \quad J^{\mathrm{T}} = -J = J^{-1}. \tag{2.10}$$

Such a matrix A is said to be *Hamiltonian.* Then A and $-A$ have the same eigenvalues. If 0 is among them its multiplicity is even, even if it is defective for lack of an equal number of eigenvectors. A's eigenvalues are important in the discussion of the MRDE's attractive stationary points.

Next consider the Fundamental Solution of (2.3) partitioned thus:

$$\Phi = \begin{array}{c c} & \begin{array}{cc} n & n \end{array} \\ \begin{array}{c} n \\ n \end{array} & \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \end{array}. \tag{2.11}$$

Since $\frac{d}{dt}(J^{-1}\Phi^{\mathrm{T}}J) = -(J^{-1}\Phi^{\mathrm{T}}J)A$,

$$\Phi^{-1} = J^{-1}\Phi^{\mathrm{T}}J = \begin{pmatrix} \Phi_{22}^{\mathrm{T}} & -\Phi_{12}^{\mathrm{T}} \\ -\Phi_{21}^{\mathrm{T}} & \Phi_{11}^{\mathrm{T}} \end{pmatrix}.$$

Therefore,

$$\begin{aligned}\Phi_{11}\Phi_{22}^{\mathrm{T}}-\Phi_{12}\Phi_{21}^{\mathrm{T}}&=\Phi_{22}^{\mathrm{T}}\Phi_{11}-\Phi_{12}^{\mathrm{T}}\Phi_{21}=I,\\ \Phi_{11}\Phi_{12}^{\mathrm{T}}-\Phi_{12}\Phi_{11}^{\mathrm{T}}&=\Phi_{21}\Phi_{22}^{\mathrm{T}}-\Phi_{22}\Phi_{21}^{\mathrm{T}}=0,\\ \Phi_{22}^{\mathrm{T}}\Phi_{12}-\Phi_{12}^{\mathrm{T}}\Phi_{22}&=\Phi_{11}^{\mathrm{T}}\Phi_{21}-\Phi_{21}^{\mathrm{T}}\Phi_{11}=0.\end{aligned}$$

Then for every symmetric initial value $X(0)=X_0$, we find that

$$\begin{aligned}X(t)&=[\Phi_{21}+\Phi_{22}X_0][\Phi_{11}+\Phi_{12}X_0]^{-1}\\ &=\left[\Phi_{11}^{\mathrm{T}}+X_0\Phi_{12}^{\mathrm{T}}\right]^{-1}[\Phi_{21}^{\mathrm{T}}+X_0\Phi_{22}^{\mathrm{T}}]\\ &=X(t)^{\mathrm{T}}.\end{aligned} \tag{2.12}$$

Theorem 2.4 (Most symmetry-preserving bilinear rational functions are two-sided)**.** *Let Φ be a $2n$-by-$2n$ matrix partitioned as in* (2.11)*. Suppose that there is at least one $n\times n$ matrix G such that $\Phi_{11}+\Phi_{12}G$ is invertible. Then*

$$[\Phi_{21}+\Phi_{22}G][\Phi_{11}+\Phi_{12}G]^{-1}\equiv\left[\Phi_{11}^{T}+G\Phi_{12}^{T}\right]^{-1}[\Phi_{21}^{T}+G\Phi_{22}^{T}]\ \textit{is symmetric} \tag{2.13}$$

for every $n\times n$ symmetric matrix $G=G^{T}$ for which the matrix inverse exists if and only if

$$\begin{pmatrix}\Phi_{22}^{T}&-\Phi_{12}^{T}\\-\Phi_{21}^{T}&\Phi_{11}^{T}\end{pmatrix}\begin{pmatrix}\Phi_{11}&\Phi_{12}\\\Phi_{21}&\Phi_{22}\end{pmatrix}=\mu I \tag{2.14}$$

for some scalar μ.

Moreover, $\mu\neq 0$ if and only if both sides of (2.13) *actually vary with G.*

Proof. The first claim is proved in the same way as for Theorem 2.1, except for running G through all symmetric matrices of the apt dimension with at most two nonzero elements.

The second claim follows from the observation that (2.14) amounts to $(J^{-1}\Phi^{\mathrm{T}}J)\cdot\Phi=\mu I$. If $\mu=0$, then the rank of Φ cannot exceed the nullity of $J^{-1}\Phi^{\mathrm{T}}J$ which is the same as that of Φ, and therefore cannot exceed n. Since the rank of (Φ_{11},Φ_{12}) is n because $[\Phi_{11}+\Phi_{12}G]^{-1}$ exists for some G, the rank of Φ is n, too. Then $(\Phi_{21},\Phi_{22})=H(\Phi_{11},\Phi_{12})$ for some square H, and thus

$$[\Phi_{21}+\Phi_{22}G][\Phi_{11}+\Phi_{12}G]^{-1}=H$$

regardless of G. On the other hand, if both sides of (2.13) do not vary with G, then $[\Phi_{21}+\Phi_{22}G][\Phi_{11}+\Phi_{12}G]^{-1}\equiv H$ for some constant symmetric matrix H with respect to G, and therefore $\Phi_{21}+\Phi_{22}G\equiv H[\Phi_{11}+\Phi_{12}G]$ for every G in a nonempty open set of symmetric G, and consequently for every symmetric G. This leads to $\Phi_{21}=H\Phi_{11}$ and $\Phi_{22}=H\Phi_{12}$. Substitute both relations into the left-hand side of (2.14) to conclude that it holds with $\mu=0$. □

Symmetric MRDEs also have the following monotonicity property which shares some resemblance with [17, Theorem 2] due originally to [38]; but there are differences in the conditions. For example, there is no requirement for both A_{12} and A_{21} to be positive semidefinite here; though this condition would guarantee that the solution of the symmetric MRDE exist for all time [16, Proposition 1]. We shall write $M\prec W$ to mean that $W-M$ is positive definite and likewise $M\preceq W$ to mean that $W-M$ is positive semidefinite.

Theorem 2.5. *For real symmetric* (MRDE), *i.e., A is real and* (2.1) *holds, let $X(t)$ and $\widetilde{X}(t)$ be its two solutions with initial values $X(0)$ and $\widetilde{X}(0)$, respectively. If $X(0) \preceq \widetilde{X}(0)$, then $X(t) \preceq \widetilde{X}(t)$ over the interval* $[0, T)$ *of their existence.*

Proof. It can be proved similarly to [17, Theorem 2]. For completeness, we present a proof here. Let $W \equiv W(t) = \widetilde{X}(t) - X(t)$. It can be verified that

$$W' = W \left[A_{22} - \frac{X + \widetilde{X}}{2} A_{12} \right]^{\mathrm{T}} + \left[A_{22} - \frac{X + \widetilde{X}}{2} A_{12} \right] W$$

whose solution can be written as $W(t) = \Psi(t) W(0) \Psi(t)^{\mathrm{T}}$, where $\Psi(t)$ is the solution of

$$\Psi' = \left[A_{22} - \frac{X + \widetilde{X}}{2} A_{12} \right] \Psi, \quad \Psi(0) = I.$$

Therefore, $W(t)$ must remain positive semidefinite over the interval $[0, T)$ in which both $X(t)$ and $\widetilde{X}(t)$ have no singularity. □

We should prefer

> ***Numerical methods that retain these properties: Symmetry and Two-Sided Bilinear Rational dependence on*** $X(0)$ ***and Monotonicity in the sense of Theorem 2.5, in their computed solutions $\boldsymbol{X}$***, other things being equal.

Observe also that bilinear rational functions that propagate matrix symmetry are closed under composition.

Remark 2.1. For Hermitian MRDEs, everything in this subsection holds, after conjugate transposes $(\,\cdot\,)^*$ replace transposes $(\,\cdot\,)^{\mathrm{T}}$.

3. Unconventional anadromic numerical methods

We shall start by presenting two simple second order anadromic numerical methods to pave the way for our general format of higher order ones which also fall into the framework of *modified integrators* in [8].

3.1. Two simple second order methods.

These methods are based on a partition technique. Consider one step of integration from τ to $\tau + \theta$, where θ is the current step-size. Define the matrix-valued function $f(\mathcal{X}, \mathcal{Y})$ as

$$\boxed{f(\mathcal{X}, \mathcal{Y}) = \boldsymbol{A}_{21} - \mathcal{X}\boldsymbol{A}_{11} + \boldsymbol{A}_{22}\mathcal{Y} - \mathcal{X}\boldsymbol{A}_{12}\mathcal{Y}, \text{ where all } \boldsymbol{A}_{ij} = A_{ij}(\tau + \tfrac{1}{2}\theta).} \tag{3.1}$$

Let $\boldsymbol{X} \approx X(\tau)$. An approximation $\boldsymbol{Z} \approx X(\tau + \theta)$ can be computed by solving

$$\frac{\boldsymbol{Y} - \boldsymbol{X}}{\theta/2} = f(\boldsymbol{X}, \boldsymbol{Y}), \ \frac{\boldsymbol{Z} - \boldsymbol{Y}}{\theta/2} = f(\boldsymbol{Z}, \boldsymbol{Y}), \tag{3.2}$$

where $\boldsymbol{Y} \approx X(\tau + \frac{1}{2}\theta)$ and $\boldsymbol{Z} \approx X(\tau + \theta)$. This defines a relatively inexpensive second order *Anadromic* method

$$\textit{Updating Formula:} \quad \boldsymbol{Z} = \boldsymbol{Q}(\theta, \tau + \tfrac{1}{2}\theta, \boldsymbol{X}) \tag{3.3}$$

that preserves all three properties, namely, *Bilinear Relation*, *Generalized Inverse Property*, and *Symmetry*, as we shall prove. It is relatively inexpensive because the determining equations for $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are linear in $\boldsymbol{Y}$ and $\boldsymbol{Z}$, unlike preexisting implicit methods which have had to solve nonlinear matrix equations. By *Anadromic*, we mean that if we integrate the MRDE backwards from $\tau+\theta$ to τ using the same updating formula with the negative step size $-\theta$ and with $\boldsymbol{Z}$ in place of $X(\tau+\theta)$, in the absence of rounding errors $\boldsymbol{X}$ is recovered back exactly, namely

$$\boldsymbol{X} \equiv \boldsymbol{Q}(-\theta, (\tau+\theta) - \tfrac{1}{2}\theta, \boldsymbol{Z}).$$

An anadromic method has many attractive properties. It is of at least second order convergence:

$$[\boldsymbol{Q}(\theta, \tau + \tfrac{1}{2}\theta, X(\tau)) - X(\tau+\theta)]/\theta = O(\theta^2).$$

In fact, more can be said. Let $\boldsymbol{Z}(\tau)$ be the computed solution with $\boldsymbol{Z}(0) = X(0)$ at $t=\tau$ which is fixed and a multiple of θ. Then [22, p. 222]

$$\boldsymbol{Z}(\tau) - X(\tau) = E_2(\tau)\theta^2 + E_4(\tau)\theta^4 + E_6(\tau)\theta^6 + \cdots ,$$

i.e, its error's asymptotic expansion in terms of θ contains only even powers of θ. In the past, such a property has been considered ideal to apply (traditional) extrapolation methods [7, 14, 21] to achieve higher order approximations.

The simple second order method (3.2) is closely related to the implicit midpoint rule applied to (2.3), an observation E. Hairer [23] generously shared with the authors. The observation makes it possible to significantly simplify our earlier analysis [35] in constructing higher order methods.

Theorem 3.1 (Hairer)**.** *Let $\boldsymbol{P}_1 \approx P(\tau+\theta)$ be the solution obtained by applying the implicit midpoint rule to* (2.3) *from $t=\tau$ to $\tau+\theta$:*

$$\frac{\boldsymbol{P}_1 - \boldsymbol{P}}{\theta} = \boldsymbol{A}\frac{\boldsymbol{P}_1+\boldsymbol{P}}{2}, \quad \boldsymbol{P} = \begin{array}{cc} & m \\ \begin{matrix} m \\ n \end{matrix} & \begin{pmatrix} \boldsymbol{S} \\ \boldsymbol{T} \end{pmatrix} \end{array} \approx P(\tau), \quad \boldsymbol{P}_1 = \begin{array}{cc} & m \\ \begin{matrix} m \\ n \end{matrix} & \begin{pmatrix} \boldsymbol{S}_1 \\ \boldsymbol{T}_1 \end{pmatrix} \end{array} \approx P(\tau+\theta),$$

where $\boldsymbol{A} = A(\tau + \frac{1}{2}\theta)$. Assume that $\boldsymbol{S}$, $\boldsymbol{S}_1$, and $\boldsymbol{S}+\boldsymbol{S}_1$ are invertible. If $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{S}^{-1}$, then

$$\boldsymbol{Y}\frac{\boldsymbol{S}+\boldsymbol{S}_1}{2} = \frac{\boldsymbol{T}+\boldsymbol{T}_1}{2}, \quad \boldsymbol{Z} = \boldsymbol{T}_1\boldsymbol{S}_1^{-1}, \tag{3.4}$$

where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are defined by (3.2)*.*

Proof. It suffices to prove that $\widehat{\boldsymbol{Y}} = [(\boldsymbol{T}+\boldsymbol{T}_1)/2][(\boldsymbol{S}+\boldsymbol{S}_1)/2]^{-1}$ and $\widehat{\boldsymbol{Z}} = \boldsymbol{T}_1\boldsymbol{S}_1^{-1}$ satisfy both defining equations in (3.2) for $\boldsymbol{Y}$ and $\boldsymbol{Z}$. We have the following identities:

$$\begin{aligned}
(\widehat{\boldsymbol{Y}} - \boldsymbol{X})\frac{\boldsymbol{S}_1+\boldsymbol{S}}{2} &= \frac{\boldsymbol{T}+\boldsymbol{T}_1}{2} - \boldsymbol{X}\,\frac{\boldsymbol{S}_1 - \boldsymbol{S} + 2\boldsymbol{S}}{2} \\
&= \frac{\boldsymbol{T}+\boldsymbol{T}_1}{2} - \boldsymbol{X}\,\frac{\boldsymbol{S}_1-\boldsymbol{S}}{2} - \boldsymbol{T} \\
&= \frac{\boldsymbol{T}_1-\boldsymbol{T}}{2} - \boldsymbol{X}\,\frac{\boldsymbol{S}_1-\boldsymbol{S}}{2}, \qquad (3.5)\\
(\widehat{\boldsymbol{Z}} - \widehat{\boldsymbol{Y}})\frac{\boldsymbol{S}_1+\boldsymbol{S}}{2} &= \frac{\boldsymbol{T}_1-\boldsymbol{T}}{2} - \widehat{\boldsymbol{Z}}\,\frac{\boldsymbol{S}_1-\boldsymbol{S}}{2}. \qquad (3.6)
\end{aligned}$$

Partition $\boldsymbol{A} = (\boldsymbol{A}_{ij})$ in the conformal way. The implicit midpoint rule gives

$$\begin{aligned}
\boldsymbol{S}_1 - \boldsymbol{S} &= \tfrac{1}{2}\theta[\boldsymbol{A}_{11}(\boldsymbol{S}_1+\boldsymbol{S}) + \boldsymbol{A}_{12}(\boldsymbol{T}_1+\boldsymbol{T})], \\
\boldsymbol{T}_1 - \boldsymbol{T} &= \tfrac{1}{2}\theta[\boldsymbol{A}_{21}(\boldsymbol{S}_1+\boldsymbol{S}) + \boldsymbol{A}_{22}(\boldsymbol{T}_1+\boldsymbol{T})].
\end{aligned}$$

Substitute $\boldsymbol{S}_1 - \boldsymbol{S}$ and $\boldsymbol{T}_1 - \boldsymbol{T}$ into (3.5) and (3.6), and then apply $[(\boldsymbol{S} + \boldsymbol{S}_1)/2]^{-1}$ from the right, and then divide by $\frac{1}{2}\theta$ to see that $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{Z}}$ satisfy both defining equations in (3.2) for $\boldsymbol{Y}$ and $\boldsymbol{Z}$, respectively. □

Updating formula $\boldsymbol{Q}$ above is not alone in its simplicity and preservation of the desired properties. An obvious alternative is

$$\frac{\boldsymbol{Y}_{\mathrm{a}} - \boldsymbol{X}}{\theta/2} = f(\boldsymbol{Y}_{\mathrm{a}}, \boldsymbol{X}), \ \frac{\boldsymbol{Z}_{\mathrm{a}} - \boldsymbol{Y}_{\mathrm{a}}}{\theta/2} = f(\boldsymbol{Y}_{\mathrm{a}}, \boldsymbol{Z}_{\mathrm{a}}), \tag{3.7}$$

where $\boldsymbol{Y}_{\mathrm{a}} \approx X(\tau + \frac{1}{2}\theta)$ and $\boldsymbol{Z}_{\mathrm{a}} \approx X(\tau + \theta)$. Which circumstances can favor one alternative over the other is not known at this time. Both preserve the same properties of the MRDE's solution. It comes as no surprise that this alternative method (3.7) closely relates to the implicit midpoint rule as well, except this time the rule is applied to the second linear reduction (2.4). We state the theorem but omit its proof as it is similar to Theorem 3.1.

Theorem 3.2. *Let $\boldsymbol{R}_1 \approx R(\tau+\theta)$ be the solution by applying the implicit midpoint rule to* (2.4) *from $t = \tau$ to $\tau + \theta$:*

$$\frac{\boldsymbol{R}_1 - \boldsymbol{R}}{\theta} = -\frac{\boldsymbol{R}_1 + \boldsymbol{R}}{2}\boldsymbol{A}, \ \boldsymbol{R} = \begin{array}{cc} m & n \\ \big(\boldsymbol{T}, & -\boldsymbol{S}\big) \end{array} \approx R(\tau), \ \boldsymbol{R}_1 = \begin{array}{cc} m & n \\ \big(\boldsymbol{T}_1, & -\boldsymbol{S}_1\big) \end{array} \approx R(\tau+\theta),$$

where $\boldsymbol{A} = A(\tau + \frac{1}{2}\theta)$. Assume that $\boldsymbol{S}$, $\boldsymbol{S}_1$, and $\boldsymbol{S} + \boldsymbol{S}_1$ are invertible. If $\boldsymbol{X} = \boldsymbol{S}^{-1}\boldsymbol{T}$, then

$$\frac{\boldsymbol{S} + \boldsymbol{S}_1}{2}\boldsymbol{Y}_a = \frac{\boldsymbol{T} + \boldsymbol{T}_1}{2}, \quad \boldsymbol{Z}_a = \boldsymbol{S}_1^{-1}\boldsymbol{T}_1, \tag{3.8}$$

where $\boldsymbol{Y}_a$ and $\boldsymbol{Z}_a$ are defined by (3.7)*.*

The method defined by (3.2) appeared in 1980s in an unpublished note of the second author, and was also discovered *independently* by Babuška and Majer [4], where its second order convergence was proved by a brute force verification. But its many properties discussed in Section 4 and which are critical to our effort to integrate MRDEs to pass poles were not known, much less exploited.

3.2. **Higher order methods.** Higher order approximations have been derived from the two simple anadromic methods (3.2) and (3.7) in at least two different ways: extrapolation [7, 14] and composition [24, 27]. The focus of this article is about a third method given below. This third method can be cast in the framework of modified integrators in the sense of [8]. We apply the simple second order anadromic methods in Subsection 3.1 to truncated modified versions of the differential equation (MRDE). For more on numerical integrations through modified differential equations, the interested reader is referred to [8]. Specifically, for one step of integration from τ to $\tau + \theta$, we shall seek a sequence of matrices

$$\widetilde{A}_\ell = \begin{array}{cc} & \begin{array}{cc} m & \ \ n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} A_{11,\ell} & A_{12,\ell} \\ A_{21,\ell} & A_{22,\ell} \end{pmatrix} \end{array} \quad \text{for } \ell = 0, 1, 2, \ldots, \tag{3.9}$$

depending only on $A(t)$ and its derivatives at $t = \tau + \frac{1}{2}\theta$. We then define

$$f_\ell(\mathcal{X}, \mathcal{Y}) = A_{21,\ell} - \mathcal{X}A_{11,\ell} + A_{22,\ell}\mathcal{Y} - \mathcal{X}A_{12,\ell}\mathcal{Y}, \tag{3.10}$$

matrix-valued functions having two matrix arguments $\mathcal{X}$ and $\mathcal{Y}$. Let c_ℓ be the coefficient of $t^{2\ell+1}$ in the power series of $\tanh(t)$ [1, p. 85]:

$$\begin{aligned}\sum_{\ell=0}^{\infty} c_\ell t^{2\ell+1} &= \frac{\exp(2t)-1}{\exp(2t)+1} = \tanh(t)\\ &= t - \frac{1}{3}t^3 + \frac{2}{15}t^5 - \frac{17}{315}t^7 + \frac{62}{2835}t^9 + O(t^{11}).\end{aligned} \tag{3.11}$$

Finally, our $(2k)$th order numerical method takes the form

$$\frac{\boldsymbol{Y}-\boldsymbol{X}}{\theta/2} = \sum_{\ell=0}^{k-1}(\tfrac{1}{2}\theta)^{2\ell}c_\ell f_\ell(\boldsymbol{X},\boldsymbol{Y}),\ \frac{\boldsymbol{Z}-\boldsymbol{Y}}{\theta/2} = \sum_{\ell=0}^{k-1}(\tfrac{1}{2}\theta)^{2\ell}c_\ell f_\ell(\boldsymbol{Z},\boldsymbol{Y}) \tag{3.12}$$

where $\boldsymbol{Y} \approx X(\tau + \frac{1}{2}\theta)$ and $\boldsymbol{Z} \approx X(\tau+\theta)$. This defines a consistent numerical method for solving (MRDEs) so long as

$$\lim_{\theta\to 0} A_{ij,0} = A_{ij}(\tau) \quad \text{for } i,\ j \in \{1,2\}.$$

In particular, if $A_{ij,0} = A_{ij}(\tau + \frac{1}{2}\theta)$, it has second order convergence because $\boldsymbol{Z}$ by (3.12) differs from the one by (3.1) and (3.2) by $O(\theta^3)$. In seeking $A_{ij,\ell}$ later, care is taken so that (3.12) defines an anadromic method of order $2k$. In particular,

$f_0(\mathcal{X},\mathcal{Y})$ ia always taken to be the same as the $f(\mathcal{X},\mathcal{Y})$ in (3.1), i.e., $A_{ij,0}$ is $A_{ij}(t)$ evaluated at $\tau + \frac{1}{2}\theta$.

What remains is to find what $f_\ell(\mathcal{X},\mathcal{Y})$ should be for $\ell \geq 1$.

It is important to notice that the determining equations for $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are again linear in $\boldsymbol{Y}$ and $\boldsymbol{Z}$, making the method easy to implement.

The entry of the coefficients c_ℓ may seem mysterious at first. In a way, it is a demonstration of the close link between (MRDE) and the linear ODE (2.3), especially for the time-invariant A. See (3.18) below.

In accordance with [8], the modified differential equation of (MRDE) to which the application of (3.2) from $t=\tau$ to $\tau+\theta$ yields $\boldsymbol{Z} = X(\tau+\theta)$ exactly is

$$\widetilde{X}' = \widetilde{A}_{21} - \widetilde{X}\widetilde{A}_{11} + \widetilde{A}_{22}\widetilde{X} - \widetilde{X}\widetilde{A}_{12}\widetilde{X},\ \widetilde{X}(\tau) = X(\tau), \tag{3.13}$$

where

$$\widetilde{A} = \sum_{\ell=0}^{\infty}(\tfrac{1}{2}\theta)^{2\ell}c_\ell \widetilde{A}_\ell \equiv \begin{pmatrix}\widetilde{A}_{11} & \widetilde{A}_{12}\\ \widetilde{A}_{21} & \widetilde{A}_{22}\end{pmatrix}. \tag{3.14}$$

The method (3.12) is simply a result of an application of (3.2) to (3.13) with $\widetilde{A}$ truncated. The corresponding modified differential equation of (2.3) for the implicit midpoint rule is

$$\frac{d\widetilde{P}}{dt} = \widetilde{A}\widetilde{P}, \tag{3.15}$$

and at the same time (3.12) is related, in much the same way as stated in Theorem 3.1 for (3.2), to the implicit midpoint rule applied to (3.15) with $\widetilde{A}$ truncated.

The method (3.12) is not alone in its simplicity and preservation of the desired properties, either, as we commented before about (3.2). An obvious alternative is

$$\frac{\boldsymbol{Y}_{\rm a}-\boldsymbol{X}}{\theta/2} = \sum_{\ell=0}^{k-1}(\tfrac{1}{2}\theta)^{2\ell}c_\ell f_\ell(\boldsymbol{Y}_{\rm a},\boldsymbol{X}),\ \frac{\boldsymbol{Z}_{\rm a}-\boldsymbol{Y}_{\rm a}}{\theta/2} = \sum_{\ell=0}^{k-1}(\tfrac{1}{2}\theta)^{2\ell}c_\ell f_\ell(\boldsymbol{Y}_{\rm a},\boldsymbol{Z}_{\rm a}), \tag{3.16}$$

where $\boldsymbol{Y}_{\mathrm{a}} \approx X(\tau + \frac{1}{2}\theta)$ and $\boldsymbol{Z}_{\mathrm{a}} \approx X(\tau + \theta)$. This alternative method can also be cast as a modified integrator of (3.7) for (MRDE). It is, not surprisingly, related to the implicit midpoint rule applied to

$$\frac{d\widetilde{R}}{dt} = -\widetilde{R}\widetilde{A} \tag{3.17}$$

after $\widetilde{A}$ is truncated.

This close relationship between the proposed higher order methods and the implicit midpoint rule plays an instrumental role in simplifying our earlier constructions in [35] of these higher order methods. That is, the sought $\widetilde{A}$ for our methods (3.12) (and for (3.16), too) is the same as the one in the modified differential equation (3.15) for the implicit midpoint rule. The latter is detailed in [36]. We summarize the findings in what follows, broken into two cases, the simpler case first.

The time-invariant case. Now A does not depend on time t. It is found in [36] that

$$\widetilde{A} = \frac{1}{\frac{1}{2}\theta} \cdot \left(e^{\theta A} - I\right)\left(e^{\theta A} + I\right)^{-1} = \sum_{\ell=0}^{\infty} (\tfrac{1}{2}\theta)^{2\ell} c_\ell A^{2\ell+1}. \tag{3.18}$$

Theorem 3.3. *Suppose A is constant. Then with $\widetilde{A}_\ell = A^{2\ell+1}$ for all ℓ,* (3.12) *defines an anadromic method of order $2k$ for* (MRDE). *Moreover, if $X(\tau) = \boldsymbol{X}$,*

$$X(\tau + \theta) = \boldsymbol{Z} + 2\,(\tfrac{1}{2}\theta)^{2k+1} c_k f_k(\boldsymbol{X}, \boldsymbol{X}) + O(\theta^{2k+2}). \tag{3.19}$$

Proof. The first claim has been proved already. The second claim follows by comparing $\boldsymbol{Z}$ to the one in (3.12) after letting $k = \infty$. □

Example 3.1. Consider a scalar time-invariant RDE, i.e., $m = n = 1$ and $X' = \alpha X^2 + \beta X + \gamma$ which can be written in the form of (MRDE):

$$X' = \gamma - X(-\beta/2 + \delta) + (\beta/2 + \delta)X - X(-\alpha)X,$$

where δ is a constant and arbitrary. The corresponding matrix is

$$A = \begin{pmatrix} -\beta/2 + \delta & -\alpha \\ \gamma & \beta/2 + \delta \end{pmatrix}. \tag{3.20}$$

It can be verified that

$$A^2 = \begin{pmatrix} (-\beta/2 + \delta)^2 - \alpha\gamma & -2\alpha\delta \\ 2\gamma\delta & (\beta/2 + \delta)^2 - \alpha\gamma \end{pmatrix}.$$

In particular, $A^2 = (\beta^2/4 - \alpha\gamma)I$ if $\delta = 0$. Then for $\delta = 0$,

$$\begin{aligned} A^{2\ell+1} &= (\beta^2/4 - \alpha\gamma)^\ell A, \\ f_0(\mathcal{X}, \mathcal{Y}) &= \gamma - \mathcal{X}(-\beta/2) + (\beta/2)\mathcal{Y} - \mathcal{X}(-\alpha)\mathcal{Y}, \\ f_\ell(\mathcal{X}, \mathcal{Y}) &= (\beta^2/4 - \alpha\gamma)^\ell f_0(\mathcal{X}, \mathcal{Y}). \end{aligned}$$

Let $\Theta = \beta^2/4 - \alpha\gamma$. The first equation in (3.12) gives

$$\boldsymbol{Y} = G(\tfrac{1}{2}\theta, \boldsymbol{X}) \overset{\text{def}}{=} \frac{\left[1 + (\beta/2)\sum_{\ell=0}^{k}(\frac{1}{2}\theta)^{2\ell+1} c_\ell \Theta^\ell\right] X_0 + \gamma \sum_{\ell=0}^{k}(\frac{1}{2}\theta)^{2\ell+1} c_\ell \Theta^\ell}{1 - (\alpha X_0 + \beta/2)\sum_{\ell=0}^{k}(\frac{1}{2}\theta)^{2\ell+1} c_\ell \Theta^\ell}. \tag{3.21}$$

Since all $f_\ell(\mathcal{X},\mathcal{Y}) \equiv f_\ell(\mathcal{Y},\mathcal{X})$, the second equation in (3.12) must yield $\boldsymbol{Z} = G(\frac{1}{2}\theta,\boldsymbol{Y})$. In the case of $k=\infty$, $\boldsymbol{Z} \equiv X(\tau+\theta)$ provided $\boldsymbol{X} = X(\tau)$ by the idea of the modified differential equation. In fact, more can be said for the example (with $\delta = 0$), namely also $\boldsymbol{Y} \equiv X(\frac{1}{2}\theta)$ for $k=\infty$, too (which may fail for nonscalar RDE, however) [35, Example 7.1]. In view of this fact,

$$\boldsymbol{X} \to \frac{\left[1+(\beta/2)\sum_{\ell=0}^{k-1}\theta^{2\ell+1}c_\ell\Theta^\ell\right]\boldsymbol{X}+\gamma\sum_{\ell=0}^{k-1}\theta^{2\ell+1}c_\ell\Theta^\ell}{1-(\alpha\boldsymbol{X}+\beta/2)\sum_{\ell=0}^{k-1}\theta^{2\ell+1}c_\ell\Theta^\ell} \tag{3.22}$$

is an order $2k$ updating formula for scalar RDE $X' = \alpha X^2+\beta X+\gamma$.

Closed formulas for $\boldsymbol{Y}$, $\boldsymbol{Z}$ for the case $\delta \neq 0$ can also be established; but they are much more complicated. The key lies in computing $A^{2\ell+1}$. The matrix A in (3.20) can be written as

$$A = \delta I + B, \quad B = \begin{pmatrix} -\beta/2 & -\alpha \\ \gamma & \beta/2 \end{pmatrix}, \quad B^2 = \Theta I.$$

Therefore, we have for $m = 2\ell+1$,

$$A^m = \sum_{j=0}^{m}\binom{m}{j}\delta^{m-j}B^j = \sum_{j=0}^{\ell}\binom{m}{2j}\delta^{m-2j}\Theta^j I + \sum_{j=0}^{\ell}\binom{m}{2j+1}\delta^{m-2j-1}\Theta^j B$$

from which each entry of A^m can be explicitly written out and so is $f_\ell(\mathcal{X},\mathcal{Y})$ for each and every ℓ. We omit the detail. ◇

The time-varying case. This is a much more complicated situation than the time-invariant case. With the help of *Mathematica*, in [36] we have found $\widetilde{A}_\ell$ in terms of values of $A(t)$ and its derivatives at $t = \tau + \frac{1}{2}\theta$ for ℓ up to 4, and they yield methods in the form of (3.12) for $k = 1,2,3,4,5$ corresponding to orders 2, 4, 6, 8, and 10 of convergence. However, the complexity of $\widetilde{A}_\ell$ measured by the number of summands grows exponentially (in fact it has 4^ℓ terms). Therefore, anadromic methods of orders higher than 6 are probably impractical in general. In what follows we present $\widetilde{A}_\ell$ for ℓ up to 2. Denote for $\ell \geq 0$,

$$\boldsymbol{A}_\ell = \left.\frac{d^\ell}{dt^\ell}A(t)\right|_{t=\tau+\frac{1}{2}\theta}. \tag{3.23}$$

We have from [36]

$$\widetilde{A}_0 = \boldsymbol{A}_0 = A(\tau+\tfrac{1}{2}\theta), \tag{3.24}$$

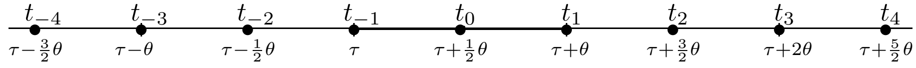
$$\widetilde{A}_1 = \boldsymbol{A}_0^3 + (\boldsymbol{A}_0\boldsymbol{A}_1 - \boldsymbol{A}_1\boldsymbol{A}_0) - \frac{1}{2}\boldsymbol{A}_2, \tag{3.25}$$

$$\begin{aligned}\widetilde{A}_2 = {} & \boldsymbol{A}_0^5 - \frac{1}{2}\boldsymbol{A}_0(\boldsymbol{A}_0\boldsymbol{A}_1 - \boldsymbol{A}_1\boldsymbol{A}_0)\boldsymbol{A}_0 + (\boldsymbol{A}_0^3\boldsymbol{A}_1 - \boldsymbol{A}_1\boldsymbol{A}_0^3) \\ & + \frac{1}{2}\left[\boldsymbol{A}_0(\boldsymbol{A}_1)^2 - 2\boldsymbol{A}_1\boldsymbol{A}_0\boldsymbol{A}_1 + (\boldsymbol{A}_1)^2\boldsymbol{A}_0\right] \\ & - \frac{1}{4}(\boldsymbol{A}_0^2\boldsymbol{A}_2 + 3\boldsymbol{A}_0\boldsymbol{A}_2\boldsymbol{A}_0 + \boldsymbol{A}_2\boldsymbol{A}_0^2) + \frac{1}{4}(\boldsymbol{A}_1\boldsymbol{A}_2 - \boldsymbol{A}_2\boldsymbol{A}_1) \\ & - \frac{1}{4}(\boldsymbol{A}_0\boldsymbol{A}_3 - \boldsymbol{A}_3\boldsymbol{A}_0) + \frac{1}{16}\boldsymbol{A}_4.\end{aligned} \tag{3.26}$$

These formulas for $\widetilde{A}_\ell$ contain (higher order) derivatives which can be hard or expensive to evaluate sometimes. In such situations, naturally, we may approximate the derivatives by divided differences. There are many ways to do so. But care must be taken in order to retain the anadromic property of the method and at the same time maintain the claimed order of convergence. Another consideration is to maximize the usage of any evaluated $A(t)$ (and possibly its low order derivatives) between consecutive steps of integration. Define

$$t_0 = \tau + \tfrac{1}{2}\theta, \quad t_{-i} = t_0 - i(\tfrac{1}{2}\theta), \quad t_i = t_0 + i(\tfrac{1}{2}\theta), \tag{3.27}$$

whose layout is shown in the following picture, where $[t_{-1}, t_1]$ is the current interval of integration:



The formulas (3.24)–(3.26) call for evaluating $A(t)$ and its derivatives at

$$\ldots, t_{-4}, t_{-2}, t_0, t_2, t_4, \ldots.$$

Our goal is to revise these formulas so that derivatives beyond (or even) the first order derivatives, are not needed. Define the first and second order divided-differences

$$A^{\dagger}(\{\alpha,\beta\}) \overset{\text{def}}{=} \frac{A(\alpha)-A(\beta)}{\alpha-\beta}, \ A^{\dagger\dagger}(\{\alpha,\beta,\gamma\}) \overset{\text{def}}{=} \frac{A^{\dagger}(\{\alpha,\beta\})-A^{\dagger}(\{\beta,\gamma\})}{\alpha-\gamma}.$$

In considering the objectives we mentioned above, we proposed in [36] the following sets of methods for derivative approximations:

$$A'(t_0) \approx A^{\dagger}(\{t_{-1}, t_1\}), \tag{3.28a}$$

$$A''(t_0) \approx 2\,A^{\dagger\dagger}(\{t_{-1}, t_0, t_1\}), \tag{3.28b}$$

$$A'(t_0) \approx A^{\dagger}(\{t_{-2}, t_2\}), \tag{3.29a}$$

$$A''(t_0) \approx 2\,A^{\dagger\dagger}(\{t_{-2}, t_0, t_2\}), \tag{3.29b}$$

$$A'''(t_0) \approx \frac{12}{\theta^2}\left[\frac{A'(t_1)+A'(t_{-1})}{2} - A^{\dagger}(\{t_{-1}, t_1\})\right], \tag{3.30a}$$

$$A^{(4)}(t_0) \approx \frac{48}{\theta^2}\left[A'^{\dagger}(\{t_{-1}, t_1\}) - 2\,A^{\dagger\dagger}(\{t_{-1}, t_0, t_1\})\right], \tag{3.30b}$$

$$A'''(t_0) \approx \frac{8}{\theta^2}\left[A^{\dagger}(\{t_{-2}, t_2\}) - A^{\dagger}(\{t_{-1}, t_1\})\right], \tag{3.31a}$$

$$A^{(4)}(t_0) \approx \frac{32}{\theta^4}\left(\frac{A(t_2)+A(t_{-2})}{2} - 4\,\frac{A(t_1)+A(t_{-1})}{2} + 3A(t_0)\right), \tag{3.31b}$$

$$A'''(t_0) \approx \frac{8}{(2\theta)^2}\left[A^{\dagger}(\{t_{-4}, t_4\}) - A^{\dagger}(\{t_{-2}, t_2\})\right], \tag{3.32a}$$

$$A^{(4)}(t_0) \approx \frac{32}{(2\theta)^4}\left(\frac{A(t_4)+A(t_{-4})}{2} - 4\,\frac{A(t_2)+A(t_{-2})}{2} + 3A(t_0)\right). \tag{3.32b}$$

Table 3.1. Anadromic method (3.12) is of order $2k$

	k	$\widetilde{A}_\ell$ and $\widetilde{b}_\ell$ for $0 \le \ell \le k-1$ given by
odr2	1	(3.24)
odr4	2	(3.24), (3.25)
odr4a	2	(3.24), (3.25) with (3.28)
odr4b	2	(3.24), (3.25) with (3.29)
odr6	3	(3.24), (3.25), (3.26)
odr6a	3	(3.24), (3.25) with (3.28), (3.34) with (3.30)
odr6b	3	(3.24), (3.25) with (3.29), (3.35) with (3.31)
odr6c	3	(3.24), (3.25) with (3.29), (3.34) with (3.32)

With them, we readily define a new $\widetilde{A}_1$ as

(3.33) New $\widetilde{A}_1$ is obtained by (3.25) with (3.28) or with (3.29) (by which we mean the first and second order derivatives $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ in (3.25) are approximated by either (3.28) or (3.29)).

Then (3.12) with $k = 2$, (3.24), and (3.33) defines anadromic methods of order 4. However, for methods of orders higher than 4, simply taking these $\widetilde{A}_\ell$ given in (3.24), (3.25), and (3.26) and replacing the derivatives of $A(t)$ at $t = t_0$ by the corresponding approximations above would not work. For example, (3.12) with $k = 2$, (3.24), (3.33), and (3.26) after (all or some of) the derivatives approximated *does not* have order 6. It turns out that using (3.25) with (3.28) or (3.29) for $\widetilde{A}_1$ affects the next $\widetilde{A}_2$.

(3.34) ***For*** (3.25) ***with*** (3.28): new $\widetilde{A}_2$ is obtained by replacing the last line of (3.26) by $\frac{1}{6}(\boldsymbol{A}_0\boldsymbol{A}_3 - \boldsymbol{A}_3\boldsymbol{A}_0) - \frac{1}{24}\boldsymbol{A}_4$;

(3.35) ***For*** (3.25) ***with*** (3.29): new $\widetilde{A}_2$ is obtained by replacing the last line of (3.26) by $\frac{17}{12}(\boldsymbol{A}_0\boldsymbol{A}_3 - \boldsymbol{A}_3\boldsymbol{A}_0) - \frac{17}{48}\boldsymbol{A}_4$.

In theory, sixth order anadromic methods are now readily available by approximating the derivatives in the new $\widetilde{A}_2$ given in (3.34) or (3.35) by any combinations of the approximation methods in (3.28)—(3.32). But again in choosing the approximations, we should keep in mind reusing the values of $A(t)$ (and possibly its first order derivatives) between consecutive steps of integration as much as possible. Table 3.1 gives our suggested anadromic methods of second, fourth, and sixth orders of convergence.

4. Preserved properties of proposed methods

In this section we shall prove that any scheme defined by (3.12) preserves all three properties discussed in Section 2, namely, *Bilinear Rational Relation*, *Generalized Inverse Property*, and *Symmetry.*

4.1. **Bilinear rational property.** Define H_{ij}, functions of θ, by

$$\sum_{\ell=0}^{k-1}(\tfrac{1}{2}\theta)^{2\ell}c_\ell\begin{pmatrix}A_{11,\ell} & A_{12,\ell}\\ A_{21,\ell} & A_{22,\ell}\end{pmatrix} = \begin{array}{cc} & \begin{array}{cc} m & \quad n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix}H_{11} & H_{12}\\ H_{21} & H_{22}\end{pmatrix}\end{array}. \tag{4.1}$$

Solving for $\boldsymbol{Y}$ and $\boldsymbol{Z}$ in (3.12) yields

$$\boldsymbol{Y} = [(2/\theta)I - (H_{22} - \boldsymbol{X}H_{12})]^{-1}[(2/\theta)\boldsymbol{X} + (H_{21} - \boldsymbol{X}H_{11})], \tag{4.2a}$$

$$\boldsymbol{Z} = [(2/\theta)\boldsymbol{Y} + (H_{21} + H_{22}\boldsymbol{Y})][(2/\theta)I + (H_{11} + H_{12}\boldsymbol{Y})]^{-1}. \tag{4.2b}$$

The formulas (4.2a) and (4.2b) (combined with Theorem 2.1 and the closure of bilinear rational functions under composition) show that *in the absence of rounding errors, the numerical solution preserves the bilinear rational property* discussed in Subsection 2.1.

Two consequences result from such preservation. The first one is the conservation of the change in rank. If $\boldsymbol{X}$ is changed, the rank of change is conserved in $\boldsymbol{Y}$ and $\boldsymbol{Z}$, like the change in the exact solution of (MRDE). See Theorem 2.2. Therefore, *the approximate solutions computed numerically by any of our anadromic methods conserve the rank of change to the initial value* provided that they use the same stepsizes θ for one approximate solution as for another approximate solution, and provided roundoff does not interfere. The second consequence is that the solution monotonicity property as in Theorem 2.5 is retained by $\boldsymbol{Z}$ (see Theorem 4.3 below).

4.2. **Generalized inverse property.** For convenience, we identify (MRDE) by its defining parameters $\{m, n, A, X_0\}$. Doing so allows us to identify its complementary (cMRDE) as one of MRDE in the form (MRDE) but with the defining parameters

$$\{n, m, A_{\mathrm{c}}, U_0\}\,, \quad \text{where } A_{\mathrm{c}} \overset{\mathrm{def}}{=} \begin{array}{cc} & \begin{array}{cc} n & \quad m \end{array} \\ \begin{array}{c} n \\ m \end{array} & \begin{pmatrix}A_{22} & A_{21}\\ A_{12} & A_{11}\end{pmatrix}\end{array}.$$

Notice that A_{c} relates to A through permuting symmetrically the blocked columns and blocked rows of A. Through identifying (cMRDE) as another (MRDE), we can apply the numerical scheme (3.12) to (cMRDE) with f_ℓ defined by

$$\begin{pmatrix}A_{22,\ell} & A_{21,\ell}\\ A_{12,\ell} & A_{11,\ell}\end{pmatrix}$$

obtained through again permuting symmetrically the blocked columns and blocked rows of matrices in (3.9). The application will lead to a numerical method for (cMRDE) as follows:

$$\boldsymbol{V} = [(2/\theta)I - (H_{11} - \boldsymbol{U}H_{21})]^{-1}[(2/\theta)\boldsymbol{U} + (H_{12} - \boldsymbol{U}H_{22})], \tag{4.3a}$$

$$\boldsymbol{W} = [(2/\theta)\boldsymbol{V} + (H_{12} + H_{11}\boldsymbol{V})][(2/\theta)I + (H_{22} + H_{21}\boldsymbol{V})]^{-1}, \tag{4.3b}$$

where $\boldsymbol{U} \approx U(\tau)$, $\boldsymbol{V} \approx U(\tau + \frac{1}{2}\theta)$ and $\boldsymbol{W} \approx U(\tau + \theta)$, and matrices H_{ij} are still defined in (4.1). Theorem 4.1 below shows that the generalized inverse property is always preserved.

Theorem 4.1. *Let $\boldsymbol{Y}$ and $\boldsymbol{Z}$ be defined by* (4.2), *and let $\boldsymbol{V}$ and $\boldsymbol{W}$ be defined by* (4.3).

(1) *If $\boldsymbol{UX} = I$ (and thus $n \geq m$), then $\boldsymbol{VY} = \boldsymbol{WZ} = I$.*

(2) *If $\boldsymbol{XU} = I$ (and thus $n \leq m$), then $\boldsymbol{YV} = \boldsymbol{ZW} = I$.*

Proof. We shall only prove item 1 since the other one can be dealt with similarly. Suppose that $\boldsymbol{UX} = I$. Write $Q_{ii} = (2/\theta)I - H_{ii}$. Then

$$\boldsymbol{VY} = [Q_{11} + \boldsymbol{U}H_{21}]^{-1}[\boldsymbol{U}Q_{22} + H_{12}][Q_{22} + \boldsymbol{X}H_{12}]^{-1}[\boldsymbol{X}Q_{11} + H_{21}].$$

We have $\boldsymbol{U}[Q_{22} + \boldsymbol{X}H_{12}] = \boldsymbol{U}Q_{22} + H_{12}$ because of $\boldsymbol{UX} = I$. Thus

$$[\boldsymbol{U}Q_{22} + H_{12}][Q_{22} + \boldsymbol{X}H_{12}]^{-1} = \boldsymbol{U},$$

and therefore

$$\boldsymbol{VY} = [Q_{11} + \boldsymbol{U}H_{21}]^{-1}\boldsymbol{U}[\boldsymbol{X}Q_{11} + H_{21}] = I.$$

To prove $\boldsymbol{WZ} = I$. Set $P_{ii} = (2/\theta)I + H_{ii}$. Then

$$\boldsymbol{WZ} = [P_{11}\boldsymbol{V} + H_{12}][P_{22} + H_{21}\boldsymbol{V}]^{-1}[P_{22}\boldsymbol{Y} + H_{21}][P_{11} + H_{12}\boldsymbol{Y}]^{-1}.$$

We have $[P_{22} + H_{21}\boldsymbol{V}]\boldsymbol{Y} = P_{22}\boldsymbol{Y} + H_{21}$ because $\boldsymbol{VY} = I$. Thus

$$[P_{22} + H_{21}\boldsymbol{V}]^{-1}[P_{22}\boldsymbol{Y} + H_{21}] = \boldsymbol{Y},$$

and therefore

$$\boldsymbol{WZ} = [P_{11}\boldsymbol{V} + H_{12}]\boldsymbol{Y}[P_{11} + H_{12}\boldsymbol{Y}]^{-1} = I,$$

as will be shown. □

4.3. **Symmetry property.** Our first few proofs, independently by the authors and David Bindel [5], of the *Symmetry Property* summarized in the following theorem were long and complicated. The proof given below is due to Hairer [23]. Also in the theorem, we impose conditions on $A_{ij,\ell}$ in (4.4). These conditions are automatically satisfied by the formulas in Subsection 3.2 if they are true for A_{ij}, as proved in [36].

Theorem 4.2. *In* (3.12), *if*

$$A_{21,\ell} = A_{21,\ell}^T,\ A_{11,\ell} = -A_{22,\ell}^T,\ A_{12,\ell} = A_{12,\ell}^T \text{ for all } \ell, \tag{4.4}$$

and $\boldsymbol{X} = \boldsymbol{X}^T$, *then* $\boldsymbol{Z} = \boldsymbol{Z}^T$.

Proof. Consider the associated linear differential equation

$$\frac{d\widehat{P}}{dt} = \widehat{A}\widehat{P}, \quad \widehat{P}(\tau) = \begin{pmatrix} \boldsymbol{S} \\ \boldsymbol{T} \end{pmatrix}, \tag{4.5}$$

where $\widehat{A} = \sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell}c_\ell\widetilde{A}_\ell$ obtained after truncating $\widetilde{A}$ in (3.14), and $\boldsymbol{X} = \boldsymbol{TS}^{-1}$. $\boldsymbol{Z}$ relates to the implicit midpoint rule solution

$$\frac{\boldsymbol{P}_1 - \boldsymbol{P}}{\theta} = \widehat{A}\frac{\boldsymbol{P}_1 + \boldsymbol{P}}{2}, \quad \boldsymbol{P} = \begin{array}{cc} & m \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} \boldsymbol{S} \\ \boldsymbol{T} \end{pmatrix} \end{array}, \quad \boldsymbol{P}_1 = \begin{array}{cc} & m \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} \boldsymbol{S}_1 \\ \boldsymbol{T}_1 \end{pmatrix} \end{array}$$

for (4.5) by $\boldsymbol{Z} = \boldsymbol{T}_1\boldsymbol{S}_1^{-1}$. The assumptions in (4.4) imply $\widehat{A}^{\mathrm{T}}J = -J\widehat{A}$, where J is defined in (2.10). It is not hard to verify that $\frac{d}{dt}[\widehat{P}(t)^{\mathrm{T}}J\widehat{P}(t)] = 0$ which says $\widehat{P}(t)^{\mathrm{T}}J\widehat{P}(t)$ is a quadratic first integral of (4.5). Since the implicit midpoint rule preserves quadratic first integral[3] [24, p.101], we have

$$\boldsymbol{P}_1^{\mathrm{T}}J\boldsymbol{P}_1 = \boldsymbol{P}^{\mathrm{T}}J\boldsymbol{P} \quad \Rightarrow \quad -\boldsymbol{S}_1^{\mathrm{T}}\boldsymbol{T}_1 + \boldsymbol{T}_1^{\mathrm{T}}\boldsymbol{S}_1 = -\boldsymbol{S}^{\mathrm{T}}\boldsymbol{T} + \boldsymbol{T}^{\mathrm{T}}\boldsymbol{S}.$$

$\boldsymbol{X}$ is symmetric; so is $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{XS} = \boldsymbol{S}^{\mathrm{T}}\boldsymbol{T}$. Thus $-\boldsymbol{S}_1^{\mathrm{T}}\boldsymbol{T}_1 + \boldsymbol{T}_1^{\mathrm{T}}\boldsymbol{S}_1 = 0$, i.e., $\boldsymbol{S}_1^{\mathrm{T}}\boldsymbol{T}_1$ is symmetric; so is $\boldsymbol{Z} = \boldsymbol{T}_1\boldsymbol{S}_1^{-1} = \boldsymbol{S}_1^{-\mathrm{T}}(\boldsymbol{S}_1^T\boldsymbol{T}_1)\boldsymbol{S}_1^{-1}$, as was to be shown. □

[3] That $\boldsymbol{P}_1^{\mathrm{T}}J\boldsymbol{P}_1 = \boldsymbol{P}^{\mathrm{T}}J\boldsymbol{P}$ can also be directly verified, by noting that $B = (I - \frac{1}{2}\theta\widehat{A})^{-1}(I + \frac{1}{2}\theta\widehat{A})$ is symplectic, i.e., $B^{\mathrm{T}}JB = J$, and $\boldsymbol{P}_1 = B\boldsymbol{P}$.

It is worth emphasizing that despite this symmetry property of our methods in the absence of rounding errors, numerically computed solutions often deviate from being symmetric (Hermitian) after many integration steps. This is what we observed in our numerical experiments and drove us to seek many different proofs of Theorem 4.2. Therefore, it is recommended to symmetrize the computed $\boldsymbol{Z}$ every few steps in implementation. Fortunately the cost of doing so is marginal, relative to the overall cost of integration.

The next theorem says $\boldsymbol{Z}$ defined by (4.2) retains the solution monotonicity property as given in Theorem 2.5. Let $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{Z}}$ be defined by (4.2) after $\boldsymbol{X}$ is changed to $\widehat{\boldsymbol{X}}$; and let θ_0 be the smallest θ for which one of the following fails to be nonsingular:

$$\begin{aligned} &(2/\theta)I - (H_{22} - \boldsymbol{X}H_{12}), \quad (2/\theta)I + (H_{11} + H_{12}\boldsymbol{Y}),\\ &(2/\theta)I - (H_{22} - \widehat{\boldsymbol{X}}H_{12}), \quad (2/\theta)I + (H_{11} + H_{12}\widehat{\boldsymbol{Y}}). \end{aligned}$$

Theorem 4.3. *For a real symmetric* (MRDE)*, assume* (4.4) *and that all $A_{ij,\ell}$ are real. If both $\boldsymbol{X}$ and $\widehat{\boldsymbol{X}}$ are real and $\boldsymbol{X} \preceq \widehat{\boldsymbol{X}}$, then $\boldsymbol{Z} \preceq \widehat{\boldsymbol{Z}}$ for $\theta \in [0, \theta_0)$.*

Proof. By Theorem 4.2, both $\boldsymbol{Z}$ and $\widehat{\boldsymbol{Z}}$ are real symmetric. For $\theta \in [0, \theta_0)$, $\widehat{\boldsymbol{Z}} - \boldsymbol{Z}$ is continuous in θ; so are its eigenvalues. Since $\operatorname{rank}(\widehat{\boldsymbol{Z}} - \boldsymbol{Z}) = \operatorname{rank}(\widehat{\boldsymbol{X}} - \boldsymbol{X})$ as θ increases from 0 to any number that is less than θ_0. Thus the eigenvalues of $\widehat{\boldsymbol{Z}} - \boldsymbol{Z}$ cannot change their signs. □

We point out that this theorem is intrinsically different from the obvious conclusion that *if $\boldsymbol{X} \prec \widehat{\boldsymbol{X}}$, then $\boldsymbol{Z} \prec \widehat{\boldsymbol{Z}}$ for sufficiently small θ.* This is so for approximate solutions because it is so for the exact solutions of a real symmetric MRDE. First for this obvious conclusion to hold, one must assume strictly $\boldsymbol{X} \prec \widehat{\boldsymbol{X}}$; and secondly how small a θ is sufficiently small depends on the smallest eigenvalue of the difference $\widehat{\boldsymbol{X}} - \boldsymbol{X}$. For this second point, θ_0 in Theorem 4.3 can be taken to be

$$\theta_0 = \min\left\{\frac{2}{\|\boldsymbol{A}_0\|(1+\|\boldsymbol{X}\|)}, \frac{2}{\|\boldsymbol{A}_0\|(1+\|\widehat{\boldsymbol{X}}\|)}\right\} \tag{4.6}$$

for the second order method, where $\|\cdot\|$ is any consistent matrix norm. For methods of order $2k$, it is

$$\theta_0 = \min\{\delta(\|\boldsymbol{X}\|), \delta(\|\widehat{\boldsymbol{X}}\|)\}, \tag{4.7}$$

where $\delta(\eta)$ is the smallest positive root of

$$1 - (1+\eta)\sum_{\ell=0}^{k-1} |c_\ell| \|\widetilde{A}_\ell\| (\tfrac{1}{2}\theta)^{2\ell+1} = 0.$$

Neither θ_0 in (4.6) nor (4.7) depends on the difference $\widehat{\boldsymbol{X}} - \boldsymbol{X}$.

Remark 4.1. For Hermitian MRDEs, everything in this subsection holds, after conjugate transposes $(\,\cdot\,)^*$ replace transposes $(\,\cdot\,)^{\mathrm{T}}$.

5. The group of two-sided bilinear rational functions

Given an n-by-m matrix G of fixed perhaps unequal dimensions but indeterminate (variable) elements, the set of all invertible bilinear rational n-by-m matrix-valued functions

$$\mathscr{F}(G) \overset{\text{def}}{=} [\Phi_{21} + \Phi_{22}G][\Phi_{11} + \Phi_{12}G]^{-1},$$

where constant matrices Φ_{ij} are the submatrices of the *invertible* matrix

$$\Phi = \begin{array}{cc} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \end{array},$$

constitute a *group* $\mathfrak{BR}_{n,m}$. Its operation is composition, as we shall see in a moment. The domain of $\mathscr{F}(G)$ is nonempty because the m rows of $(\Phi_{11}\ \Phi_{12})$ must be linearly independent. Note that the semi-group of all n-by-m bilinear rational functions includes noninvertible functions corresponding to noninvertible matrices Φ, but none of these are related to MRDE all of whose linear reductions generate invertible *matrizants* Φ. Consequently, what follows is confined to the group of invertible bilinear rational functions.

Given three invertible $(m+n)$-by-$(m+n)$ matrices

$$\Phi_j = \begin{array}{cc} & \begin{array}{cc} m & n \end{array} \\ \begin{array}{c} m \\ n \end{array} & \begin{pmatrix} \Phi_{11,j} & \Phi_{12,j} \\ \Phi_{21,j} & \Phi_{22,j} \end{pmatrix} \end{array} \quad \text{for } j = 1,\ 2 \text{ and } 3$$

corresponding respectively to three bilinear rational matrix functions

$$\mathscr{F}_j(G) \overset{\text{def}}{=} [\Phi_{21,j} + \Phi_{22,j}G][\Phi_{11,j} + \Phi_{12,j}G]^{-1},$$

we find that if $\Phi_1 = \Phi_2\Phi_3$, then $\mathscr{F}_1(G) = \mathscr{F}_2(\mathscr{F}_3(G))$. Conversely, if $\mathscr{F}_1(G) = \mathscr{F}_2(\mathscr{F}_3(G))$, then we find $\Phi_1 = \beta\Phi_2\Phi_3$ for some nonzero scalar β. In general each bilinear rational matrix function $\mathscr{F}(G)$ in $\mathfrak{BR}_{n,m}$ is associated with a ray of invertible matrices $\beta\Phi$ generated by running β through all nonzero scalars. Therefore, the group $\mathfrak{BR}_{n,m}$ of n-by-m bilinear rational matrix functions $\mathscr{F}(G)$ is isomorphic to the *Multiplicative Quotient Group* of invertible $(m+n)$-by-$(m+n)$ matrices Φ by nonzero scalars β. This quotient group is denoted by PGL_{m+n} , which stands for the *Projective General Linear Group.* We shall drop the dimensions m and n henceforth since they will not change.

One connection between $\mathscr{F}$ in group $\mathfrak{BR}$ and $\beta\Phi$ in PGL is an equation

$$\begin{pmatrix} I \\ \mathscr{F}(G) \end{pmatrix} = \beta\Phi \begin{pmatrix} I \\ G \end{pmatrix} S^{-1} = \beta \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \begin{pmatrix} I \\ G \end{pmatrix} S^{-1}$$

in which $S = \beta[\Phi_{11} + \Phi_{12}G]$. Here the nonzero scalar β cancels away, and S is invertible except when G falls at a pole of $\mathscr{F}(G) = [\Phi_{21} + \Phi_{22}G][\Phi_{11} + \Phi_{12}G]^{-1}$, where $\det(\Phi_{11} + \Phi_{12}G) = 0$, which turns out to force $\mathscr{F}(G)$ to ∞. A second connection is an unobvious equation

$$\begin{pmatrix} \mathscr{F}(G), & -I \end{pmatrix} = \widetilde{S}^{-1} \begin{pmatrix} G, & -I \end{pmatrix} \begin{pmatrix} \widetilde{\Phi}_{11} & \widetilde{\Phi}_{12} \\ \widetilde{\Phi}_{21} & \widetilde{\Phi}_{22} \end{pmatrix},$$

where

$$\begin{pmatrix} \widetilde{\Phi}_{11} & \widetilde{\Phi}_{12} \\ \widetilde{\Phi}_{21} & \widetilde{\Phi}_{22} \end{pmatrix} = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}^{-1} \quad \text{and} \quad \widetilde{S} = -G\widetilde{\Phi}_{12} + \widetilde{\Phi}_{22},$$

which exhibits the same function $\mathscr{F}(G) = [G\widetilde{\Phi}_{12} - \widetilde{\Phi}_{22}]^{-1}[-G\widetilde{\Phi}_{11} + \widetilde{\Phi}_{21}]$ now associated with $\beta\Phi^{-1}$ in PGL. The two formulas for $\mathscr{F}$ justify the term "*two-sided*". In other words, there are two isomorphisms between the group $\mathfrak{BR}$ of bilinear rational matrix functions $\mathscr{F}$ and the rays of matrices $\beta\Phi$ and $\beta\Phi^{-1}$ (the two sets coincide) in the quotient group PGL.

The two isomorphisms supply two ways to compute $\mathscr{F}(G)$ numerically for any particular G. Sometimes one way is far more accurate than the other. For instance, if one of the matrices

$$\begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \widetilde{\Phi}_{11} & \widetilde{\Phi}_{12} \\ \widetilde{\Phi}_{21} & \widetilde{\Phi}_{22} \end{pmatrix} = \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}^{-1}$$

has just one relatively tiny singular value, this matrix can usually provide a more accurately computed value of $\mathscr{F}(G)$ than can the other matrix, all but one of whose singular values must be relatively tiny. However, sometimes neither matrix provides accurate results; such can be the case when both matrices have too many relatively tiny singular values. Such is sometimes the case for bilinear rational matrix functions that solve matrix Riccati differential equations.

Solutions $X(t)$ of the matrix Riccati differential equation (MRDE) can be regarded not so much as functions of t selected by an initial value $X(0)$, but rather as sample-values $X(t) = \mathscr{F}_t(X(0))$ of members of the group $\mathfrak{BR}$ selected by t and sampled at an indeterminate $X(0)$. Thus, as t increases from 0, the differential equation's solution $\mathscr{F}_t(\,\cdot\,)$ traces a trajectory through the group $\mathfrak{BR}$ of bilinear rational matrix functions $\mathscr{F}$ starting at the identity function $\mathscr{F}_0(\,\cdot\,)$.

But numerical methods that would compute $\mathscr{F}_t(\,\cdot\,)$ from its matrix Φ_t often encounter numerical instability. Instead, numerical methods try to compute $X(t) = \mathscr{F}_t(X(0))$ well only for a specific initial value $X(0)$. Our anadromic methods approximate $\mathscr{F}_t(X(0))$ by a sequence of sample-values $\mathscr{F}(X(0))$ each drawn from the same group $\mathfrak{BR}$ of functions $\mathscr{F}(\,\cdot\,)$, and all intended to follow nearly along the trajectory traced by $\mathscr{F}_t(X(0))$ regardless of $X(0)$.

6. Marching over poles

The updating formula defined by (3.12) preserves a generalized inverse relation between solutions of complementary MRDEs so long as their numerical solutions are computed with matching step-sizes θ small enough that all needed matrix inverses exist. But what do we gain from this? When $\boldsymbol{X}$ is square naturally we could switch between the given MRDE for $\boldsymbol{X}$ and its complementary MRDE for $\boldsymbol{U}$, whichever has no poles. This should allow us to march over poles without stopping the numerical integration unnecessarily, unlike existing methods!

What about nonsquare cases? Then such a switch mechanism cannot possibly work because neither of two nonsquare generalized inverses determines the other uniquely. Lemmas 6.1 and 6.2 below show that the updating formula will work just fine as long as we do not accidentally step on any poles, or too close to the poles lest the associated linear matrix equations in the formulas (4.2) would be too ill-conditioned to let us solve them with adequate accuracy.

Lemma 6.1. *Suppose $n \times m$ matrix $\mathcal{X}$ has full column rank, and let $\mathscr{U} = \{\mathcal{U} : \mathcal{U}\mathcal{X} = I\}$. If $\mathcal{U}\widehat{\mathcal{X}} = I$ for all $\mathcal{U}$ in a nonempty relatively open set of $\mathscr{U}$, then $\widehat{\mathcal{X}} = \mathcal{X}$, or in other words, $\mathcal{X}$ is uniquely determined by a nonempty relatively open set in the collection of its (left) generalized inverses.*

Proof. No proof is necessary if $\mathcal{X}$ is square. Suppose $\mathcal{X}$ is not square, and let $\mathcal{V}$ be from the nonempty open set of $\mathscr{U}$. So $\mathcal{V}$ is $m \times n$ and $\mathcal{V}\mathcal{X} = I$. We claim both $\mathcal{X}$ and $\mathcal{V}$ can be embedded in invertible matrices such that

$$\begin{pmatrix} \mathcal{V} \\ \check{\mathcal{V}} \end{pmatrix} (\mathcal{X}, \quad \check{\mathcal{X}}) = \begin{pmatrix} I_m & 0 \\ 0 & I_{n-m} \end{pmatrix}. \tag{6.1}$$

This is because both $\mathcal{V}$ and $\mathcal{X}$ have full rank, and therefore they can be embedded in invertible matrices, respectively,

$$\begin{pmatrix} \mathcal{V} \\ \mathcal{Y} \end{pmatrix}, \quad \text{and} \quad (\mathcal{X}, \quad \mathcal{Z}).$$

The matrix

$$\begin{pmatrix} \mathcal{V} \\ \mathcal{Y} \end{pmatrix} (\mathcal{X} \quad \mathcal{Z}) = \begin{pmatrix} I_m & \mathcal{V}\mathcal{Z} \\ \mathcal{Y}\mathcal{X} & \mathcal{Y}\mathcal{Z} \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ \mathcal{Y}\mathcal{X} & I_{n-m} \end{pmatrix} \begin{pmatrix} I_m & \mathcal{V}\mathcal{Z} \\ 0 & \mathcal{Y}\mathcal{Z} - \mathcal{Y}\mathcal{X}\mathcal{V}\mathcal{Z} \end{pmatrix}$$

is invertible, and

$$\begin{aligned} I_n &= \begin{pmatrix} I_m & 0 \\ \mathcal{Y}\mathcal{X} & I_{n-m} \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{V} \\ \mathcal{Y} \end{pmatrix} (\mathcal{X} \quad \mathcal{Z}) \begin{pmatrix} I_m & \mathcal{V}\mathcal{Z} \\ 0 & \mathcal{Y}\mathcal{Z} - \mathcal{Y}\mathcal{X}\mathcal{V}\mathcal{Z} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} I_m & 0 \\ -\mathcal{Y}\mathcal{X} & I_{n-m} \end{pmatrix} \begin{pmatrix} \mathcal{V} \\ \mathcal{Y} \end{pmatrix} (\mathcal{X} \quad \mathcal{Z}) \begin{pmatrix} I_m & -\mathcal{V}\mathcal{Z}(\mathcal{Y}\mathcal{Z} - \mathcal{Y}\mathcal{X}\mathcal{V}\mathcal{Z})^{-1} \\ 0 & (\mathcal{Y}\mathcal{Z} - \mathcal{Y}\mathcal{X}\mathcal{V}\mathcal{Z})^{-1} \end{pmatrix}. \end{aligned}$$

Now take

$$\check{\mathcal{V}} = \mathcal{Y} - \mathcal{Y}\mathcal{X}\mathcal{V}, \quad \check{\mathcal{X}} = (\mathcal{Z} - \mathcal{X}\mathcal{V}\mathcal{Z})(\mathcal{Y}\mathcal{Z} - \mathcal{Y}\mathcal{X}\mathcal{V}\mathcal{Z})^{-1}$$

to get (6.1). We have, by (6.1), $\mathcal{X}\mathcal{V} + \check{\mathcal{X}}\check{\mathcal{V}} = I$. If $\mathcal{U}\mathcal{X} = I$ too, then

$$\mathcal{U} = \mathcal{U}(\mathcal{X}\mathcal{V} + \check{\mathcal{X}}\check{\mathcal{V}}) = \mathcal{V} + (\mathcal{U}\check{\mathcal{X}})\check{\mathcal{V}}.$$

On the other hand every matrix of the form $\mathcal{U} \overset{\text{def}}{=} \mathcal{V} + G\check{\mathcal{V}}$ satisfies $\mathcal{U}\mathcal{X} = I$. Therefore, the nonempty relatively open set of all solutions $\mathcal{U}$ of $\mathcal{U}\mathcal{X} = I$ contains all $\mathcal{V} + G\check{\mathcal{V}}$ as G runs through a corresponding nonempty open set of all $n \times (n-m)$ matrices. If $\widehat{\mathcal{X}}$ satisfies

$$(\mathcal{V} + G\check{\mathcal{V}})\widehat{\mathcal{X}} = I = (\mathcal{V} + G\check{\mathcal{V}})\mathcal{X}$$

for all G in the nonempty open set of $n \times (n-m)$ matrices, then these equations must hold for all $n \times (n-m)$ matrices because they constrain in the entries of G linearly. Therefore, $\mathcal{V}(\widehat{\mathcal{X}} - \mathcal{X}) = 0$ and $G\check{\mathcal{V}}(\widehat{\mathcal{X}} - \mathcal{X}) = 0$ for all $n \times (n-m)$ matrices G, whence $\check{\mathcal{V}}(\widehat{\mathcal{X}} - \mathcal{X}) = 0$, and finally

$$\widehat{\mathcal{X}} - \mathcal{X} = I(\widehat{\mathcal{X}} - \mathcal{X}) = (\mathcal{X}\mathcal{V} + \check{\mathcal{X}}\check{\mathcal{V}})(\widehat{\mathcal{X}} - \mathcal{X}) = 0,$$

as expected. □

Similarly we have

Lemma 6.2. *Suppose $n \times m$ matrix $\mathcal{X}$ has full row rank, and let $\mathscr{U} = \{\mathcal{U} : \mathcal{X}\mathcal{U} = I\}$. If $\widehat{\mathcal{X}}\mathcal{U} = I$ for all $\mathcal{U}$ in a nonempty relatively open set of $\mathscr{U}$, then $\widehat{\mathcal{X}} = \mathcal{X}$, or in other words, $\mathcal{X}$ is uniquely determined by a nonempty relatively open set in the collection of its (right) generalized inverses.*

Lemmas 6.1 and 6.2, together with the generalized inverse property, guarantee that ***the computed approximation at $\tau + \theta$ by* (4.2) *is still meaningful, even if there are poles in between τ and $\tau + \theta$!*** We shall now explain. Suppose that $\boldsymbol{X} \approx X(\tau)$ has full rank, and let $\mathscr{U}$ be a nonempty relatively open

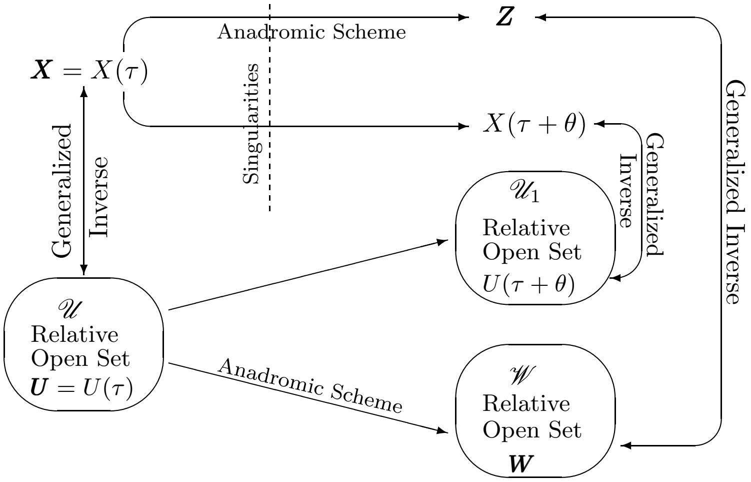


FIGURE 6.1. $\mathscr{U}_1$ and $\mathscr{W}$ being pointwise close and the generalized inverse relations imply that $\boldsymbol{Z}$ and $X(\tau+\theta)$ must be close as well, despite the fact that $X(t)$ may have pole singularities in the interval $(\tau, \tau+\theta)$.

set of $\boldsymbol{X}$'s generalized inverses. Assume that the solution to the complementary (cMRDE) from $t = \tau$ to $\tau+\theta$ with $U(\tau)$ taking any matrix in $\mathscr{U}$ has no singularity. Formulas (4.2) produce $\boldsymbol{Z}$ while formulas (4.3) with any $\boldsymbol{U}$ in $\mathscr{U}$ produces $\boldsymbol{W}$ which approximates $U(\tau+\theta)$ well as determined by the size of θ, provided $\boldsymbol{U} = U(\tau)$. If θ is small enough, the continuity of the solution of the complementary (cMRDE) with respect to the initial values and the continuity of $\boldsymbol{W}$ as defined by (4.3) with respect to $\boldsymbol{U}$ imply that

$$\mathscr{U}_1 \overset{\text{def}}{=} \{U(\tau+\theta) : U(\tau) = \boldsymbol{U} \in \mathscr{U}\} \quad \text{and} \quad \mathscr{W} \overset{\text{def}}{=} \{\boldsymbol{W} : \boldsymbol{U} \in \mathscr{U}\}$$

are nonempty relatively open sets of the generalized inverses of $X(\tau+\theta)$ and of $\boldsymbol{Z}$, respectively, because of Theorems 2.3 and 4.1. Since $\mathscr{W}$ approximates $\mathscr{U}_1$ pointwise as determined by the size of θ, and $X(\tau+\theta)$ is uniquely determined by $\mathscr{U}_1$ and $\boldsymbol{Z}$ by $\mathscr{W}$, we conclude that $\boldsymbol{Z}$ must approximate $X(\tau+\theta)$ well. What happens if $\boldsymbol{X}$ does not have full rank? We can then perturb $\boldsymbol{X}$ arbitrarily little so that the perturbed $\boldsymbol{X}$ has full rank. Assume the solution to (MRDE) from $t = \tau$ to $\tau+\theta$ is continuously dependent on $X(\tau)$ even across the possible singularities within the interval. Then the limiting argument (i.e., by letting the perturbations to $\boldsymbol{X}$ go to zero) will lead to the same conclusion, i.e., $\boldsymbol{Z}$ approximates $X(\tau+\theta)$ well. Figure 6.1 presents a pictorial view of our argument here, where the relatively open sets $\mathscr{U}_1$ and $\mathscr{W}$ are pointwise close and the generalized inverse relations imply that $\boldsymbol{Z}$ and $X(\tau+\theta)$ must be close to the extent comparable to the pointwise closeness between $\mathscr{U}_1$ and $\mathscr{W}$.

In this explanation of (4.2) being able to march over the poles, we made two critical assumptions:

Assumption 1: From $t = \tau$ to $\tau+\theta$ the solutions of the complementary (cMRDE) are well behaved with $U(\tau)$ taking values in a nonempty relatively open set of the generalized inverses of $\boldsymbol{X}$.

Assumption 2: The solution to (MRDE) from $t=\tau$ to $\tau+\theta$ is continuously dependent on $X(\tau)=\boldsymbol{X}$ even across the possible singularities within the interval. This is needed only when $\boldsymbol{X}$ does not have full rank.

We point out that the conditioning of associated linear systems must be monitored to prevent them from becoming too ill-conditioned. This is a well-understood issue in Numerical Linear Algebra and there are existing tools for the purpose; see [13, 2, 20, 25, 26, 43, 45] and references therein.

7. A linear stability theory

A linear stability theory for our methods can be established. Not surprisingly, it coincides with the one in [36] for the modified implicit midpoint rules. Consider

$$x' = \lambda x \equiv -x(-\lambda_1) + x\lambda_2, \quad x(0)=x_0, \tag{7.1}$$

where $\lambda=\lambda_1+\lambda_2$ and both λ_i have the same sign. It takes the form of (MRDE) with

$$m=n=1, \quad A=\begin{pmatrix} -\lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix},$$

and it is a time-invariant MRDE for which we know all $A_{ij,\ell}$ to give anadromic numerical methods of any even order. Since A is diagonal,

$$A^j=\begin{pmatrix} (-\lambda_1)^j & 0 \\ 0 & \lambda_2^j \end{pmatrix}, \quad f_\ell(\mathcal{X},\mathcal{Y}) = -\mathcal{X}\,(-\lambda_1)^{2\ell+1}+\lambda_2^{2\ell+1}\mathcal{Y}.$$

Therefore (3.12) yields

$$\begin{aligned} \boldsymbol{Y} &= \frac{1-\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell(-\lambda_1)^{2\ell+1}}{1-\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell\lambda_2^{2\ell+1}}\boldsymbol{X}, \\ \boldsymbol{Z} &= \frac{1+\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell\lambda_2^{2\ell+1}}{1+\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell(-\lambda_1)^{2\ell+1}}\boldsymbol{Y} \\ &= \frac{1+\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell\lambda_2^{2\ell+1}}{1-\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell\lambda_2^{2\ell+1}}\cdot\frac{1-\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell(-\lambda_1)^{2\ell+1}}{1+\sum_{\ell=0}^{k-1}(\frac{1}{2}\theta)^{2\ell+1}c_\ell(-\lambda_1)^{2\ell+1}}\boldsymbol{X}. \end{aligned} \tag{7.2}$$

As $k\to\infty$, $\boldsymbol{Z}$ goes to

$$\begin{aligned} &\left[\mathscr{C}\left(\sum_{\ell=0}^{\infty}(\tfrac{1}{2}\theta)^{2\ell+1}c_\ell\lambda_2^{2\ell+1}\right)\right]^{-1}\mathscr{C}\left(\sum_{\ell=0}^{\infty}(\tfrac{1}{2}\theta)^{2\ell+1}c_\ell(-\lambda_1)^{2\ell+1}\right)\boldsymbol{X} \\ &\qquad = e^{\theta\lambda_2}\,e^{-[\theta(-\lambda_1)]}\boldsymbol{X} = e^{\theta(\lambda_1+\lambda_2)}\boldsymbol{X}, \end{aligned}$$

where $\mathscr{C}(\cdot)$ denotes the Cayley transform[4]. This suggests that the magnitude of

$$\rho_{2k}(\mu) \overset{\text{def}}{=} \frac{1+\sum_{\ell=0}^{k-1}c_\ell\mu^{2\ell+1}}{1-\sum_{\ell=0}^{k-1}c_\ell\mu^{2\ell+1}} \tag{7.3}$$

should provide a quantitative measure to the linear stability of our methods. Thus we define the *region of stability* of the $(2k)$th order method to be

$$\mathscr{R}_{2k} = \{\mu : |\rho_{2k}(\mu)|\le 1\} \tag{7.4}$$

[4] Given square matrix Γ such that $I+\Gamma$ is nonsingular, the *Cayley Transform* of Γ is defined as $\mathscr{C}(\Gamma) \overset{\text{def}}{=} (I-\Gamma)(I+\Gamma)^{-1}$. The Cayley transform is an involution, i.e., $\mathscr{C}(\mathscr{C}(\Gamma))=\Gamma$.

which is exactly the same as the one for the modified implicit midpoint rules in [36]. Ideally, $\mathscr{R}_{2k}$ should contain only those μ with $\Re(\mu) \le 0$, and none of those μ with $\Re(\mu) > 0$. It turns out that only $\mathscr{R}_2$ has this property. When

$$\tfrac{1}{2}\theta\lambda = \frac{\theta\lambda}{2} \in \mathscr{R}_{2k}, \tag{7.5}$$

$\boldsymbol{Z}$ by (7.2) has nonincreasing magnitude. Therefore it is important to make sure (7.5) holds for λ with $\Re(\lambda) \le 0$ by using the appropriate step-size θ. What this means for (MRDE) is that at any point τ of integration, we need to use these θ satisfying (7.5) for all

$$\lambda = -\lambda_1 + \lambda_2, \quad \lambda_1 \in \operatorname{eig}(A_{11} + A_{12}X), \quad \lambda_2 \in \operatorname{eig}(A_{22} - XA_{12}),$$

with $\Re(\lambda) \le 0$, where $\operatorname{eig}(\,\cdot\,)$ is the set of eigenvalues of a matrix.

The contour plots of $|\rho_{2k}(\mu)|$ lying in $\mathscr{R}_{2k}$ can be found in [36]. They imply

- $\mathscr{R}_2$ contains[5] only those μ with $\Re(\mu) \le 0$, and none of those μ with $\Re(\mu) > 0$, while all other $\mathscr{R}_{2k}$ for $k \ge 1$ do not have this property, unfortunately.
- Restricted to real μ, $\mathscr{R}_{2k}$ with odd k has the perfect stability property, i.e., $|\rho_{2k}(\mu)| \le 1$ for $\mu \le 0$, and $|\rho_{2k}(\mu)| > 1$ for $\mu > 0$. This is important for the case when all eigenvalues are real because then only real μ is of interest. One of our examples in the next section does have all real negative eigenvalues.

8. Numerical examples

In this section, we shall present a few numerical tests to demonstrate the capability of our methods, especially when the solution has singularities. The tests are for constant step-sizes. In a way, a robust variable step-size implementation would be more efficient. As this paper is already long, we decide to explore robust variable step-size implementations elsewhere. Notice that what we have established in this paper naturally offers two approaches to pursue—using local truncation errors or running two methods of different orders (e.g., second and fourth, fifth and sixth) to estimate local errors to vary the stepsizes.

Example 1. This is a scalar time-varying MRDE

$$x' = t + x^2 \quad \text{for all } t \ge 0,\ x(0) = 0 \tag{8.1}$$

whose solution is expressible in terms of Bessel functions:

$$x(t) = -\frac{d}{dt}\ln\left(\sqrt{t}J_{-1/3}(2t^{3/2}/3)\right) = \sqrt{t}\frac{J_{2/3}(2t^{3/2}/3)}{J_{-1/3}(2t^{3/2}/3)}, \tag{8.2}$$

and it has lots of poles. See the top left plot in Figure 8.1. The corresponding A is linear, making our methods for the time-varying case rather easy to apply. In fact, the matrices in (3.24), (3.25), and (3.26) are

$$\begin{pmatrix} 0 & -1 \\ t & 0 \end{pmatrix}, \quad \begin{pmatrix} -1 & t \\ -t^2 & 1 \end{pmatrix}, \quad \begin{pmatrix} 3t/2 & -t^2 \\ t^3+1 & -3t/2 \end{pmatrix},$$

respectively. We shall pretend not to know the solution to formula (8.2) but we try to compute $x(t)$ for $0 \le t \le 10$. The results are shown in Figure 8.1 whose bottom left and right plots are for the absolute and relative errors:

$$|\boldsymbol{X}_i - x(\tau_i)|, \qquad \left|\frac{\boldsymbol{X}_i - x(\tau_i)}{x(\tau_i)}\right|$$

[5] This can be easily shown by noting that $\rho_2(\mu) = (1+\mu)/(1-\mu)$.

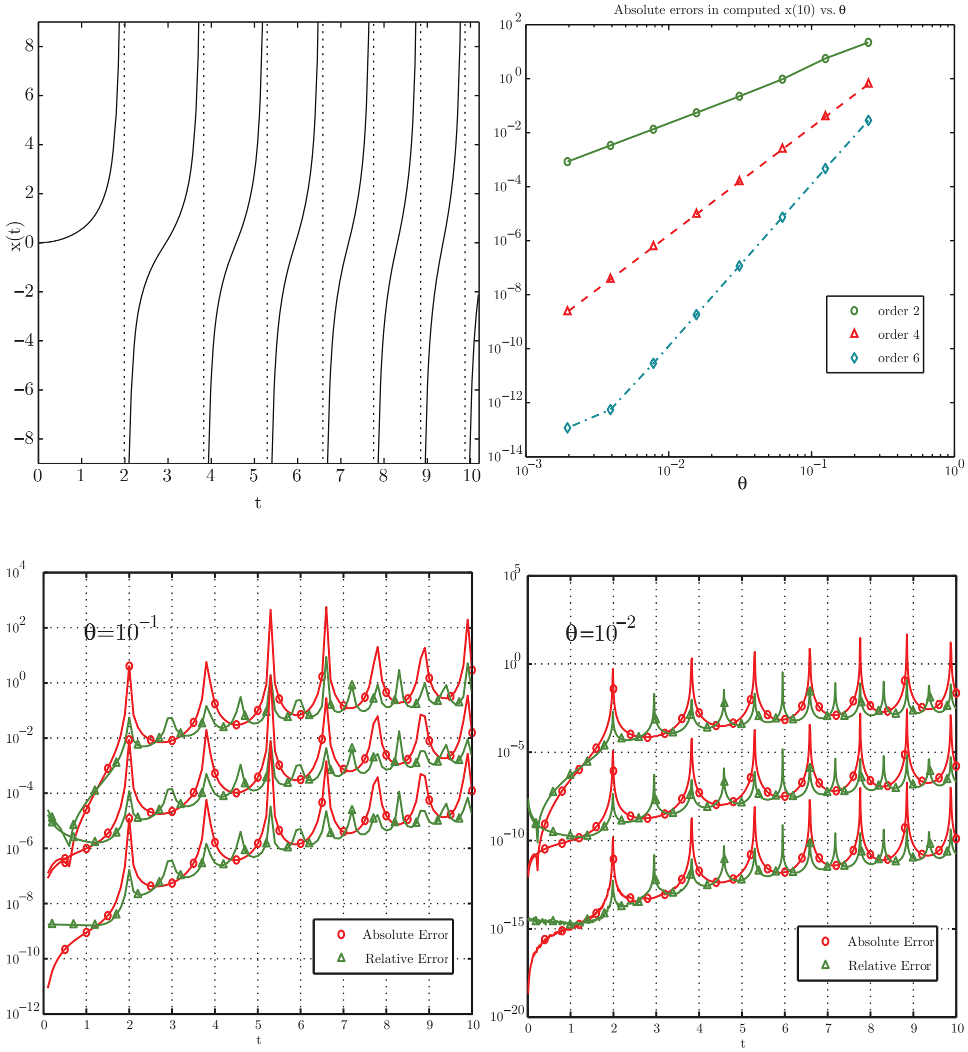


FIGURE 8.1. ***Top Left:*** the solution to Riccati Equation $x' = t + x^2$ with $x(0) = 0$. Note its many poles. ***Top Right:*** The absolute errors in the computed $x(10)$ as θ varies by odr2, odr4, and odr6. ***Bottom left and right:*** Absolute and relative errors of computed solutions. In each of the two, the top two error curves are for odr2, the middle two curves for odr4, and the bottom two curves for odr6.

at the integration points. The top right plot shows the absolute errors in the computed $x(10)$ as θ varies. Figure 8.1 clearly shows that our proposed numerical methods are able to compute $x(\tau_i)$ with accuracy dictated by the step-size θ and the order convergence, despite the singularity poles. It is not surprising that both errors get bigger near singularity poles. Also notice that the relative errors increase, too, near zeros of $x(t)$. $\diamond$

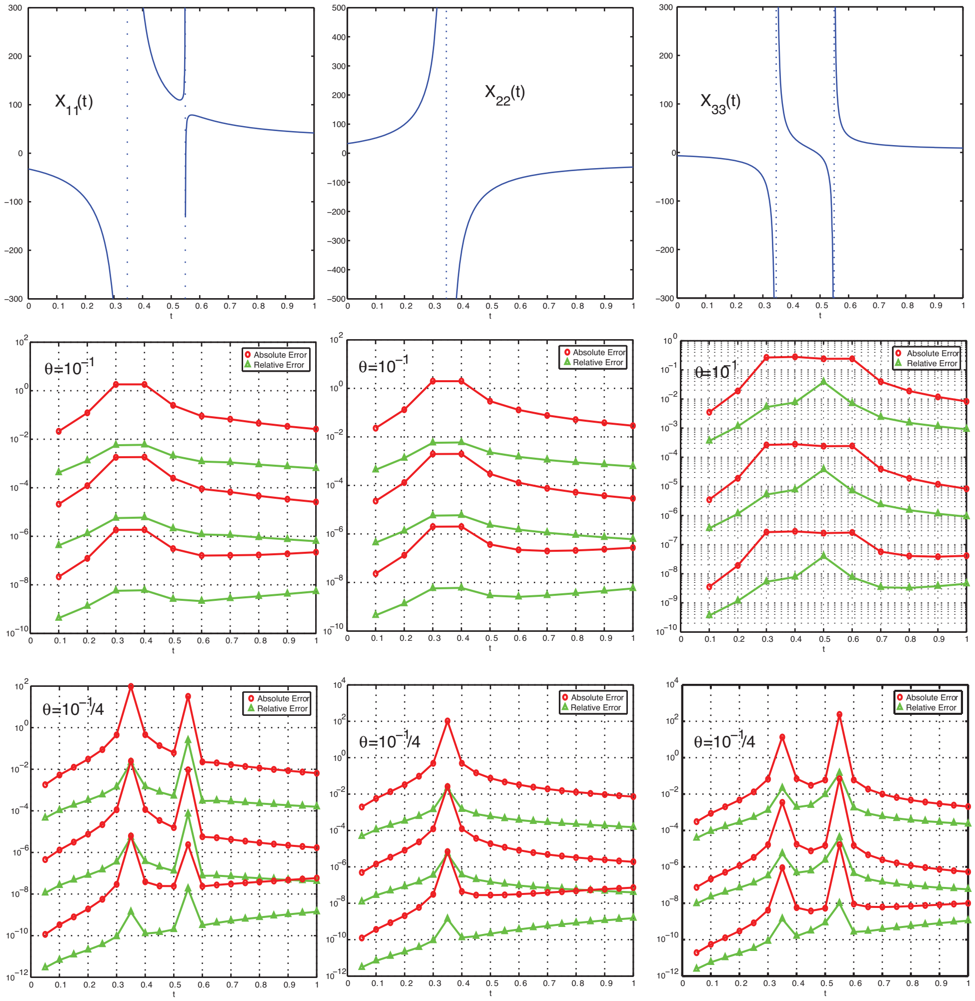


FIGURE 8.2. Diagonal entries of the solution to $X' = -X^2 + I$. The first row plots the diagonal entries of the exact solution, and the last two rows plot the absolute and relative errors of the computed diagonal entries with two different step sizes by odr2, odr4, and odr6. To see which error curve is for which, one simply sees that errors decrease with orders. It is interesting to note that the singularity at $(\ln 3)/2 = 0.54931$ in other entries did not spill over to the computed second diagonal entry.

Example 2. It is taken from [9], where a few time-invariant cases with known solutions are examined. The MRDE we tested is

$$X' = -X^2 + I, \quad X(0) = P\,\mathrm{diag}(-1,-2,-3)P^{-1}, \quad P = \begin{pmatrix} 4 & -5 & 9 \\ -8 & 18 & -17 \\ 4 & -37 & 9 \end{pmatrix}.$$

Notice that this MRDE is not really symmetric by definition because its initial value is not. The exact solution is known to be

$$X(t) = P \operatorname{diag}\left(\frac{\sinh t - \cosh t}{\cosh t - \sinh t}, \frac{\sinh t - 2\cosh t}{\cosh t - 2\sinh t}, \frac{\sinh t - 3\cosh t}{\cosh t - 3\sinh t}\right) P^{-1}, \tag{8.3}$$

and thus there are two poles $(\ln 2)/2 = 0.34657$ and $(\ln 3)/2 = 0.54931$ in the solution. Again we shall pretend not to know the solution and compute $X(t)$ for $0 \le t \le 1$. The first row of Figure 8.2 plots the diagonal entries of $X(t)$ (Note there is only one pole presented in $x_{22}(t)$, as can be verified from the analytic solution), while the second and third rows plot the absolute and relative errors of the corresponding computed diagonal entries by odr2, odr4, and odr6. Errors in other computed off-diagonal entries behave similarly. Once again the computed $X(\tau_i)$ have accuracy dictated by the step-size θ and the order of the method, despite the singularity poles. Note that the poles are more eminently displayed in the computed solution when the step-size θ is made smaller. This is expected because the smaller θ is, the closer some of integration points τ_i get to the poles. We also examined the absolute errors in the computed $X(1)$ as θ varies. A plot similar to the top right one in Figure 8.1 was obtained but omitted here to save space. This means that our methods display the claimed rate of convergence at $t = 1$, even after marching over two poles between $t = 0$ and $t = 1$. It is interesting to note that the singularity at $(\ln 3)/2 = 0.54931$ in other entries did not get spilled over to the computed second diagonal entry. $\diamond$

Example 3. This example originates from a stiff two-point boundary value problem in [6]. It was used as a test example in [15]. Here $m = n = 2$, $X(-1) = 0$, and

$$A = \begin{pmatrix} -t/(2\epsilon) & 0 & 1/\epsilon & 0 \\ 0 & 0 & 0 & 1/\epsilon \\ 1/2 & 1 & 0 & t/(2\epsilon) \\ 0 & 1 & 0 & 0 \end{pmatrix},$$

where $0 < \epsilon \ll 1$. According to Dieci [15] who was interested in $-1 \le t \le 1$, this MRDE's solution has an initial layer at $t = -1$ and then it approaches

$$X(t) \approx \begin{pmatrix} -\epsilon/t & 2\left(\sqrt{\epsilon} + t\right)/\left[2\left(1 - t\sqrt{\epsilon}\right)\right] \\ 0 & \sqrt{\epsilon} \end{pmatrix}$$

for t away from zero; then there is a smooth transition around the origin and then

$$X(t) \approx \begin{pmatrix} t/2 & \sqrt{\epsilon} \\ 0 & \sqrt{\epsilon} \end{pmatrix}$$

for $t > 0$. In [15], tests were done primarily with $\epsilon = 10^{-5}$ and for $-1 \le t \le 1$. We shall also take $\epsilon = 10^{-5}$. This problem is very stiff, and we use it to show suitability of the linear stability theory outlined in Section 7. Figure 8.3 plots computed $X(t)$ that can be repeated with smaller step-sizes to make sure its correctness, as well as the four eigenvalues along the solution trajectory: $\lambda = -\lambda_1 + \lambda_2$ with $\lambda_1 \in \operatorname{eig}(A_{11} + A_{12}X)$ and $\lambda_2 \in \operatorname{eig}(A_{22} - XA_{12})$. All eigenvalues are real and negative with $\min \lambda$ about $-10^5/2$. The linear stability regions shown in [36] suggest that our methods of orders 4, 8, 12, $\ldots$ would have to use very small step-sizes θ in order not to encounter any stability problem, while our methods of order 2, 6, 10, $\cdots$ do not have the same problem. But this is a time-varying MRDE for which the highest order of our methods given here is 6. Our findings are as follows.

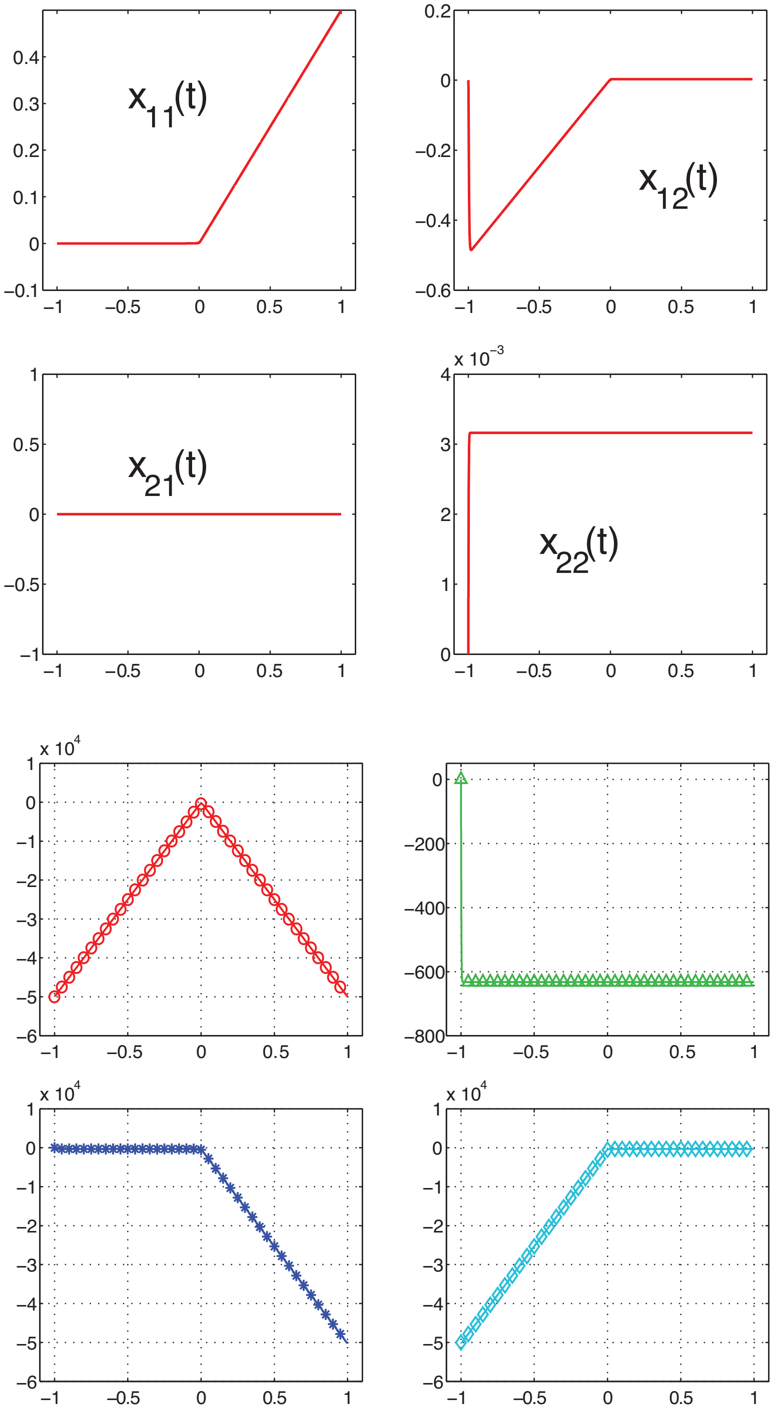


FIGURE 8.3. *Top four plots:* the solution to Example 3. *Bottom four plots:* Four eigenvalue trajectories $\lambda = -\lambda_1 + \lambda_2$ (all real and negative with $\min \lambda$ about $-10^5/2$), where $\lambda_1 \in \text{eig}(A_{11} + A_{12}X)$ and $\lambda_2 \in \text{eig}(A_{22} - XA_{12})$.

(1) We need to use $\theta = 10^{-2}/2$ or smaller for odr2 and odr6 to have reasonable numerical results, i.e., each entry of $X(t)$ behaves like what's being plotted in Figure 8.3.

(2) Our odr2 and odr6 encounter no stability problem with $\theta = 10^{-2}/2$ or smaller. For odr6, it is because its stability region $\mathscr{R}_6$ contains all $(-\infty, 0]$, even though it does not contain the entire left half-plane.

(3) Our odr4 produces erroneously numerical solutions for θ bigger than $10^{-4}/2$. This is very much consistent with what (7.5) suggests. Note the stability region $\mathscr{R}_4$ contains $[-\sqrt{3}, 0]$ but not $(-\infty, -\sqrt{3})$. So (7.5) suggests the step-size θ should satisfy

$$\frac{\theta}{2} \cdot \frac{10^5}{2} \le \sqrt{3} \quad \Rightarrow \quad \theta \le 6.9282 \times 10^{-5}.$$

When we use $\theta = 10^{-4}/2$ or smaller, odr4 does integrate the MRDE beautifully. ◇

9. Conclusions

We have derived a family of unconventional anadromic numerical methods for MRDEs. These methods preserve three important properties of MRDE, namely, *Bilinear Rational Relation*, *Generalized Inverse Property*, and *Symmetry*, as well as *conserving the Rank of Change to the initial value* and *Solution's Monotonicity* in the sense of Theorem 4.3. Among them, only *Symmetry* is widely preserved by pre-existing methods. The other two properties are mostly ignored thus far. Some of the advantages of our methods are:

(1) They have a distinctive capability that is able to march over solution singularities to render meaningful numerical results.

(2) They are anadromic, which implies they have even orders of convergence.

(3) They are semi-implicit, involving only linear systems of matrix equations to solve, and thus easily implementable.

Methods of any even order of convergence for time-invariant MRDEs, and of orders up to 10 for time-varying MRDEs are established. But the methods for time-varying MRDEs get complicated quickly as the order of convergence increases; so only methods up to 6 are stated in detail. Our methods can be cast into the framework of modified integrators in the sense of [8].

A linear stability theory is established to validate the suitability of our methods for stiff MRDE, especially those with real eigenvalues by which we mean all $\lambda = -\lambda_1 + \lambda_2$ are real, where $\lambda_1 \in \text{eig}(A_{11} + A_{12}X)$ and $\lambda_2 \in \text{eig}(A_{22} - XA_{12})$, and $X \equiv X(t)$ is the solution.

Our numerical tests are currently done within MATLAB and for constant step-size implementation. In a way, a robust variable step-size implementation would be more efficient. Since this paper is already lengthy, we shall investigate various varying step-size strategies elsewhere. Using the local truncation error formulas we have gotten in Theorem 3.3 for time-invariant MRDEs and running two methods of consecutive even orders are two natural strategies that we will be exploiting.

In explaining why our methods can march over solution poles and still correctly render meaningful numerical results, we made two crucial assumptions in Section 6. In Theorem 2.3, it is assumed that the solutions $X(t)$ to (MRDE) and $U(t)$ to (cMRDE) have only isolated singularities and share none in common. A better quantitative understanding of these assumptions is conceivably important for a better implementation of our methods.

Acknowledgment

The authors wish to thank Professor E. Hairer for his insightful comments [23], especially the connection between our methods and the implicit midpoint rule which led to a dramatic simplification to our proof of Theorem 3.3 in the earlier version of this article [35]. The connection also makes it possible for our *Mathematica* program to run more efficiently the time-varying MRDEs.

References

[1] M. Abramowitz and I. A. Stegun (editors), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, 9th printing ed., Dover Publications, Inc., New York, 1970.

[2] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK users' guide*, 3rd ed., SIAM, Philadelphia, 1999.

[3] U. M. Ascher, R. M. Mattheij, and R. D. Russell, *Numerical solution of boundary value problems for ordinary differential equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988. MR1000177 (90h:65120)

[4] I. Babuška and V. Majer, *The factorization method for the numerical solution of two point boundary value problems for linear ODE's*, SIAM J. Numer. Anal. **24** (1987), 1301–1334. MR917454 (88m:65114)

[5] David Bindel, *Private communication*, 2003.

[6] David L. Brown and Jens Lorenz, *A high-order method for stiff boundary value problems with turning points*, SIAM J. Sci. Statist. Comput. **8** (1987), no. 5, 790–805. MR902743 (90a:34030)

[7] R. Bulirsch and J. Stoer, *Numerical treatment of ordinary differential equations by extrapolation methods*, Numer. Math. **8** (1966), 1–13. MR0191095 (32:8504)

[8] Philippe Chartier, Ernst Hairer, and Gilles Vilmart, *Numerical integrators based on modified differential equations*, Math. Comp. **76** (2007), no. 260, 1941–1953. MR2336275 (2008g:65082)

[9] C. H. Choi and A. J. Laub, *Constructing Riccati differential equations with known analytic solutions for numerical experiments*, IEEE Trans. Automat. Control **35** (1990), 437–439. MR1047997

[10] ———, *Efficient matrix-valued algorithm for solving stiff Riccati differential equations*, IEEE Trans. Automat. Control **35** (1990), 770–776. MR1058361 (91f:93038)

[11] Chia-Chun Chou and Robert E. Wyatt, *Computational method for the quantum Hamilton-Jacobi equation: Bound states in one dimension*, J. Chem. Phys. **125** (2006), no. 17, 174103.

[12] E. J. Davison and M. C. Maki, *The numerical solution of the matrix differential Riccati equation*, IEEE Trans. Automat. Control **AC-18** (1973), 71–73.

[13] J. Demmel, *Applied numerical linear algebra*, SIAM, Philadelphia, PA, 1997. MR1463942 (98m:65001)

[14] P. Deuflhard, *Recent progress in extrapolation methods for ordinary differential equations*, SIAM Rev. **27** (1985), 505–535. MR812452 (86m:65075)

[15] L. Dieci, *Numerical integration of the differential Riccati equation and some related issues*, SIAM J. Numer. Anal. **29** (1992), no. 3, 781–815. MR1163357 (93b:65093)

[16] L. Dieci and T. Eirola, *Positive definiteness in the numerical solution of Riccati differential equations*, Numer. Math. **67** (1994), 303–313. MR1269499 (95b:65099)

[17] ———, *Preserving monotonicity in the numerical solution of Riccati differential equations*, Numer. Math. **74** (1996), 35–47. MR1400214 (97j:65114)

[18] L. Dieci, M. R. Osborne, and R. D. Russell, *A Riccati transformation method for solving linear BVPs. I: Theoretical aspects*, SIAM J. Numer. Anal. **25** (1988), no. 5, 1055–1073. MR960866 (90b:65153)

[19] ———, *A Riccati transformation method for solving linear BVPs. II: Computational aspects*, SIAM J. Numer. Anal. **25** (1988), no. 5, 1074–1092. MR960867 (90b:65154)

[20] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed., Johns Hopkins University Press, Baltimore, Maryland, 1996. MR1417720 (97g:65006)

[21] W. Gragg, *On extrapolation methods for ordinary initial-value problems*, SIAM J. Numer. Anal. **2** (1965), 384–403. MR0202318 (34:2191)

[22] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations I*, 2nd ed., Springer-Verlag, New York, 1992.

[23] Ernst Hairer, *Private communication*, December 10, 2009.

[24] Ernst Hairer, Christian Lubich, and Gerhard Wanner, *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, 2nd ed., Springer Series in Computational Mathematics, no. 31, Springer, Berlin, 2006. MR2221614 (2006m:65006)

[25] N. J. Higham, *FORTRAN codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation*, ACM Trans. Math. Software **14** (1988), 381–396. MR1062484

[26] ———, *Accuracy and stability of numerical algorithms*, SIAM, Philadephia, 1996. MR1368629 (97a:65047)

[27] W. Kahan and Ren-Cang Li, *Composition constants for raising the orders of unconventional schemes for ordinary differential equations*, Math. Comp. **66** (1997), 1089–1099. MR1423077 (97m:65120)

[28] ———, *Unconventional schemes for a class of ordinary differential equations–with applications to the Korteweg-de Vries*, J. Comput. Phys. **134** (1997), 316–331. MR1458831 (98b:65078)

[29] C. S. Kenney and R. B. Leipnik, *Numerical integration of the differential matrix Riccati equation*, IEEE Trans. Automat. Control **AC-30** (1985), no. 10, 962–970. MR804133 (87a:34018)

[30] H. Kwakernaak and R. Sivan, *Linear optimal control systems*, Wiley Interscience, New York, 1972. MR0406607 (53:10394)

[31] D. G. Lainiotis, *Generalized Chandrasekhar algorithms: Time-varying models*, IEEE Trans. Automat. Control **AC-21** (1976), 728–732. MR0421806 (54:9800)

[32] ———, *Partitioned Riccati solutions and integration-free doubling algorithms*, IEEE Trans. Automat. Control **AC-21** (1976), 677–689. MR0421805 (54:9799)

[33] A. J. Laub, *Schur techniques for Riccati differential equations*, Feedback Control of Linear and Nonlinear Systems (New York) (D. Hinrichsen and A. Isidori, eds.), Springer-Verlag, 1982. MR837458

[34] R. B. Leipnik, *A canonical form and solution for the matrix Riccati differential equation*, Bull. Australian Math. Soc. **26** (1985), 355–361. MR776321 (86f:34005)

[35] Ren-Cang Li and William Kahan, *A family of anadromic numerical methods for matrix riccati differential equations*, Tech. Report 2009-20, Department of Mathematics, University of Texas at Arlington, 2009, Available at `http://www.uta.edu/math/preprint/`.

[36] ———, *Modifying implicit midpoint rules for linear ordinary differential equation*, Tech. Report 2010-02, Department of Mathematics, University of Texas at Arlington, 2010, Available at `http://www.uta.edu/math/preprint/`.

[37] D. W. Rand and P. Winternitz, *Nonlinear superposition principles: A new numerical method for solving matrix Riccati equations*, Comput. Phys. Commun. **33** (1984), 305–328. MR770094 (86i:58118)

[38] W. T. Reid, *Monotoneity properties of solutions of Hermitian Riccati differential equations*, SIAM J. Math. Anal. **1** (1970), no. 2, 195–213. MR0262596 (41:7202)

[39] ———, *Riccati differential equations*, Academic Press, New York, 1972. MR0357936 (50:10401)

[40] I. Rusnak, *Almost analytic representation for the solution of the differential matrix Riccati equation*, IEEE Trans. Automat. Control **33** (1988), 191–193. MR922796

[41] Jeremy Schiff and S. Shnider, *A natural approach to the numerical integration of Riccati differential equations*, SIAM J. Numer. Anal. **36** (1999), no. 5, 1392–1413. MR1706774 (2000d:34024)

[42] M. Sorine and P. Winternitz, *Superposition laws for solutions of differential matrix Riccati equations arising in control theory*, IEEE Trans. Automat. Control **AC-30** (1985), 266–272. MR778430 (86e:34013)

[43] Lloyd N. Trefethen and David Bau, III, *Numerical linear algebra*, SIAM, Philadelphia, 1997. MR1444820 (98k:65002)

[44] D. R. Vaughan, *A negative exponential solution for the matrix Riccati equation*, IEEE Trans. Automat. Control **AC-14** (1969), 72–75. MR0250727 (40:3959)

[45] D. S. Watkins, *Fundamentals of matrix computations*, 2nd ed., John Wiley & Sons, New York, 2002. MR1899577 (2003a:65002)

[46] M. I. Zelikin, *Control theory and optimization i: Homogeneous spaces and the Riccati equation in calculus of variations*, Encyclopaedia of Mathematical Sciences, vol. 86, Springer, Berlin, 2000, Originally published in Russian in 1998 and translated into English by S. A. Vakhrameev. MR1739679 (2001a:49002)

Department of Mathematics, University of Texas at Arlington, P.O. Box 19408, Arlington, Texas 76019-0408

E-mail address: rcli@uta.edu

Department of Mathematics, and of Electric Engineering & Computer Science, University of California at Berkeley, Berkeley, California 94720

E-mail address: wkahan@eecs.berkeley.edu