J Math Chem (2011) 49:1026–1041
DOI 10.1007/s10910-010-9794-z

ORIGINAL PAPER

On some discretization methods for solving a linear matrix ordinary differential equation

Hao Zheng · Weimin Han

Received: 4 August 2010 / Accepted: 10 December 2010 / Published online: 13 February 2011

Abstract In this paper, some discretization methods are considered for solving a linear matrix ordinary differential equation. Discussion is focused on a family of one step methods which include Euler, backward Euler, and Crank–Nicolson schemes as special cases, as well as the Runge–Kutta methods. As an illustration, detailed convergence and error analysis are given for the family of one step methods. Some numerical examples are provided to show the good performance of the methods.

Keywords Ordinary differential equations · Discretization methods · Convergence · Error estimates

1 Introduction

Many of the general laws of nature in chemistry, other physical sciences and engineering are expressed by differential equations (cf. e.g., [3,8,11,12]). In two recent papers [1,2], Altınbaşak and Demiralp discuss solutions to the following initial value problem of a linear matrix ordinary differential equation

$$\mathbf{X}'(t) = \mathbf{A}(t)\,\mathbf{X}(t), \tag{1}$$

$$\mathbf{X}(0) = \mathbf{I}. \tag{2}$$

H. Zheng (✉)
Department of Chemistry, Zhejiang University, 310027 Hangzhou, China
e-mail: zhenghao@zju.edu.cn

W. Han
Department of Mathematics and Program in Applied Mathematical and Computational Sciences, University of Iowa, Iowa City, IA 52242, USA
e-mail: whan@math.uiowa.edu

Here $\mathbf{X}(t)$ and $\mathbf{A}(t)$ are $m \times m$ matrix-valued functions of a real variable t, $\mathbf{A}(t)$ is given, $\mathbf{X}(t)$ is the unknown, and $\mathbf{I} := \mathbf{I}_m$ is the $m \times m$ identity (unit) matrix. In [1,2], the given matrix function $\mathbf{A}(t)$ is assumed to be a polynomial of t. A series expansion solution and corresponding truncation approximation around the initial time $t = 0$ are the topic of [2]. As a sequel to [2], a perturbation technique is used in [1] that allows the construction of a perturbation expansion solution around values of the independent variable t other than $t = 0$. The purpose of this paper is to study some numerical methods through discretization to solve the initial value problem (1)–(2). As will be evident, numerical methods provide an efficient, general way of solving the initial value problem.

Unlike [1,2], we do not assume the matrix function $\mathbf{A}(t)$ to be a polynomial of t. The numerical methods can be applied to solve the problem (1)–(2) with a general continuous coefficient matrix $\mathbf{A}(t)$. Throughout the paper, we assume $\mathbf{A}(t)$ is a continuous function of the variable t. Of course, to have optimal convergence order for the numerical solutions and to prove rigorously such optimal convergence order, we need smoothness assumptions on the coefficient matrix $\mathbf{A}(t)$ which in turn implies smoothness of the exact solution (cf. Proposition 1 below for this implication). We will consider a family of one step methods as well as the Runge–Kutta methods for the initial value problem. As an illustration, we will provide a rigorous convergence and error analysis for the one step methods. We will also give examples of numerical results and discuss the performance of some representative methods. For definiteness, we solve the differential equation on an interval $[0, T]$ where $T < \infty$ is arbitrary but fixed. Therefore, the problem we discuss is

$$\mathbf{X}'(t) = \mathbf{A}(t)\,\mathbf{X}(t), \quad t \in [0, T], \tag{3}$$

$$\mathbf{X}(0) = \mathbf{I}. \tag{4}$$

By the standard theory on ordinary differential equations (e.g., [13] or any advanced text on ordinary differential equations), if $\mathbf{A}(t)$ is a continuous function of t, then the problem (3)–(4) admits a unique solution $\mathbf{X}(t)$, which is continuously differentiable.

As is typical in doing error analysis of a numerical method for solving a differential equation (e.g., [4–7,10]), we need certain smoothness condition on the exact solution of the problem (3)–(4). Differentiating the equation (3) with respect to t and using the equation (3), we have

$$\begin{aligned}
\mathbf{X}''(t) &= \mathbf{A}'(t)\,\mathbf{X}(t) + \mathbf{A}(t)\,\mathbf{X}'(t) = \left[\mathbf{A}'(t) + \mathbf{A}(t)^2\right]\mathbf{X}(t),\\
\mathbf{X}^{(3)}(t) &= \left[\mathbf{A}''(t) + \mathbf{A}'(t)\,\mathbf{A}(t) + \mathbf{A}(t)\,\mathbf{A}'(t)\right]\mathbf{X}(t) + \left[\mathbf{A}'(t) + \mathbf{A}(t)^2\right]\mathbf{X}'(t)\\
&= \left[\mathbf{A}''(t) + \mathbf{A}(t)\,\mathbf{A}'(t) + 2\,\mathbf{A}'(t)\,\mathbf{A}(t) + \mathbf{A}(t)^3\right]\mathbf{X}(t).
\end{aligned}$$

Higher order derivatives of the solution $\mathbf{X}(t)$ can be treated inductively. Thus, the following result holds.

Proposition 1 *For the problem (3)–(4), if* $\mathbf{A} \in C^1([0,T])^{m\times m}$, *then the solution* $\mathbf{X} \in C^2([0,T])^{m\times m}$ *and for some constant c depending on* $\|\mathbf{A}\|_{C^1([0,T])^{m\times m}}$,

$$\|\mathbf{X}\|_{C^2([0,T])^{m\times m}} \le c; \tag{5}$$

moreover, if $\mathbf{A} \in C^2([0,T])^{m\times m}$, *then* $\mathbf{X} \in C^3([0,T])^{m\times m}$ *and for some constant c depending on* $\|\mathbf{A}\|_{C^2([0,T])^{m\times m}}$,

$$\|\mathbf{X}\|_{C^3([0,T])^{m\times m}} \le c. \tag{6}$$

In general, for a non-negative integer k, if $\mathbf{A} \in C^k([0,T])^{m\times m}$, *then* $\mathbf{X} \in C^{k+1}([0,T])^{m\times m}$ *and for some constant c depending on* $\|\mathbf{A}\|_{C^k([0,T])^{m\times m}}$,

$$\|\mathbf{X}\|_{C^{k+1}([0,T])^{m\times m}} \le c.$$

In Proposition 1 and later in the paper, for a k-times continuously differentiable matrix-valued function $\mathbf{B}(t)$, $t \in [0,T]$, the norm $\|\mathbf{B}\|_{C^k([0,T])^{m\times m}}$ is defined as follows:

$$\|\mathbf{B}\|_{C^k([0,T])^{m\times m}} := \max_{0\le j\le k} \max_{0\le t\le T} \|\mathbf{B}^{(j)}(t)\|,$$

where $\|\cdot\|$ is any operator matrix norm [4] such as the 1-norm, 2-norm, or maximum norm, or the Frobenius norm. Since any two norms are equivalent over a finite dimensional space (e.g., the space of all $m \times m$ matrices), any matrix norm can be used.

The rest of the paper is organized as follows. In Sect. 2, some numerical methods are introduced for solving the problem (3)–(4), including a family of one step methods and the Runge–Kutta methods. The family of one step methods include such methods as Euler, backward Euler, and Crank–Nicolson schemes as special cases. In Sect. 3, rigorous error bounds for the one step methods are derived. In Sect. 4, numerical results from some examples are presented and discussion is given on performance of some representative methods. The final section, Sect. 5, contains several remarks.

2 Some numerical methods

For the matrix ordinary differential equation, it is possible to extend the discretization methods for solving scalar ordinary differential equations covered in textbooks and monographs on the topic, e.g. [6,7,10].

For a positive integer N, let $h = T/N$ be the step-size and denote the node points

$$t_n = n\,h, \quad 0 \le n \le N.$$

We will use the short-hand notation $\mathbf{A}_n := \mathbf{A}(t_n)$, $\mathbf{X}_n := \mathbf{X}(t_n)$ for $n = 0, 1, \ldots, N$, and use $\mathbf{Y}_n$ to denote a numerical approximation of $\mathbf{X}_n$.

First, we consider a family of one step methods. For a parameter θ, in the range $\theta \in [0, 1]$, we introduce the following one step numerical method for the problem (3)–(4):

$$\mathbf{Y}_{n+1} = \mathbf{Y}_n + h\left[\theta\,\mathbf{A}_{n+1}\mathbf{Y}_{n+1} + (1-\theta)\,\mathbf{A}_n\mathbf{Y}_n\right], \quad 0 \le n \le N-1, \tag{7}$$

$$\mathbf{Y}_0 = \mathbf{I}. \tag{8}$$

When $\theta = 0$, the scheme (7) reduces to

$$\mathbf{Y}_{n+1} = (\mathbf{I} + h\,\mathbf{A}_n)\,\mathbf{Y}_n, \quad 0 \le n \le N-1, \tag{9}$$

and we recover the Euler method. The Euler method is an explicit method, in the sense that once an approximate solution $\mathbf{Y}_n$ at t_n is known, we can compute the approximate solution $\mathbf{Y}_{n+1}$ at the next node point t_{n+1} by the formula (9) through addition and multiplication.

For $\theta \neq 0$, we deduce from (7) that

$$(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})\,\mathbf{Y}_{n+1} = [\mathbf{I} + (1-\theta)\,h\,\mathbf{A}_n]\,\mathbf{Y}_n. \tag{10}$$

Thus, to compute $\mathbf{Y}_{n+1}$ from $\mathbf{Y}_n$, one needs to solve a linear system with the coefficient matrix $(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})$. The method is thus an implicit method. Note that if the step-size h is small enough, e.g., if

$$|\theta|\,h\,\|\mathbf{A}\|_{C([0,T])^{m\times m}} < 1,$$

then the linear system (10) is uniquely solvable for $\mathbf{Y}_{n+1}$. Actually, the linear system (10) is uniquely solvable as long as $\det(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1}) \neq 0$, or in other words, as long as h^{-1} is not an eigenvalue of the matrix $\theta\,\mathbf{A}_{n+1}$. For a general coefficient matrix $\mathbf{A}(t)$ arising in applications, it is unlikely for h^{-1} to happen to be an eigenvalue of the matrix $\theta\,\mathbf{A}_{n+1}$. In the particular case where $\theta = 1$, we get a formula for the backward Euler method

$$(\mathbf{I} - h\,\mathbf{A}_{n+1})\,\mathbf{Y}_{n+1} = \mathbf{Y}_n, \tag{11}$$

whereas if $\theta = 1/2$, a formula for the Crank–Nicolson scheme is obtained,

$$\left(\mathbf{I} - \frac{h}{2}\,\mathbf{A}_{n+1}\right)\mathbf{Y}_{n+1} = \left(\mathbf{I} + \frac{h}{2}\,\mathbf{A}_n\right)\mathbf{Y}_n. \tag{12}$$

Next, we consider using the Runge–Kutta methods (cf. e.g., [6,7,10]). For the initial value problem of a first order scalar differential equation

$$y' = f(t, y),$$

with a numerical solution y_n at t_n known, an s-stage Runge–Kutta method to compute the numerical solution y_{n+1} at t_{n+1} has the following general form:

$$z_{n,i} = y_n + h\sum_{j=1}^{s} a_{i,j} f(t_n + c_j h,\ z_{n,j}), \quad i = 1, \ldots, s,$$
$$y_{n+1} = y_n + h\sum_{j=1}^{s} b_j f(t_n + c_j h,\ z_{n,j}).$$

Here, the coefficients $a_{i,j}$, b_j and c_j, $1 \le i, j \le s$, are selected so that the method has a prescribed convergence order and satisfies certain stability property. The method is usually represented by the Butcher tableau

$$\begin{array}{c|cccc} c_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,s-1} & a_{1,s} \\ c_2 & a_{2,1} & a_{2,1} & \cdots & a_{2,s-1} & a_{2,s} \\ c_3 & a_{3,1} & a_{3,2} & \cdots & a_{3,s-1} & a_{3,s} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_s & a_{s,1} & a_{s,2} & \cdots & a_{s,s-1} & a_{s,s} \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array}$$

Applied to the problem (3)–(4), the Runge–Kutta method starts with

$$\mathbf{Y}_0 = \mathbf{I}.$$

For $n = 0, 1, \ldots,$ with $\mathbf{Y}_n$ known, we first determine s matrices of order $m \times m$, $\{\mathbf{Z}_{n,i}\}_{1 \le i \le s}$, from the system

$$\mathbf{Z}_{n,i} = \mathbf{Y}_n + h\sum_{j=1}^{s} a_{i,j}\mathbf{A}(t_n + c_j h)\,\mathbf{Z}_{n,j}, \quad i = 1, \ldots, s,$$

and then compute the approximate solution $\mathbf{Y}_{n+1}$ at t_{n+1} by the formula

$$\mathbf{Y}_{n+1} = \mathbf{Y}_n + h\sum_{j=1}^{s} b_j\mathbf{A}(t_n + c_j h)\,\mathbf{Z}_{n,j}.$$

In Sect. 4, we will report numerical results based on a popular 4 stage 4th order Runge–Kutta method with the Butcher-tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

For this method, to compute $\mathbf{Y}_{n+1}$ from $\mathbf{Y}_n$, we do the following, with $\mathbf{A}_{n+1/2} \equiv \mathbf{A}(t_n + h/2)$,

$$\mathbf{B}_{n,1} = \mathbf{A}_n \mathbf{Y}_n, \quad \mathbf{Z}_{n,1} = \mathbf{Y}_n + \frac{1}{2} h \, \mathbf{B}_{n,1}, \tag{13}$$

$$\mathbf{B}_{n,2} = \mathbf{A}_{n+1/2} \mathbf{Z}_{n,1}, \quad \mathbf{Z}_{n,2} = \mathbf{Y}_n + \frac{1}{2} h \, \mathbf{B}_{n,2}, \tag{14}$$

$$\mathbf{B}_{n,3} = \mathbf{A}_{n+1/2} \mathbf{Z}_{n,2}, \quad \mathbf{Z}_{n,3} = \mathbf{Y}_n + h \, \mathbf{B}_{n,3}, \tag{15}$$

$$\mathbf{Y}_{n+1} = \mathbf{Y}_n + \frac{h}{6} \left(\mathbf{B}_{n,1} + 2 \, \mathbf{B}_{n,2} + 2 \, \mathbf{B}_{n,3} + \mathbf{A}_{n+1} \mathbf{Z}_{n,3} \right). \tag{16}$$

3 Convergence and error bounds

As an example for a mathematical analysis of a numerical method for solving the linear matrix ordinary differential equation problem (3)–(4), in this section, we provide a rigorous convergence and error analysis for the method (7)–(8). For this purpose, introduce the truncation error

$$\boldsymbol{\tau}_n(\mathbf{X}) := \frac{\mathbf{X}_{n+1} - \mathbf{X}_n}{h} - \left[\theta \, \mathbf{A}_{n+1} \mathbf{X}_{n+1} + (1 - \theta) \, \mathbf{A}_n \mathbf{X}_n \right], \quad 0 \le n \le N - 1. \tag{17}$$

We distinguish two cases, according to whether $\theta = 1/2$. Recall that $\| \cdot \|$ is any operator matrix norm or the Frobenius norm.

We will need the following result (cf. [4, Theorem 7.11] or [5, Theorem 2.3.1] in a general abstract setting).

Lemma 2 *If* $\|\mathbf{B}\| < 1$, *then* $(\mathbf{I} - \mathbf{B})^{-1}$ *exists and*

$$\|(\mathbf{I} - \mathbf{B})^{-1}\| \le \frac{1}{1 - \|\mathbf{B}\|}.$$

First consider the case of a general value $\theta \in [0, 1]$. We assume $\mathbf{A} \in C^1([0, T])^{m \times m}$. Then according to Proposition 1, the solution $\mathbf{X} \in C^2([0, T])^{m \times m}$ and we have the bound (5). By the Taylor expansion, we have

$$\max_{0 \le n \le N-1} \|\boldsymbol{\tau}_n(\mathbf{X})\| \le c_1 h \, \|\mathbf{X}\|_{C^2([0,T])^{m \times m}} \le c_2 h \tag{18}$$

for some constant c_2 depending on $\|\mathbf{A}\|_{C^1([0,T])^{m\times m}}$. From the definition (17), we have

$$\mathbf{X}_{n+1} = \mathbf{X}_n + h\left[\theta\,\mathbf{A}_{n+1}\mathbf{X}_{n+1} + (1-\theta)\,\mathbf{A}_n\mathbf{X}_n\right] + h\,\boldsymbol{\tau}_n(\mathbf{X}), \quad 0 \le n \le N-1. \tag{19}$$

Denote the approximation errors

$$\mathbf{E}_n := \mathbf{X}_n - \mathbf{Y}_n, \quad 0 \le n \le N. \tag{20}$$

Then,

$$\mathbf{E}_0 = \mathbf{O} \tag{21}$$

is the zero matrix of order $m \times m$. Subtract (7) from (19) to obtain

$$\mathbf{E}_{n+1} = \mathbf{E}_n + h\left[\theta\,\mathbf{A}_{n+1}\mathbf{E}_{n+1} + (1-\theta)\,\mathbf{A}_n\mathbf{E}_n\right] + h\,\boldsymbol{\tau}_n(\mathbf{X}), \quad 0 \le n \le N-1,$$

i.e.,

$$(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})\,\mathbf{E}_{n+1} = [\mathbf{I} + (1-\theta)\,h\,\mathbf{A}_n]\,\mathbf{E}_n + h\,\boldsymbol{\tau}_n(\mathbf{X}), \quad 0 \le n \le N-1. \tag{22}$$

Denote

$$a_0 := \|\mathbf{A}\|_{C([0,T])^{m\times m}}.$$

Then by Lemma 2, if $\theta\,h < 1/a_0$ (this condition is valid for any step-size h for the Euler method which corresponds to the choice $\theta = 0$), $(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})$ is invertible and

$$\left\|(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})^{-1}\right\| \le \frac{1}{1-\theta\,a_0 h}.$$

We obtain from (22) that

$$\mathbf{E}_{n+1} = (\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})^{-1}\,[\mathbf{I} + (1-\theta)\,h\,\mathbf{A}_n]\,\mathbf{E}_n + h\,(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})^{-1}\,\boldsymbol{\tau}_n(\mathbf{X}).$$

Apply the matrix norm to the above equality,

$$\begin{aligned}\|\mathbf{E}_{n+1}\| &\le \left\|(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})^{-1}\right\|\,\|[\mathbf{I} + (1-\theta)\,h\,\mathbf{A}_n]\|\,\|\mathbf{E}_n\| \\ &\quad + h\left\|(\mathbf{I} - \theta\,h\,\mathbf{A}_{n+1})^{-1}\right\|\,\|\boldsymbol{\tau}_n(\mathbf{X})\| \\ &\le \frac{1 + (1-\theta)\,a_0 h}{1-\theta\,a_0 h}\,\|\mathbf{E}_n\| + \frac{c_2 h^2}{1-\theta\,a_0 h},\end{aligned} \tag{23}$$

where we have used the bound (18). Apply (23) recursively to obtain

$$\|\mathbf{E}_n\| \le \left[\frac{1+(1-\theta)\,a_0h}{1-\theta\,a_0h}\right]^n \|\mathbf{E}_0\| + \frac{c_2h^2}{1-\theta\,a_0h}\sum_{j=0}^{n-1}\left[\frac{1+(1-\theta)\,a_0h}{1-\theta\,a_0h}\right]^j. \quad (24)$$

By (21), $\|\mathbf{E}_0\| = 0$. Also recall the formula

$$\sum_{j=0}^{n-1} r^j = \frac{r^n-1}{r-1}, \quad r \ne 1.$$

We deduce from (24) that

$$\|\mathbf{E}_n\| \le \frac{c_2h}{a_0}\left\{\left[\frac{1+(1-\theta)\,a_0h}{1-\theta\,a_0h}\right]^n - 1\right\}. \quad (25)$$

Now recall the inequalities

$$1+t \le e^t \quad \text{for any real } t,$$
$$\frac{1}{1-t} \le 1+2t \quad \text{for any } t \in [0,1/2].$$

Then from (25), we have, for $\theta\,a_0h \le 1/2$,

$$\|\mathbf{E}_n\| \le \frac{c_2h}{a_0}\left[e^{(1+\theta)\,a_0n\,h} - 1\right] \le \frac{c_2h}{a_0}\left[e^{(1+\theta)\,a_0T} - 1\right].$$

So for a constant c_3 depending on $\|\mathbf{A}\|_{C^1([0,T])^{m\times m}}$, θ and T, we have the error bound

$$\max_{0\le n\le N}\|\mathbf{E}_n\| \le c_3h. \quad (26)$$

Thus, the method is first order convergent.

Now consider the scheme with $\theta = 1/2$. We assume $\mathbf{A} \in C^2([0,T])^{m\times m}$. Then by Proposition 1, the solution $\mathbf{X} \in C^3([0,T])^{m\times m}$ and we have the bound (6). By Taylor expansion, the truncation error

$$\boldsymbol{\tau}_n(\mathbf{X}) := \frac{\mathbf{X}_{n+1}-\mathbf{X}_n}{h} - \frac{1}{2}\left(\mathbf{A}_{n+1}\mathbf{X}_{n+1} + \mathbf{A}_n\mathbf{X}_n\right), \quad 0 \le n \le N-1$$

is bounded as follows:

$$\max_{0\le n\le N-1}\|\boldsymbol{\tau}_n(\mathbf{X})\| \le c_4h^2\|\mathbf{X}\|_{C^3([0,T])^{m\times m}} \le c_5h^2 \quad (27)$$

for some constant c_5 depending on $\|\mathbf{A}\|_{C^2([0,T])^{m\times m}}$. We still have (22) with $\theta = 1/2$:

$$\left(\mathbf{I} - \frac{h}{2}\mathbf{A}_{n+1}\right)\mathbf{E}_{n+1} = \left(\mathbf{I} + \frac{h}{2}\mathbf{A}_n\right)\mathbf{E}_n + h\,\boldsymbol{\tau}_n(\mathbf{X}), \quad 0 \le n \le N-1.$$

Assume $h \le 1/a_0$. Similar to (23), we have

$$\|\mathbf{E}_{n+1}\| \le \frac{2 + a_0 h}{2 - a_0 h}\|\mathbf{E}_n\| + \frac{2\,c_5 h^3}{2 - a_0 h}.$$

The analogue of (25) is

$$\|\mathbf{E}_n\| \le \frac{c_5 h^2}{a_0}\left[\left(\frac{2 + a_0 h}{2 - a_0 h}\right)^n - 1\right],$$

and then

$$\|\mathbf{E}_n\| \le \frac{c_5 h^2}{a_0}\left(e^{3\,a_0 T/2} - 1\right).$$

So for a constant c_6 depending on $\|\mathbf{A}\|_{C^2([0,T])^{m\times m}}$ and T, we have the error bound

$$\max_{0\le n\le N}\|\mathbf{E}_n\| \le c_6 h^2. \tag{28}$$

Thus, the method is second order convergent.

We summarize the results in the form of a theorem.

Theorem 3 *For h small enough, the family of numerical methods (7)–(8) is well defined. If $\mathbf{A} \in C^1([0,T])^{m\times m}$, then the methods are first order accurate and the error bound (26) holds. For $\theta = 1/2$, assume further that $\mathbf{A} \in C^2([0,T])^{m\times m}$. Then the method is second order accurate and the error bound (28) holds.*

4 Numerical examples

In this section, we present numerical results on the two examples given in [1,2]. The results we report are based on the second order Crank–Nicolson scheme (8) and (12), and on the fourth order Runge–Kutta method (8) and (13)–(16). As in [1,2], we compare the Frobenius norms of the exact solution and the numerical solutions. We also show numerical results on the relative error $\|\mathbf{X}_n - \mathbf{Y}_n\|/\|\mathbf{X}_n\|$ in the Frobenius norm. We use MATLAB for the computation in these examples.

Example 4 In the first example, the coefficient matrix is

$$\mathbf{A}(t) = \begin{pmatrix} -\frac{19}{2}t - 12 & -14\,t - \frac{35}{2} \\ \frac{20}{3}t + \frac{25}{3} & \frac{59}{6}t + \frac{73}{6} \end{pmatrix}.$$

Table 1 Relative errors of Crank–Nicolson solutions for Example 4

h	$t = 1$	Ratio	$t = 5$	Ratio	$t = 10$	Ratio
1/10	1.882×10^{-3}		4.614×10^{-2}		5.408×10^{-1}	
1/20	4.697×10^{-4}	4.01	1.126×10^{-2}	4.10	1.113×10^{-1}	4.86
1/40	1.174×10^{-4}	4.00	2.797×10^{-3}	4.02	2.657×10^{-2}	4.19
1/80	2.933×10^{-5}	4.00	6.982×10^{-4}	4.01	6.568×10^{-3}	4.05
1/160	7.333×10^{-6}	4.00	1.745×10^{-4}	4.00	1.637×10^{-3}	4.01

The solution of the initial value problem (1)–(2) is given by the formula

$$\mathbf{X}(t) = \begin{pmatrix} 15\,x_1(t) - 14\,x_2(t) & 21\,x_1(t) - 21\,x_2(t) \\ 10\,x_2(t) - 10\,x_1(t) & 15\,x_2(t) - 14\,x_1(t) \end{pmatrix},$$

where

$$x_1(t) = e^{-t^2/12 - t/3}, \quad x_2(t) = e^{t^2/4 + t/2}.$$

Numerical results from the Crank–Nicolson scheme are reported in Figs. 1 and 2. In Fig. 1, we show the values of the Frobenius norm of the exact solution and of the numerical solutions corresponding to stepsizes $h = 1/10$, 1/20, and 1/40. Since the exact solution grows exponentially as t increases, even for a moderate size of t, values of the Frobenius norm for either the exact solution or the numerical solutions are of very large size. So Fig. 1 only shows the results on the time interval [0,3]; if the results are shown on a larger time interval, due to the exponentially fast growth of the exact solution, all the curves for most part of the time interval will be visually nearly horizontal, coinciding with the x-axis of the figure (see Fig. 3 for such a phenomenon). To have some idea on the fast growth of the solution, we note that

$$\|\mathbf{X}(1)\| \doteq 4.524 \times 10^1, \quad \|\mathbf{X}(3)\| \doteq 1.313 \times 10^3,$$
$$\|\mathbf{X}(5)\| \doteq 1.957 \times 10^5, \quad \|\mathbf{X}(10)\| \doteq 3.315 \times 10^{14}.$$

Figure 2 shows the relative errors of the numerical solutions. Table 1 provides relative numerical solution errors at three representative times $t = 1$, 5, and 10. For a given meshsize h, the corresponding component in the column "ratio" provides the ratio of the numerical solution error for $2h$ with that for h. Second order convergence is evident from the table: When h is small and is halved, the error is reduced by a factor approximately 4.

We then apply the Runge–Kutta method (8) and (13)–(16). The method is of 4th order and provides substantially more accurate numerical solutions than the Crank–Nicolson scheme. Some numerical values are reported in Table 2. Fourth order convergence is evident from Table 2: When h is small and is halved, the error is reduced by a factor approximately 16.

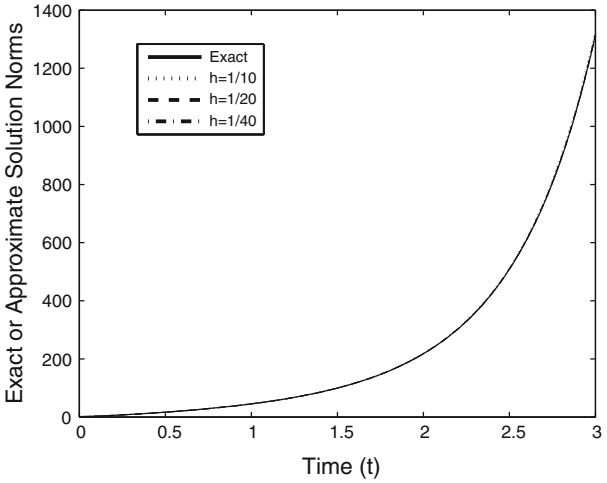

Fig. 1 Crank–Nicolson scheme for Example 4: comparison of the Frobenius norms of the exact solution and numerical solutions for several step-sizes

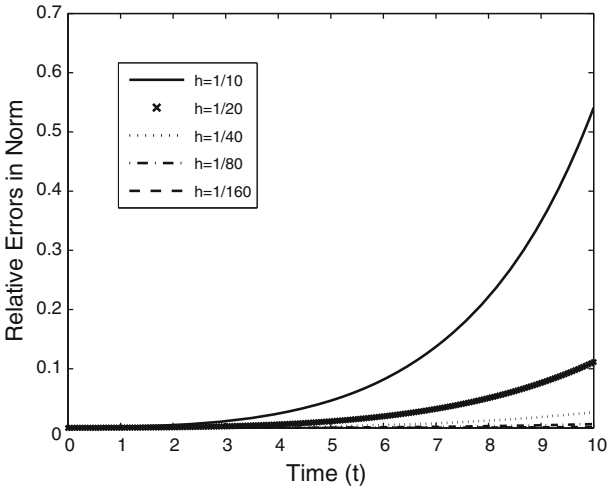

Fig. 2 Crank–Nicolson scheme for Example 4: relative errors of numerical solutions in the Frobenius norm

Table 2 Relative errors of Runge–Kutta solutions for Example 4

h	$t = 1$	Ratio	$t = 5$	Ratio	$t = 10$	Ratio
1/10	3.939×10^{-7}		1.639×10^{-4}		5.200×10^{-3}	
1/20	2.490×10^{-8}	15.8	1.139×10^{-5}	14.4	3.952×10^{-4}	13.2
1/40	1.564×10^{-9}	15.9	7.502×10^{-7}	15.2	2.723×10^{-5}	14.5
1/80	9.798×10^{-11}	16.0	4.815×10^{-8}	15.6	1.787×10^{-6}	15.2
1/160	6.126×10^{-12}	16.0	3.049×10^{-9}	15.8	1.145×10^{-7}	15.6

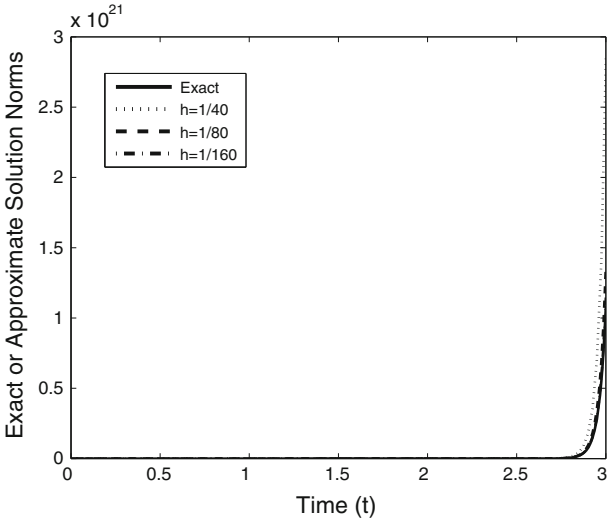

Fig. 3 Crank–Nicolson scheme for Example 5: comparison of the Frobenius norms of the exact solution and numerical solutions for several step-sizes

We note that at each step, the main costs of the Runge–Kutta method are the two matrix-valued function evaluations, $\mathbf{A}_{n+1/2}$ and $\mathbf{A}_{n+1}$ ($\mathbf{A}_n$ is available from the previous step), and four matrix multiplications. In comparison, for the Crank–Nicolson scheme, the main costs at each step are one matrix-valued function evaluation, $\mathbf{A}_{n+1}$, one matrix multiplication for the right side of (12), and one linear system solving. So at each step, the Runge–Kutta method is about twice as expensive as the Crank–Nicolson scheme. Since the numerical solutions from the Runge–Kutta method are substantially more accurate than that from the Crank–Nicolson scheme for a same value of the step-size, the Runge–Kutta method is a better choice for this example.

Example 5 The second example is constructed in such a way that it is more difficult to apply analytic approximation methods [1,2]. The increased degree of difficulty is due to the even faster exponential growth of the exact solution. In this example, the coefficient matrix is

$$\mathbf{A}(t) = \begin{pmatrix} -202\,t - 117 & -294\,t - 168 \\ 140\,t + 80 & 204\,t + 115 \end{pmatrix}.$$

The solution of the initial value problem (1)–(2) is given by the formula

$$\mathbf{X}(t) = \begin{pmatrix} 15\,x_1(t) - 14\,x_2(t) & 21\,x_1(t) - 21\,x_2(t) \\ 10\,x_2(t) - 10\,x_1(t) & 15\,x_2(t) - 14\,x_1(t) \end{pmatrix},$$

where

$$x_1(t) = e^{-3t^2-5t}, \quad x_2(t) = e^{4t^2+3t}.$$

Table 3 Relative errors of Crank–Nicolson solutions for Example 5, part I

h	$t = 1$	$t = 5$	$t = 10$
1/10	8.331×10^{-1}	1.000	1.000
1/20	1.439×10^{-1}	1.337×10^{24}	1.000
1/40	3.327×10^{-2}	6.636×10^{2}	–
1/80	8.162×10^{-3}	3.350	3.268×10^{9}
1/160	2.031×10^{-3}	4.324×10^{-1}	1.470×10^{2}

Table 4 Relative errors of Crank–Nicolson solutions for Example 5, part II

h	$t = 1$	Ratio	$t = 5$	Ratio	$t = 10$	Ratio
1/160	2.031×10^{-3}		4.324×10^{-1}		1.470×10^{2}	
1/320	5.072×10^{-4}	4.01	9.346×10^{-2}	4.63	2.399	61.2
1/640	1.268×10^{-4}	4.00	2.256×10^{-2}	4.14	3.557×10^{-1}	6.74
1/1280	3.169×10^{-5}	4.00	5.590×10^{-3}	4.04	7.895×10^{-2}	4.51
1/2560	7.922×10^{-6}	4.00	1.395×10^{-3}	4.01	1.917×10^{-2}	4.12

Table 5 Relative errors of Runge–Kutta solutions for Example 5, part I

h	$t = 1$	Ratio	$t = 5$	Ratio	$t = 10$	Ratio
1/10	1.425×10^{-2}		9.996×10^{-1}		1.000	
1/20	1.307×10^{-3}	10.9	7.989×10^{-1}	1.25	1.000	1.00
1/40	9.901×10^{-5}	13.2	1.831×10^{-1}	4.36	9.956×10^{-1}	1.00
1/80	6.814×10^{-6}	14.5	1.815×10^{-2}	10.1	4.886×10^{-1}	2.04
1/160	4.470×10^{-7}	15.2	1.382×10^{-3}	13.1	5.824×10^{-2}	8.39

The solution in this example grows even more rapidly than the one in Example 4:

$$\|\mathbf{X}(1)\| \doteq 3.401 \times 10^{4}, \quad \|\mathbf{X}(3)\| \doteq 1.084 \times 10^{21},$$
$$\|\mathbf{X}(5)\| \doteq 2.726 \times 10^{51}, \quad \|\mathbf{X}(10)\| \doteq 1.731 \times 10^{188}.$$

Numerical results from the Crank–Nicolson scheme for the problem are reported in Fig. 3 over the time interval $[0, 3]$, and in Tables 3 and 4 for relative numerical solution errors at $t = 1, 5$, and 10. The exact solution grows so fast that the numerical solution from the Crank–Nicolson scheme has a reasonably good accuracy only at a small time such as $t = 1$ when the stepsize h is relatively large (say, for h around 1/40). Note that with $h = 1/40$, the MATLAB program fails to provide a solution value at $t = 10$. To have better numerical solutions, we need to decrease the value of the step-size h. Table 4 shows the improvement in the solution accuracy when h becomes smaller.

Table 6 Relative errors of Runge–Kutta solutions for Example 5, part II

h	$t = 1$	Ratio	$t = 5$	Ratio	$t = 10$	Ratio
1/160	4.470×10^{-7}		1.382×10^{-3}		5.824×10^{-2}	
1/320	2.862×10^{-8}	15.6	9.510×10^{-5}	14.5	4.491×10^{-3}	13.0
1/640	1.810×10^{-9}	15.8	6.235×10^{-6}	15.3	3.084×10^{-4}	14.6
1/1280	1.139×10^{-10}	15.9	3.991×10^{-7}	15.6	2.019×10^{-5}	15.3
1/2560	6.982×10^{-12}	16.3	2.525×10^{-8}	15.8	1.291×10^{-6}	15.6

More dramatic improvement on the numerical solution accuracy is achieved when the Runge–Kutta method (8) and (13)–(16) is applied. The corresponding numerical results are reported in Tables 5 and 6. When t is larger, we also need to use a smaller h to get numerical solutions with acceptable accuracy. Again, for this example, considering both the numerical solution accuracy and the costs, we see that the 4th order Runge–Kutta method is prefered to the Crank–Nicolson scheme.

5 Concluding remarks

In this paper, we study the numerical solution through discretization methods for an initial value problem of a linear matrix ordinary differential equation of the form (3). A family of one step methods and the Runge–Kutta methods are considered. As an illustration of the theoretical study of the methods, rigorous convergence and error analysis is provided for the family of one step methods. The family of the methods include as special cases the Euler, the backward Euler, and the Crank–Nicolson schemes. Under appropriate assumptions on the smoothness of the coefficient matrix function, we prove that the Crank–Nicolson scheme has a convergence order 2 and for the remaining methods in the family, the convergence order is 1. The Crank–Nicolson scheme and a 4th order Runge–Kutta method are tested on two examples whose exact solutions display fast exponential growth. For such problems with fast growing solutions, a numerical method may not produce good numerical results when the step-size is not small enough. With proper choice of the step-size, however, both numerical methods show satisfactory performance on the two examples. The numerical results also indicate that for problems with smooth exact solutions, higher order methods are generally prefered.

As is noted in [2], choice of the simple initial value condition (2) is not a restriction. Consider the problem

$$\begin{aligned} \mathbf{Z}'(t) &= \mathbf{B}(t)\,\mathbf{Z}(t), \\ \mathbf{Z}(0) &= \mathbf{Z}_0, \end{aligned}$$

where the initial value $\mathbf{Z}_0$ is non-singular. Let $\mathbf{A}(t) = \mathbf{Z}_0^{-1}\mathbf{B}(t)\,\mathbf{Z}_0$ and define $\mathbf{X}(t)$ to be the solution of (1)–(2). Then it is easy to verify the relation

$$\mathbf{Z}(t) = \mathbf{Z}_0\mathbf{X}(t).$$

The methods and their convergence analysis in this paper can be extended to solving the initial value problem of a general first order matrix ordinary differential equation

$$\begin{aligned}\mathbf{X}'(t) &= \mathbf{F}(t, \mathbf{X}(t)), \quad t \in [0, T],\\ \mathbf{X}(0) &= \mathbf{I},\end{aligned}$$

where $\mathbf{F}(\cdot, \cdot)$ satisfies a uniform Lipschitz condition with respect to its second argument:

$$\|\mathbf{F}(t, \mathbf{Y}) - \mathbf{F}(t, \mathbf{Z})\| \leq c_0\|\mathbf{Y} - \mathbf{Z}\|$$

for any $m \times m$ matrices $\mathbf{Y}$ and $\mathbf{Z}$, with the Lipschitz constant c_0 independent of t. This can be accomplished in a similar spirit as that for the case of solving a scalar ordinary differential equation; however, rather delicate analysis will be required to accommodate the complications caused by the fact that the unknown function is matrix-valued. One may even try to develop structure-preserving numerical algorithms [9] to solve matrix ordinary differential equations as long as the equations arise in an application area where preserving the structure is important. These topics are worth further studying.

Acknowledgments We thank the referees for their valuable comments and suggestions on the previous manuscript.

References

1. S.Ü. Altınbaşak, M. Demiralp, Solutions to linear matrix ordinary differential equations via minimal, regular, and excessive space extension based universalization: Perturbation matrix splines, convergence and error estimate issues for polynomial coefficients in the homogeneous case. J. Math. Chem. **48**, 253–265 (2010)
2. S.Ü. Altınbaşak, M. Demiralp, Solutions to linear matrix ordinary differential equations via minimal, regular, and excessive space extension based universalization: Convergence and error estimates for truncation approximants in the homogeneous case with premultiplying polynomial coefficient matrix. J. Math. Chem. **48**, 266–286 (2010)
3. P.W. Atkins, J.D. Paula, *Physical Chemistry*, 7th edn. (Oxford University Press, Oxford, 2002)
4. Atkinson, *An Introduction to Numerical Analysis*, 2nd edn. (Wiley, New York, 1989)
5. K. Atkinson, W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, 3rd edn. (Springer, Berlin, 2009)
6. K. Atkinson, W. Han, D. Stewart, *Numerical Solution of Ordinary Differential Equations* (Wiley, New York, 2009)
7. J.C. Butcher, *Numerical Methods for Ordinary Differential Equations*, 2nd edn. (Wiley, New York, 2008)
8. A. Canada, P. Drábek, A. Fonda (eds.), *Handbook of Differential Equations: Ordinary Differential Equations*, vol. 3 (North-Holland, 2006)
9. E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration*, 2nd edn. (Springer, Berlin, 2006)
10. E. Hairer, S.P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I*, 2nd edn. (Springer, Berlin, 1993)
11. D.A. McQuarrie, *Quantum Chemistry*, 2nd edn. (University Science Books, California, 2008)

12. A.C. Norris, *Computational Chemistry: An Introduction to Numerical Methods* (Wiley, New York, 1981)
13. E. Zeidler, *Nonlinear Functional Analysis and its Applications. I: Fixed-point Theorems* (Springer, New York, 1985)