



Continuous numerical solutions of coupled mixed partial differential systems using Fer's factorization

Sergio Blanes^{a,*}, Lucas Jódar^b

^aDepartament de Física Teòrica, Universitat de València, Dr. Moliner 50, 46100-Burjassot, València, Spain

^bDepartamento de Matemática Aplicada, Universidad Politécnica, Camino de Vera, 22 012-Valencia, Spain

Received 4 May 1998; received in revised form 18 September 1998

Abstract

In this paper continuous numerical solutions expressed in terms of matrix exponentials are constructed to approximate time-dependent systems of the type $\mathbf{u}_t - \mathbf{A}(t)\mathbf{u}_{xx} - \mathbf{B}(t)\mathbf{u} = \mathbf{0}$, $0 < x < p$, $t > 0$, $\mathbf{u}(0, t) = \mathbf{u}(p, t) = \mathbf{0}$, $\mathbf{u}(x, 0) = \mathbf{f}(x)$, $0 \leq x \leq p$. After truncation of an exact series solution, the numerical solution is constructed using Fer's factorization. Given $\varepsilon > 0$ and t_0, t_1 , with $0 < t_0 < t_1$ and $D(t_0, t_1) = \{(x, t); 0 \leq x \leq p, t_0 \leq t \leq t_1\}$ the error of the approximated solution with respect to the exact series solution is less than ε uniformly in $D(t_0, t_1)$. An algorithm is also included. © 1999 Elsevier Science B.V. All rights reserved.

AMS classification: 65M15, 34A50, 35C10, 35A50

Keywords: Mixed time-dependent partial differential systems; Accurate solution; A priori error bounds; Fer's factorization; Algorithm

1. Introduction

Systems of partial differential equations are frequent in many different problems such as the study of heat conduction and diffusion problems [21], or in the analysis of pollutant migration through soil modelling coupled thermoelastoplastic hydraulic response clays [22]. In the evaluation of coupled microwave heating processes the constant coefficient model often leads to misleading results due to the complexity of the field distribution within the over and the variation in dielectric properties of material with temperature, moisture content, density and other properties [23, 15, Ch. 3].

In this paper we consider mixed problems for time-dependent systems of the type

$$\mathbf{u}_t(x, t) - \mathbf{A}(t)\mathbf{u}_{xx}(x, t) - \mathbf{B}(t)\mathbf{u}(x, t) = \mathbf{0}, \quad 0 < x < p, \quad t > 0, \quad (1)$$

* Corresponding author. Tel.: 34-6-3983148; fax: 34-6-3642345; e-mail: blanes@evalvx.uv.es.

$$u(0, t) = u(p, t) = 0, \quad t > 0, \quad (2)$$

$$u(x, 0) = f(x), \quad 0 \leq x \leq p, \quad (3)$$

where the unknown $u(x, t)$ and the right-hand side $f(x)$ are vectors in \mathcal{C}^r , and $A(t)$, $B(t)$ are continuous $\mathcal{C}^{r \times r}$ valued functions such that

there exists a positive number ϱ such that for all $t \geq 0$

$$\text{and every eigenvalue } z \text{ of } (A(t) + A^H(t))/2, \quad z \geq \varrho > 0 \quad (4)$$

where $A^H(t)$ denotes the Hermitian conjugate of the matrix $A(t)$. Problem (1)–(3) has been treated in [25] for the case where $A(t) = A$ and $B(t) = B$ are constant matrices, and in [19] for the case where $B(t) = 0$ and $A(t)$ is an analytic matrix function.

The aim of this paper is not the comparison with respect to other discrete methods but the construction of exact and continuous numerical solutions of problem (1)–(3) in terms of matrix exponentials, with a prefixed accuracy in a bounded subdomain. In spite of the expensive cost of the numerical computation of matrix exponentials the proposed method has several possible important advantages:

- (i) It permits the determination of a priori error bounds for the constructed analytic-numerical solutions, in terms of the available information of the problem.
- (ii) The matrix coefficient $A(t)$ does not need to be an analytic function, but only continuous, see [19].
- (iii) With respect to discrete methods, the constructed approximation is defined simultaneously for all the points (x, t) of the prefixed subdomain, and not only at a discrete mesh of points.

The organization of the paper is as follows. Section 2 deals with a revisited version of the error analysis developed in [4], adapted to the problem

$$V'(t) = [B(t) - \lambda^2 A(t)]V(t), \quad V(a_0) = V_0, \quad a_0 \leq t \leq a_1, \quad (5)$$

using Fer's factorization to approximate the solution of (5) by matrix exponentials. In particular, using the concept of logarithmic norm, error bounds given in [4] are improved. In Section 3 an exact series solution of Problem (1)–(3), is constructed under hypothesis (4), using a separation of variables technique. Given an admissible error $\varepsilon > 0$ and $t_1 > t_0 > 0$ we propose a truncation strategy so that the error of the truncated series be less than ε in

$$D(t_0, t_1) = \{(x, t); \quad 0 \leq x \leq p, \quad 0 < t_0 \leq t \leq t_1\}. \quad (6)$$

Section 4 deals with the construction of Fer's approximations to each of the exact solutions of vector problems of the type

$$T'_n(t) = \left[B(t) - \left(\frac{n\pi}{p} \right)^2 A(t) \right] T_n(t), \quad T_n(0) = c_n, \quad 1 \leq n \leq n_0, \quad (7)$$

where

$$c_n = \frac{2}{p} \int_0^p f(x) \sin \left(\frac{n\pi x}{p} \right) dx, \quad (8)$$

is the sine Fourier series coefficient of $f(x)$ and n_0 is the truncation index. Given $\varepsilon > 0$ we determine the index m of Fer's approximations $T_n^{[m]}(t)$, so that the error of the numerical approximation of Problem (1)–(3), after replacing the exact solution $T_n(t)$ of Problem (7)–(8) by its Fer's approximation $T_n^{[m]}(t)$, be smaller than ε uniformly in $D(t_0, t_1)$, when the whole interval of integration is split in subintervals. Section 5 provides the algorithm and an illustrative example.

Throughout this paper the set of all the eigenvalues of a matrix D in $\mathcal{C}^{r \times r}$ is denoted by $\sigma(D)$ and the spectral radius of D , denoted by $\rho(D)$ is the maximum of the set $\{|z|; z \in \sigma(D)\}$. We denote by $\|D\|$ the 2-norm of D , [13, p. 56; 16, p. 295]:

$$\|D\| = \sup_{y \neq 0} \frac{\|Dy\|_2}{\|y\|_2} = \max\{|\omega|^{1/2}; \omega \in \sigma(D^H D)\},$$

where for a vector $y \in \mathcal{C}^r$, $\|y\|_2 = (y^H y)^{1/2}$ is the usual euclidean norm of y . In accordance with [12, p. 110; 14, p. 59], the logarithmic norm $\mu(D)$ is defined by

$$\mu(D) = \lim_{h \rightarrow 0, h > 0} \frac{\|I + hD\| - 1}{h},$$

and satisfies

$$|\mu(D)| \leq \|D\|, \quad (9)$$

$$\mu(\alpha D) = \alpha \mu(D) \quad \text{for } \alpha \geq 0, \quad (10)$$

$$\mu(D) = \max \left\{ \omega; \omega \in \sigma \left(\frac{D + D^H}{2} \right) \right\}. \quad (11)$$

2. Fer approximation to $V' = [B(t) - \lambda^2 A(t)]V$

The interest in recovering qualitative features of the exact solution of time dependent matrix differential equations, and in particular the fundamental solution of a time-dependent linear system, has claimed the attention of many authors who have developed methods based on Lie groups. Although Lie groups are not mentioned in [17], the well-known method of the iterated commutators is a rediscovering by Iserles of Fer's approach, and is now being intensively investigated as a powerful tool to treat both linear and nonlinear differential equations on Lie groups and other manifolds [3, 27]. Application of Fer's approximation as a symplectic integrator may be found in [6], and may be found as a tool for solving certain initial value problems for linear partial differential equations in [7]. Apart from [11] other recent relevant works related to Fer's method are [4, 8, 20].

In this section, starting from recent results of [4], we introduce some improvements in the error analysis of Fer's method addressed to find a connection between the order of approximation and a prefixed accuracy.

Fer's algorithm approximates the solution $V(t)$ of the matrix initial problem

$$V'(t) = S(t)V(t) \quad V(0) = I, \quad (12)$$

by a product of matrix exponentials. Convergence of the approximations appear already in Fer's original work [11]. Our starting point is Section 3 of [4] where, using a slightly different argument,

the convergence region is enlarged and upper error bounds are improved. The aim of this section is to improve some of the results given in [4] using the concept of the logarithmic norm of a matrix and introducing a different iterative mapping. We recall that Fer's expansion is generated by the following recursive scheme:

$$\begin{aligned} V &= e^{F_1} e^{F_2} \dots e^{F_m} V_m, \\ V'_m &= S_m(t) V_m, \quad V_m(0) = I, \quad m = 1, 2, 3, \dots, \end{aligned} \quad (13)$$

where $F_m(t)$, $S_m(t)$ are given by

$$\begin{aligned} F_{m+1}(t) &= \int_0^t S_m(s) ds, \quad S_0(t) = S(t), \quad m = 0, 1, 2, \dots \\ S_{m+1} &= \int_0^1 dx \int_0^x du e^{-(1-u)F_{m+1}} [S_m, F_{m+1}] e^{(1-u)F_{m+1}} \end{aligned} \quad (14)$$

where $[P, Q] = PQ - QP$.

When after m steps we impose $V_m(t) = I$ we are left with an approximation $V^{[m]}(t)$ to the exact solution $V(t)$. Let us consider that the matrix $S(t)$ is bounded and $\|S(t)\|$ is a piecewise continuous function such that $\|S(t)\| \leq k(t) \equiv k^{(0)}(t)$. Fer's algorithm, Eqs. (13)–(14), provides then a recursive relation among corresponding bounds $k^{(m)}(t)$ for $\|S_m(t)\|$. Let us denote $K^{(m)}(t, 0) \equiv \int_0^t k^{(m)}(s) ds$. Using the mapping

$$M(x) = \int_0^x \frac{1 - e^{2s}(1 - 2s)}{2s} ds,$$

presented in [4], we can take as bounds (for $\|F_{m+1}(t)\|$), $K^{(m+1)}(t, 0) = M(K^{(m)}(t, 0))$ and the convergence is assured if $K^{(m)}(t, 0) < \xi$ with $\xi = 0.8604065$. But, considering that for $0 < x < \xi$

$$M(x) < x^2/\xi < x, \quad (15)$$

we can use the mapping

$$K^{(m+1)}(t, 0) = G(K^{(m)}(t, 0)), \quad G(x) = x^2/\xi. \quad (16)$$

Then $\lim_{m \rightarrow \infty} K^{(m)}(t, 0) = 0$ if 0 is a stable fixed point for the iteration and $K^{(0)}$ is within its basin of attraction. It is clear that $x = 0$ is a stable fixed point of $x = G(x)$ and $x = \xi$ is the next, unstable, fixed point [10]. Thus we have still assured the convergence of Fer's expansion for values of time t such that

$$\int_0^t \|S(s)\| ds \leq K^{(0)}(t, 0) < 0.8604065. \quad (17)$$

Note that if $K^{(m)}(t, 0) < \xi$ then $K^{(m+1)}(t, 0) < K^{(m)}(t, 0)$. Expression (16) is simpler than using the mapping $M(x)$. Further it has the same convergence domain and provides the iteration

$$K^{(m+1)} = \xi(K^{(0)}/\xi)^{2^m}.$$

The next result provides a priori error bounds of the theoretical solution of the matrix problem

$$V'(t) = [B(t) - \lambda^2 A(t)]V(t), \quad V(a_0) = V_0, \quad a_0 \leq t \leq a_1, \quad \lambda > 0. \quad (18)$$

Theorem 1. Under hypothesis (4), let $\alpha(a_0, a_1)$, $\beta(a_0, a_1)$ be positive numbers defined by

$$\alpha(a_0, a_1) = \min \left\{ z \in \sigma \left(\frac{A(t) + A^H(t)}{2} \right); \quad a_0 \leq t \leq a_1 \right\}, \tag{19}$$

$$\beta(a_0, a_1) \geq \max \left\{ \rho \left(\frac{B(t) + B^H(t)}{2} \right); \quad a_0 \leq t \leq a_1 \right\}. \tag{20}$$

Then the solution $V(t)$ of (18) satisfies

$$\|V(t)\| \leq \|V_0\| e^{(t-a_0)[\beta(a_0, a_1) - \lambda^2 \alpha(a_0, a_1)]}, \quad a_0 \leq t \leq a_1. \tag{21}$$

Proof. By [12, p. 114] the solution $V(t)$ of (18) satisfies

$$\|V(t)\| \leq \|V_0\| \exp \left(\int_{a_0}^t \mu(B(s) - \lambda^2 A(s)) ds \right), \quad a_0 \leq t \leq a_1. \tag{22}$$

By (11) one gets

$$\mu(B(s) - \lambda^2 A(s)) = \max \left\{ z \in \sigma \left(\frac{B(t) + B^H(t)}{2} - \lambda^2 \frac{A(t) + A^H(t)}{2} \right) \right\}, \tag{23}$$

and by [2, p. 246] it follows that

$$\sigma \left(\frac{B(t) + B^H(t)}{2} - \lambda^2 \frac{A(t) + A^H(t)}{2} \right) \subset \bigcup_{i=1}^r G_i(s), \quad 1 \leq i \leq r, \tag{24}$$

$$G_i(s) = \left\{ \omega \in \mathcal{C}; \quad \left| \omega + \lambda^2 \lambda_i \left(\frac{A(t) + A^H(t)}{2} \right) \right| \leq \rho \left(\frac{B(t) + B^H(t)}{2} \right) \right\}, \tag{25}$$

where $\lambda_i(A(t) + A^H(t)/2)$ is the i th eigenvalue of the matrix $(A(t) + A^H(t))/2$. By (22)–(25) and (19)–(20), it follows that

$$\begin{aligned} \mu(B(s) - \lambda^2 A(s)) &\leq \rho \left(\frac{B(t) + B^H(t)}{2} \right) - \lambda^2 \min \left\{ z \in \sigma \left(\frac{A(t) + A^H(t)}{2} \right) \right\} \\ &\leq \beta(a_0, a_1) - \lambda^2 \alpha(a_0, a_1), \quad a_0 \leq s \leq a_1. \end{aligned} \tag{26}$$

Thus the result is established. \square

Now from [4], see also [3], if $V(t)$ is the solution of

$$V'(t) = [B(t) - \lambda^2 A(t)]V(t), \quad V(a_0) = V_0, \quad a_0 \leq t \leq a_1 = a_0 + h,$$

and $V^{[m]}(t)$ is the Fer approximation of order m of $V(t)$, for $a_0 \leq t \leq a_1$ one gets

$$\|V(t) - V^{[m]}(t)\| \leq \|V_0\| K^{(m)}(a_1, a_0) e^{K^{(0)}(a_1, a_0) + 2K^{(m)}(a_1, a_0)}, \tag{27}$$

but in the problem we are considering it is preferable to take a refined bound. From Fer’s algorithm we have $V = V^{[m]}V_m$, so $V - V^{[m]} = VD_m$ with $D_m = I - V_m^{-1}$. If we consider that

$$(V_m^{-1})' = -V_m^{-1}S_m, \tag{28}$$

then

$$D_m(t) = - \int_{a_0}^t V_m^{-1}(s)S_m(s) ds. \tag{29}$$

Taking norms and considering $\|V_m^{-1}(s)\| \leq e^{K^{(m)}(a_1, a_0)s}$ we have $\|D_m(t)\| \leq e^{K^{(m)}(a_1, a_0)t}K^{(m)}(a_1, a_0)$, then

$$\|V(t) - V^{[m]}(t)\| \leq \|V\|e^{K^{(m)}(a_1, a_0)t}K^{(m)}(a_1, a_0), \tag{30}$$

where

$$K^{(0)}(s, a_0) \geq \int_{a_0}^s \|B(t) - \lambda^2 A(t)\| dt. \tag{31}$$

By (30) and Theorem 1 for $a_0 \leq t \leq a_1$ one gets

$$\|V(t) - V^{[m]}(t)\| \leq \|V\|_0 e^{(t-a_0)[\beta(a_0, a_1) - \lambda^2 \alpha(a_0, a_1)]} e^{K^{(m)}(a_1, a_0)t} K^{(m)}(a_1, a_0). \tag{32}$$

Remark 2. Once we approximate the solution of (12) by the Fer approximation $V^{[m]}(t) = e^{F_1(t)} \dots e^{F_m(t)}$, it is necessary to compute the matrix exponentials $e^{F_i(t)}$, where matrices $F_i(t)$ are related by (14). Numerous algorithms for computing matrix exponentials have been proposed, but most of them are of dubious numerical quality, as is pointed in [24]. In accordance with [24], scaling and squaring with Padé approximants and a careful implementation of Parlett’s Schur decomposition method, see [13, Ch. 11], were found the less dubious of the nineteen methods scrutinized. A promising method for computing matrix exponentials has been recently proposed in [9].

3. Exact and approximated theoretical solutions

The eigenfunction method suggests to seek a series solution of Problem (1)–(3) of the form

$$u(x, t) = \sum_{n \geq 1} T_n(t) \sin\left(\frac{n\pi x}{p}\right), \tag{33}$$

where $T_n(t)$ is the \mathcal{C}^r -valued solution of the initial value Problem (7)–(8). The solution of the vector problem can be written in the form

$$T_n(t) = U_n(t)c_n, \tag{34}$$

where U_n is the solution of the matrix initial value problem

$$U_n'(t) = \left[B(t) - \left(\frac{n\pi}{p}\right)^2 A(t) \right] U_n(t), \quad U_n(0) = I. \tag{35}$$

By the Riemann–Lebesgue lemma, there exists a constant M such that

$$\|c_n\| \leq M, \quad n \geq 1, \tag{36}$$

and by Theorem 1 and (34), solution of (7)–(8) satisfies

$$\|T_n(t)\| \leq M e^{t[\beta(0,t_1) - (n\pi/p)^2 \alpha(0,t_1)]} \tag{37}$$

$$\leq M e^{t_1 \beta(0,t_1)} e^{-t_0 (n\pi/p)^2 \alpha(0,t_1)}, \quad 0 < t_0 \leq t \leq t_1, \tag{38}$$

$$\|T'_n(t)\| \leq (\|B(t)\| + (n\pi/p)^2 \|A(t)\|) \|T_n(t)\| \tag{39}$$

$$\leq \left(b(t_1) + \left(\frac{n\pi}{p}\right)^2 a(t_1) \right) M e^{t_1 \beta(0,t_1)} e^{-t_0 (n\pi/p)^2 \alpha(0,t_1)},$$

where

$$\begin{aligned} a(t_1) &= \max\{\|A(t)\|; 0 \leq t \leq t_1\}, \\ b(t_1) &= \max\{\|B(t)\|; 0 \leq t \leq t_1\}. \end{aligned} \tag{40}$$

Hence the series appearing taking termwise partial differentiation in (33), once with respect to t ,

$$\sum_{n \geq 1} T'_n(t) \sin\left(\frac{n\pi x}{p}\right),$$

and twice with respect to x ,

$$\begin{aligned} &\sum_{n \geq 1} \frac{n\pi}{p} T_n(t) \cos\left(\frac{n\pi x}{p}\right), \\ &- \sum_{n \geq 1} \left(\frac{n\pi}{p}\right)^2 T_n(t) \sin\left(\frac{n\pi x}{p}\right), \end{aligned}$$

are uniformly convergent in $D(t_0, t_1) = \{(x, t); 0 \leq x \leq p, 0 < t_0 \leq t \leq t_1\}$. By the derivation theorem of functional series [1, p. 402], one gets that $u(x, t)$ is termwise partially differentiable with respect to the variable t and twice termwise partially differentiable with respect to the variable x . By (7)–(8) it follows that

$$u_t(x, t) = B(t)u(x, t) + A(t)u_{xx}(x, t), \quad 0 \leq x \leq p, \quad t > 0.$$

If each component f_j of $f = (f_1, \dots, f_r)^T$ satisfies one of the conditions:

- (i) f_j is locally of bounded variation at every point x in $[0, p]$.
 - (ii) f_j admits one-side derivatives $(f'_j)_R(x)$ and $(f'_j)_L(x)$ at every point x in $[0, p]$,
- (41)

and f is continuous in $[0, p]$ with $f(0) = f(p) = \mathbf{0}$, then by [5, p. 57] it follows that

$$f(x) = \sum_{n \geq 1} c_n \sin\left(\frac{n\pi x}{p}\right) = u(x, 0), \quad 0 \leq x \leq p. \tag{42}$$

Thus the following result has been established:

Theorem 3. Let $A(t), B(t)$ continuous $\mathcal{C}^{r \times r}$ -valued functions such that condition (4) is satisfied. Let $f(x)$ be continuous in $[0, p]$ with $f(0) = f(p) = \mathbf{0}$ and let each component f_j of f satisfy one of the conditions of (41). Then $u(x, t)$ defined by (33) when $T_n(t)$ satisfies (7)–(8), is a solution of Problem (1)–(3).

Given $\varepsilon > 0$ we are interested, under hypotheses of Theorem 3, in the determination of an index n_0 so that

$$\left\| \sum_{n \geq n_0} T_n(t) \sin\left(\frac{n\pi x}{p}\right) \right\| < \frac{\varepsilon}{2}, \quad (x, t) \in D(t_0, t_1). \tag{43}$$

From (38), considering that

$$\sum_{n > n_0} e^{-t_0(n\pi/p)^2 \alpha(0, t_1)} \leq \int_{n_0}^{\infty} e^{-t_0(x\pi/p)^2 \alpha(0, t_1)} dx = \frac{p}{2\sqrt{\pi t_0 \alpha(0, t_1)}} \operatorname{erfc}\left(\frac{n_0 \pi}{p} \sqrt{t_0 \alpha(0, t_1)}\right)$$

where $\operatorname{erfc}(t) = (2/\sqrt{\pi}) \int_t^{\infty} e^{-x^2} dx$ is the complementary error function, and from (43) it follows that

$$\sum_{n > n_0} \|T_n(t)\| \leq \frac{M e^{t_1 \beta(0, t_1)} p}{2\sqrt{\pi t_0 \alpha(0, t_1)}} \operatorname{erfc}\left(\frac{n_0 \pi}{p} \sqrt{t_0 \alpha(0, t_1)}\right). \tag{44}$$

Hence, taking the first positive integer n_0 such that

$$\operatorname{erfc}\left(\frac{n\pi}{p} \sqrt{t_0 \alpha(0, t_1)}\right) < \frac{\varepsilon \sqrt{\pi t_0 \alpha(0, t_1)}}{M e^{t_1 \beta(0, t_1)} p}, \tag{45}$$

inequality (43) holds and

$$\left\| u(x, t) - \sum_{n=1}^{n_0} T_n(t) \sin\left(\frac{n\pi x}{p}\right) \right\| < \frac{\varepsilon}{2}, \quad (x, t) \in D(t_0, t_1). \tag{46}$$

Thus the following result has been established:

Corollary 4. Under hypothesis of Theorem 3, let $t_1 > t_0 > 0$, $\varepsilon > 0$ and let $D(t_0, t_1)$ be defined by (6). If $u(x, t)$ is the exact series solution of Problem (1)–(3) given by Theorem 3, and n_0 is the first positive integer satisfying (45), then

$$u(x, t, n_0) = \sum_{n=1}^{n_0} T_n(t) \sin\left(\frac{n\pi x}{p}\right) \tag{47}$$

is an approximation satisfying (46).

4. Continuous numerical solution

Section 2 was concerned with the study of the local error, in the convergence domain, using Fer’s approximation for the solution of Problem (7). If we are interested in the approximation of

Problem (7) using Fer’s algorithm in an interval $[t_0, t_1]$ where Fer’s method is not convergent, then we may split the interval in subintervals with guaranteed convergence. In this section we address the following question: Given $\varepsilon > 0$, $0 < t_0 < t_1$ and n_0 given by Corollary 4, how to determine the order m of the Fer’s approximation $T_n^{[m]}(t)$ of the exact solution $T_n(t)$ of Problem (7)–(8) such that the error with respect to the exact solution be smaller than $\varepsilon/(2n_0)$ when the whole interval of integration is split in N subintervals and the algorithm is used in each interval, i.e. to look for m such that

$$\|T_n(t) - T_n^{[m]}(t)\| < \frac{\varepsilon}{2n_0}, \quad 0 < t_0 \leq t \leq t_1, \quad 1 \leq n \leq n_0. \tag{48}$$

Let $h > 0$ and consider the partition $0 = h_0 < h_1 < \dots < h_N = t_1$, where $h_j = jh$, $0 \leq j \leq N$ and $Nh = t_1$. If $h_j \leq t \leq h_{j+1}$, then we can write

$$T_n(t) = U_n(t, 0)c_n, \tag{49}$$

$$\frac{d}{dt}U_n(t, h_j) = \left[B(t) - \left(\frac{n\pi}{p}\right)^2 A(t) \right] U_n(t, h_j), \tag{50}$$

$$U_n(h_j, h_j) = I, \quad h_j \leq t \leq h_{j+1},$$

$$U_n(t, 0) = U_n(t, h_i)U_n(h_i, h_{i-1}) \cdots U_n(h, 0), \quad h_i \leq t \leq h_{i+1}. \tag{51}$$

Let us introduce the notation

$$U_n(t, h_j) = U_{n,j}(t), \quad U_n^{[m]}(t, h_j) = U_{n,j}^{[m]}(t), \tag{52}$$

$$0 \leq j \leq N - 1, \quad h_j \leq t \leq h_{j+1},$$

where for simplicity we will write $U_{n,j}(h_{j+1}) = U_{n,j}(h)$, $U_{n,j}^{[m]}(h_{j+1}) = U_{n,j}^{[m]}(h)$. Thus for $h_i \leq t \leq h_{i+1}$, $0 \leq i \leq N - 1$ we can write

$$\begin{aligned} U_n(t, 0) - U_n^{[m]}(t, 0) &= U_{n,i}(t)U_{n,i-1}(h) \cdots U_{n,0}(h) - U_{n,i}^{[m]}(t)U_{n,i-1}^{[m]}(h) \cdots U_{n,0}^{[m]}(h) \\ &= (U_{n,i}(t) - U_{n,i}^{[m]}(t))U_{n,i-1}(h) \cdots U_{n,0}(h) \\ &\quad + U_{n,i}^{[m]}(t)(U_{n,i-1}(h) - U_{n,i-1}^{[m]}(h))U_{n,i-2}(h) \cdots U_{n,0}(h) \\ &\quad + \cdots \\ &\quad + U_{n,i}^{[m]}(t) \cdots U_{n,i-j+1}^{[m]}(h)(U_{n,i-j}(h) - U_{n,i-j}^{[m]}(h))U_{n,i-j-1}(h) \cdots U_{n,0}(h) \\ &\quad + \cdots \\ &\quad + U_{n,i}^{[m]}(t) \cdots U_{n,2}^{[m]}(h)(U_{n,0}(h) - U_{n,0}^{[m]}(h)). \end{aligned}$$

Hence

$$\begin{aligned} \|U_n(t, 0) - U_n^{[m]}(t, 0)\| &\leq \sum_{j=0}^i \|U_{n,i}^{[m]}(t)\| \cdots \|U_{n,i-j+1}^{[m]}(h)\| \|U_{n,i-j}(h) - U_{n,i-j}^{[m]}(h)\| \\ &\quad \times \|U_{n,i-j-1}(h)\| \cdots \|U_{n,0}(h)\|, \quad h_i \leq t \leq h_{i+1}. \end{aligned} \tag{53}$$

By (30), for $h_j \leq t \leq h_{j+1}$ it follows that

$$\|U_{n,j}^{[m]}(t)\| \leq \|U_{n,j}(t)\| e^{K_{n,j}^{(m)}(t,h_j)}, \tag{54}$$

$$\|U_{n,i-j}(t) - U_{n,i-j}^{[m]}(t)\| \leq \|U_{n,i-j}(t)\| K_{n,i-j}^{(m)}(t,h_{i-j}) e^{K_{n,i-j}^{(m)}(t,h_{i-j})}. \tag{55}$$

By (54)–(55) one gets

$$\begin{aligned} & \|U_{n,i}^{[m]}(t)\| \cdots \|U_{n,i-j+1}^{[m]}(h)\| \|U_{n,i-j}(h) - U_{n,i-j}^{[m]}(h)\| \\ & \leq \|U_{n,i}(t)\| \cdots \|U_{n,i-j+1}(h)\| \|U_{n,i-j}(h)\| K_{n,i-j}^{(m)}(h_{i-j+1}, h_{i-j}) \\ & \quad \times \exp(K_{n,i}^{(m)}(t, h_i) + K_{n,i-1}^{(m)}(h_i, h_{i-1}) + \cdots + K_{n,i-j}^{(m)}(h_{i-j+1}, h_{i-j})). \end{aligned} \tag{56}$$

Note that by Theorem 1, for $0 < t_0 \leq t \leq t_1$ one gets

$$\|U_{n,i}(t)\| \cdots \|U_{n,0}(h)\| \leq e^{t_1 \beta(0,t_1) - t_0(n\pi/p)^2 \alpha(0,t_1)}. \tag{57}$$

By (53)–(56) and (57) it follows

$$\begin{aligned} & \|U_n(t, 0) - U_n^{[m]}(t, 0)\| \leq e^{t_1 \beta(0,t_1) - t_0(n\pi/p)^2 \alpha(0,t_1)} \\ & \quad \times \sum_{j=0}^i K_{n,i-j}^{(m)}(h_{i-j+1}, h_{i-j}) \exp(K_{n,i}^{(m)}(t, h_i) + K_{n,i-1}^{(m)}(h_i, h_{i-1}) + \cdots + K_{n,i-j}^{(m)}(h_{i-j+1}, h_{i-j})) \\ & \quad h_i \leq t \leq h_{i+1}, N_0 = \frac{t_0}{h} \leq i \leq N - 1. \end{aligned} \tag{58}$$

Let $0 < \delta < 1$, and let n_0 be given by Corollary 4, then take $h > 0$ and select the integer N such that

$$N > t_1 \frac{a(t_1)(n_0\pi/p)^2 + b(t_1)}{\delta \xi}, \quad h = \frac{t_1}{N}. \tag{59}$$

Then taking

$$\delta_n = \frac{a(t_1)(n\pi/p)^2 + b(t_1)}{a(t_1)(n_0\pi/p)^2 + b(t_1)} \delta, \quad 1 \leq n \leq n_0, \tag{60}$$

where $\delta_n \leq \delta$ and $\delta_{n_0} = \delta$, one can consider

$$\begin{aligned} & \int_{jh}^t \left\| B(s) - \left(\frac{n\pi}{p}\right)^2 A(s) \right\| ds < K_{n,j}^{(0)}(t, h_j) = h \left[a(t_1) \left(\frac{n\pi}{p}\right)^2 + b(t_1) \right] < \delta_n \xi, \\ & jh \leq t \leq (j+1)h, \quad 0 \leq j \leq N - 1, \quad 1 \leq n \leq n_0. \end{aligned} \tag{61}$$

By (60) and (61) it follows that

$$K_{n,j}^{(0)}(t, h_j) \leq K_{n,j}^{(0)}(h_{j+1}, h_j) < \delta_n \xi, \tag{62}$$

$$K_{n,j}^{(m)}(t, h_j) = \xi \left(\frac{K_{n,j}^{(0)}(t, h_j)}{\xi} \right)^{2^m}, \tag{63}$$

$$K_{n,j}^{(m)}(t, h_j) \leq K_{n,j}^{(m)}(h_{j+1}, h_j) = \xi \left(\frac{K_{n,j}^{(0)}(h_{j+1}, h_j)}{\xi} \right)^{2^m} \leq \delta_{n,m}(\xi), \tag{64}$$

$$\delta_{n,m}(\xi) = \delta_n^{2^m} \xi, \quad \lim_{m \rightarrow \infty} K_{n,j}^{(m)}(t, h_j) = 0. \tag{65}$$

Under hypothesis (59), by (58), (65) it follows that

$$\begin{aligned} \|U_n(t, 0) - U_n^{[m]}(t, 0)\| &\leq e^{t_1 \beta(0, t_1) - t_0 (n\pi/p)^2 \alpha(0, t_1)} \delta_{n,m}(\xi) e^{\delta_{n,m}(\xi)} \sum_{j=0}^i e^{j \delta_{n,m}(\xi)}, \\ &= e^{t_1 \beta(0, t_1) - t_0 (n\pi/p)^2 \alpha(0, t_1)} \delta_{n,m}(\xi) e^{\delta_{n,m}(\xi)} \frac{e^{(i+1) \delta_{n,m}(\xi)} - 1}{e^{\delta_{n,m}(\xi)} - 1} \\ &\quad ih \leq t \leq (i+1)h, \quad N_0 \leq i \leq N-1. \end{aligned} \tag{66}$$

Note that if $\alpha > 0, x > 0$, by the mean value theorem $e^{\alpha x} - 1 = \alpha x e^{\alpha s}$ for some $s \in]0, x[$. Hence, using that $e^x - 1 \geq x$, one gets

$$\frac{e^{\alpha x} - 1}{e^x - 1} \leq \alpha e^{\alpha s} \leq \alpha e^{\alpha x}, \quad x > 0, \quad \alpha > 0. \tag{67}$$

By (66), (67) one gets

$$\|U_n(t, 0) - U_n^{[m]}(t, 0)\| \leq e^{t_1 \beta(0, t_1) - t_0 (n\pi/p)^2 \alpha(0, t_1)} (i+1) \delta_{n,m}(\xi) e^{(i+2) \delta_{n,m}(\xi)},$$

and by (34) it follows that

$$\begin{aligned} \|T_n(t) - T_n^{[m]}(t)\| &\leq \|c_n\| N e^{t_1 \beta(0, t_1) - t_0 (n\pi/p)^2 \alpha(0, t_1)} \delta_{n,m}(\xi) e^{(N+1) \delta_{n,m}(\xi)}, \\ 0 < t_0 \leq t \leq t_1, \quad 1 \leq n \leq n_0. \end{aligned} \tag{68}$$

Let ρ_n be the unique root of equation

$$x e^{(N+1)x} = \frac{\varepsilon e^{t_0 (n\pi/p)^2 \alpha(0, t_1)}}{2 \|c_n\| N e^{t_1 \beta(0, t_1)} n_0}, \quad 1 \leq n \leq n_0 \tag{69}$$

and let m_n be the first positive integer m satisfying

$$2^m > \frac{\ln(\frac{\rho_n}{\xi})}{\ln(\delta_n)}, \tag{70}$$

then

$$2^m \ln(\delta_n) < \ln\left(\frac{\rho_n}{\xi}\right), \quad \delta_{n,m}(\xi) = \xi \delta_n^{2^m} < \rho_n, \tag{71}$$

and

$$\|T_n(t) - T_n^{[m]}(t)\| \leq \frac{\varepsilon}{2n_0}, \quad 0 \leq t \leq t_1, \quad 1 \leq n \leq n_0. \tag{72}$$

5. The algorithm and an example

We begin this section summarizing the algorithm proposed in Sections 3 and 4 for the construction of a continuous numerical solution of Problem (1)–(3), such that under hypothesis (4) satisfies

$$\left\| \mathbf{u}(x, t) - \sum_{n=1}^{n_0} \mathbf{T}_n^{[m]}(t) \sin\left(\frac{n\pi x}{p}\right) \right\| \leq \varepsilon, \quad (x, t) \in D(t_0, t_1). \tag{73}$$

STEP 1. Truncated theoretical approximation.

Given $0 < t_0 < t_1$, $\varepsilon > 0$:

- Compute M given by (36) and $\alpha(0, t_1)$, $\beta(0, t_1)$ given by (19) and (20).
- Take n_0 as the first positive integer n satisfying (45).

STEP 2. Construction of continuous numerical solution.

Given n_0 , let $0 < \delta < 1$ fixed and $\xi = 0.8604065$:

- Compute $a(t_1)$, $b(t_1)$ given by (40).
- Select $h > 0$ and an integer N such that $Nh = t_1$, and satisfying (59).
- Compute the root ρ_n of Eq. (69) for $1 \leq n \leq n_0$.
- Given ρ_n take the first positive integer m_n satisfying (70) for $1 \leq n \leq n_0$.
- Compute $\mathbf{T}_n^{[m]}(t)$ for $1 \leq n \leq n_0$.
- $\mathbf{u}(x, t, n_0, m) = \sum_{n=1}^{n_0} \mathbf{T}_n^{[m]}(t) \sin(n\pi x/p)$ is an approximate solution satisfying (73) uniformly for $(x, t) \in D(t_0, t_1)$.

Example. Let us consider Problem (1)–(3) in $\mathcal{C}^{2 \times 2}$ where

$$A(t) = \begin{pmatrix} 1 & \frac{1}{4}e^t \\ \frac{1}{4}e^{-t} & 1 \end{pmatrix}; \quad B(t) = \begin{pmatrix} 1 & -\frac{1}{4}e^t \\ -\frac{1}{4}e^t & 1 \end{pmatrix}$$

and $f(x)$ any function in \mathcal{C}^2 satisfying the hypotheses of Theorem 3, such that $\|c_n\| < M$ with $M = 1$. We will consider $p = 1$, $0 < t_0 \leq t \leq t_1$ with $t_0 = \frac{1}{5}$ and $t_1 = 1$ and the error bound $\varepsilon = 10^{-3}$. Then we can choose

$$\begin{aligned} \alpha(0, 1) &= 1 - \frac{1}{8}(e + e^{-1}), \\ \beta(0, 1) &= 1 + \frac{1}{8}(e + e^{-1}). \end{aligned}$$

The next step is to look for the minimum n_0 satisfying (45). That happens for $n_0 = 3$ where

$$\operatorname{erfc}\left(\frac{n_0\pi}{p}\sqrt{t_0\alpha(0, 1)}\right) = 2.9885 \times 10^{-6} < 1.5539 \times 10^{-4} = \varepsilon \frac{\sqrt{\pi t_0 \alpha(0, 1)}}{M e^{t_1 \beta(0, 1)}}.$$

Following the second step we choose $\delta = \frac{1}{11}$, and taking

$$a(t_1) = b(t_1) = 1 + \frac{1}{4}e$$

the number of steps to consider is

$$t_1 \frac{b(t_1) + \left(\frac{n_0\pi}{p}\right)^2 a(t_1)}{\delta \xi} = 1928.8 < 2000 = N$$

or equivalently $h = 1/2000$. The values of δ_n are

$$\delta_1 = 1.100 \times 10^{-2}, \quad \delta_2 = 4.097 \times 10^{-2}, \quad \delta_3 = \delta = \frac{1}{11}.$$

Now we compute the root ρ_n of Eq. (69) for $1 \leq n \leq 3$, giving us

$$\rho_1 = 7.009 \times 10^{-8}, \quad \rho_2 = 2.649 \times 10^{-6}, \quad \rho_3 = 4.577 \times 10^{-4}.$$

Finally we take the first positive integer m_n satisfying (70) for $1 \leq n \leq 3$. We obtain

$$2^{m_1} > 3.619, \quad 2^{m_2} > 3.972, \quad 2^{m_3} > 3.144$$

and so $m_1 = m_2 = m_3 = 2$. To sum up, using the Fer's factorization to second order for the three first terms of the series, the error of the approximate solution will be smaller than $\varepsilon = 10^{-3}$. For $\varepsilon = 10^{-4}$ (with $n_0 = 3$) and taking now $\delta = 1/22.8$ we can choose $N = 4000$. Following the same calculations we find also $m_1 = m_2 = m_3 = 2$.

Remark 5. Note that previous algorithm involves a free parameter δ with $0 < \delta < 1$ and its choice is significant. In accordance with (59) if $\delta \rightarrow 0$ then h must tend to 0 also. This means that we need to apply Fer's approximation more times, in more subintervals in which the original interval $[t_0, t_1]$ is split. If $\delta \rightarrow 1$, then we apply Fer's approximation a minor number of times but, by (70), the integer m proposed by the algorithm as the order of the Fer's approximation to construct $T_n^{[m]}(t)$, increases as well as $\delta \rightarrow 1$. Taking into account the complexity of the nested integrals (13)–(14), in general, we suggest to take $\delta < \frac{1}{2}$. Of course the optimal value of δ is also depending on the interval $[t_0, t_1]$ and the minimum of the positive eigenvalues of the real part of $A(t)$, see Condition (4).

Acknowledgements

S.B. has been supported by DGICYT under contract no. PB92-0820 and by EEC network no. ERBCHRXCT940456. The work of L.J. is supported by DGICYT under contract no. PB96-1321-C02-02 and Generalitat Valenciana under contract no. GV-C-CN-10057/96.

References

- [1] T.M. Apostol, *Mathematical Analysis*, Addison Wesley, Reading, MA, 1957.
- [2] F.L. Bauer, A.S. Householder, Absolute norms and characteristic roots, *Numer. Math.* 3 (1961) 241–246.
- [3] S. Blanes, Estudio de la evolución de sistemas dinámicos clásicos y cuánticos utilizando métodos algebraicos, Ph.D. Thesis, University of Valencia, 1998.
- [4] S. Blanes, F. Casas, J.A. Oteo, J. Ros, Magnus and Fer expansions for matrix differential equations: the convergence problem, *J. Phys. A* 31 (1998) 259–268.
- [5] J.W. Brown, R.V. Churchill, *Fourier Series and Boundary Value Problems*, McGraw-Hill, New York, 1978.
- [6] F. Casas, Fer's factorization as a symplectic integrator, *Numer. Math.* 74 (1996) 283–303.
- [7] F. Casas, Solution of linear partial differential equations by Lie algebraic methods, *J. Comput. Appl. Math.* 76 (1996) 159–170.
- [8] F. Casas, J.A. Oteo, J. Ros, Lie algebraic approach to Fer's expansion for classical Hamiltonian systems, *J. Phys. A* 24 (1991) 4037–4046.

- [9] E. Celledoni, A. Iserles, Approximating the exponential from a Lie algebra to a Lie group, DAMTP 1998/NA3, University of Cambridge, Cambridge, 1998.
- [10] P. Collet, J.-P. Eckmann, Iterated maps on the interval as dynamical systems, Progress in Physics, Birkhäuser, Boston, 1980.
- [11] F. Fer, Résolution de l'équation matricielle $\dot{U} = pU$ par produit infini d'exponentielles matricielles, Bull. Classe Sci. Acad. Roy. Bel. 44 (1958) 818–829.
- [12] T.M. Flett, Differential Analysis, Cambridge University Press, Cambridge, 1980.
- [13] G. Golub, C.F. Van Loan, Matrix Computations, Johns Hopkins University Press, Baltimore, MD, 1991.
- [14] E. Hairer, S.P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I, Springer, Berlin, 1980.
- [15] A.F. Harvey, Microwave Engineering, Academic Press, New York, 1963.
- [16] R.A. Horn, Ch. R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1993.
- [17] A. Iserles, Solving linear ordinary differential equations by exponentials of iterated commutators, Numer. Math. 45 (1984) 183–199.
- [18] A. Iserles, S.P. Nørsett, On the solution of linear differential equations in Lie groups, DAMTP 1997/NA3, University of Cambridge, Cambridge, 1997.
- [19] L. Jódar, E. Defez, Analytic-numerical solutions with a priori error bounds for the time dependent mixed partial differential problems, Comput. Math. Appl. 34 (1997) 49–59.
- [20] S. Klarsfeld, J.A. Oteo, Exponential infinite-product representations of the time displacement operator, J. Phys. A 22 (1989) 2687–2694.
- [21] A.I. Lee, J.M. Hill, On the general linear coupled system for diffusion in media with two diffusivities, J. Math. Anal. Appl. 89 (1982) 530–557.
- [22] R.W. Lewis, E. Hinton, P. Bettess, B.A. Schrefler, Methods in Transient and Coupled Problems, Wiley, New York, 1987.
- [23] A.C. Metaxas, R.J. Meredith, Industrial Microwave Heating, Peter Peregrinus, London, 1983.
- [24] C. Moler y C. Van Loan, Nineteen dubious ways to compute the exponential of a matrix, SIAM Rev. 20 (1978) 801–836.
- [25] E. Navarro, E. Ponsoda, L. Jódar, A matrix approach to the analytic-numerical solution of mixed partial differential systems, Comput. Math. Appl. 30 (1995) 99–109.
- [26] D.M. Pozar, Microwave Engineering, Academic Press, New York, 1990.
- [27] A. Zanna, The method of iterated commutators for ordinary differential equations on Lie groups, DAMTP 1996/NA12, University of Cambridge, Cambridge, 1996.