Lucas Jódar; Enrique A. Navarro; M. V. Ferrer
A matrix constructive method for the analytic-numerical solution of coupled partial differential systems


Persistent URL: http://dml.cz/dmlcz/134302

Terms of use:

© Institute of Mathematics AS CR, 1995

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these Terms of use.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project DML-CZ: The Czech Digital Mathematics Library http://project.dml.cz
A MATRIX CONSTRUCTIVE METHOD
FOR THE ANALYTIC-NUMERICAL SOLUTION
OF COUPLED PARTIAL DIFFERENTIAL SYSTEMS

LUCAS JÓDAR, ENRIQUE NAVARRO, M.V. FERRER, Valencia

(Received February 22, 1994)

Summary. In this paper we construct analytic-numerical solutions for initial-boundary value systems related to the equation $u_t - Au_{xx} - Bu = 0$, where $B$ is an arbitrary square complex matrix and $A$ is a matrix such that the real part of the eigenvalues of the matrix $\frac{1}{2}(A + A^H)$ is positive. Given an admissible error $\varepsilon$ and a finite domain $G$, and analytic-numerical solution whose error is uniformly upper bounded by $\varepsilon$ in $G$, is constructed.

Keywords: Schur decomposition, partial differential system, eigenvalues bound, matrix norms, analytic-numerical solution, error bounds

AMS classification: 15A24, 15A45, 15A60, 65F15, 65N15

1. INTRODUCTION

Coupled systems of partial differential equations appear in many physical problems such as in heat diffusion [4, 10, 13], magnetohydrodynamics [3, 18], propagation of signals [20], armament models [9], neutron and nerve conduction problems [15, 16, 17], etc. Methods based on the transformation of a coupled system into a new system of uncoupled equations may be found in [7, 21], and its drawbacks have been treated in [5].

---

1 This work has been supported by the Generalitat Valenciana grant GV1118/93.
The aim of this paper is to construct analytic approximate solutions of coupled systems of the type

\begin{align}
(1.1) \quad & u_t(x, t) - Au_{xx}(x, t) - Bu(x, t), \quad 0 < x < p, \quad t > 0 \\
(1.2) \quad & u(0, t) = u(p, t) = 0, \quad t > 0 \\
(1.3) \quad & u(x, 0) = F(x), \quad 0 \leq x \leq p,
\end{align}

where \( u = (u_1, \ldots, u_m)^T \), \( F(x) \) is a vector in \( \mathbb{C}^m \) and \( A, B \) are \( m \times m \) complex matrices such that if \( A^H \) denotes the conjugate transpose matrix of \( A \), then the following spectral condition is satisfied:

\begin{equation}
(1.4) \quad \text{Every eigenvalue of the matrix } (A + A^H)/2 \text{ is positive.}
\end{equation}

For the case when the matrix \( B \) is zero, the above problem has been treated in \([11, 12]\). This paper that may be regarded as a continuation of \([11]\), is organized as follows. In Section 2 we prove that under the hypothesis \((1.4)\) the spectrum of the matrix \( B - (n\pi/p)^2 A \) is located in the left-hand side of the complex plane for large values of the positive integer \( n \). Then a convergent series solution of the problem \((1.1)-(1.3)\) is constructed. Section 3 is concerned with the construction of a finite computable analytic numerical solution of the problem such that the approximation error is smaller than \( \varepsilon \) in a prescribed domain \( G = [0, p] \times [t_0, t_1] \), where \( 0 < t_0 < t_1 \). This construction is performed by truncation of the infinite series and by approximation of certain matrix exponentials by truncation of its Taylor series expansion. Thus the approximation is only expressed in terms of the data.

Throughout this paper we denote by \( \mathbb{C}^{m \times m} \) the set of all \( m \times m \) complex matrices. For a matrix \( B \) in \( \mathbb{C}^{m \times m} \), the set of all eigenvalues of \( B \) is denoted by \( \sigma(B) \). The spectral radius of \( B \) will be denoted by \( r(B) \) and defined as the maximum of the set \( \{ |z|; \ z \in \sigma(B) \} \). If \( B \) lies in \( \mathbb{C}^{m \times m} \) we denote by \( \|B\| \) its 2-norm and by \( \|B\|_F \) its Frobenius norm, see \([8, p. 14]\), and we recall that \( \|B\| \leq \|B\|_F \leq m^{1/2} \|B\| \), \([8, p. 15]\). Finally, if \( B \) is a matrix in \( \mathbb{C}^{m \times m} \) then there exists a unitary matrix \( Q \in \mathbb{C}^{m \times m} \) such that \( Q^H B Q = D + N \), where \( D \) is a diagonal matrix and \( N \in \mathbb{C}^{m \times m} \) is strictly upper triangular. This is called the Schur decomposition of \( B \), and \( \|N\|_F \leq \|B\|_F \), see \([8, p. 192–193]\). If we denote

\begin{equation}
(1.5) \quad \alpha(B) = \max \{ \text{Re}(z); \ z \in \sigma(B) \}, \quad M_S(t) = \sum_{k=0}^{m-1} \frac{\|N^k\|}{k!},
\end{equation}

then it follows from \([8, p. 369]\) that

\begin{equation}
(1.6) \quad \| \exp(tB) \| \leq \exp\left(t\alpha(B)\right) M_S(t), \quad t \geq 0.
\end{equation}
2. ALGEBRAIC PRELIMINARIES AND CONSTRUCTION OF A SERIES SOLUTION

For the sake of clarity of the presentation we recall two results that will be used below.

**Theorem 1** ([19, p. 395]). Decomposing an arbitrary matrix $M \in \mathbb{C}^{m \times m}$ into $M = N_1 + iN_2$, where $N_1$ and $N_2$ are the hermitian matrices $N_2 = (M - M^H)/(2i)$, $N_1 = \frac{1}{2}(M + M^H)$, then for every eigenvalue $z \in \sigma(M)$ one has

$$z_{\text{min}}(N_1) \leq \Re z \leq z_{\text{max}}(N_1)$$

where $z_{\text{min}}(N_1)$ and $z_{\text{max}}(N_1)$ denote respectively the minimum and the maximum of the real eigenvalues of the matrix $N_1$.

**Theorem 2** ([2, p. 246]). Let $C$ and $D$ be complex hermitian matrices in $\mathbb{C}^{m \times m}$ and let $r(D)$ be the spectral radius of $D$. If $\sigma(C) = \{\lambda_i(C); 1 \leq i \leq m\}$ and $G_i = \{z \in \mathbb{C}; |z - \lambda_i(C)| \leq r(D)\}$, then $\sigma(C + D) \subseteq \bigcup_{i=1}^{m} G_i$.

Now we are in good position to locate the spectrum of $B - (n\pi/p)^2A$.

**Theorem 3.** Let $A$ be a complex matrix in $\mathbb{C}^{m \times m}$ satisfying the condition (1.4), and let $\gamma(A)$ be defined by

$$\gamma(A) = \min \left\{ \Re z \left| z \in \sigma\left(\frac{1}{2}(A + A^H)\right) \right. \right\}.$$

Then

$$\Re z < r\left(\frac{1}{2}(B + B^H)\right) - (n\pi/p)^2\gamma(A), \quad z \in \sigma(B - (n\pi/p)^2A), \; n \geq 1.$$

**Proof.** Let us consider the decompositions $A = H_1 + iH_2$, $B = S_1 + iS_2$, where

$$H_1 = \frac{1}{2}(A + A^H), \quad S_1 = \frac{1}{2}(B + B^H), \quad H_2 = \frac{1}{2i}(A - A^H), \quad S_2 = \frac{1}{2i}(B - B^H),$$

$$B - (n\pi/p)^2A = S_1 - (n\pi/p)^2H_1 + i[S_2 - (n\pi/p)^2H_2].$$

If we apply Theorem 2 to the hermitian matrices $C = S_1$ and $D = -(n\pi/p)^2H_1$, it follows that

$$\sigma(S_1 - (n\pi/p)^2H_1) \subseteq \bigcup_{i=1}^{m} G_i, \quad G_i = \{z \in \mathbb{C}; |z + (n\pi/p)^2\lambda_i(H_1)| \leq r(S_1)\},$$

$$\sigma(S_1 - (n\pi/p)^2H_1) \subseteq \bigcup_{i=1}^{m} G_i, \quad G_i = \{z \in \mathbb{C}; |z + (n\pi/p)^2\lambda_i(H_1)| \leq r(S_1)\},$$
where $\sigma(H_1) = \{\lambda_i(H_1); 1 \leq i \leq m\}$.

Now if we apply Theorem 1 with $M = B - (n\pi/p)^2A$, $N_1 = S_1 - (n\pi/p)^2H_1$ and take into account the inequality (2.1), we obtain

$$\text{Re } z \leq z_{\text{max}}(S_1 - (n\pi/p)^2H_1), \quad z \in \sigma(B - (n\pi/p)^2A), \quad n \geq 1.$$ 

Hence the inequality (2.3) is a consequence of the definition of the sets $G_i$ given by (2.4).

Let us seek a sequence of nonzero vector solutions of the boundary value problem (1.1)–(1.2) of the form

(2.5) \hspace{1cm} u(x, t) = T(t)X(x), \quad T(t) \in C_{m \times m}, \quad X(x) \in C^m.

Note that if $X$ is an eigenfunction of the Sturm-Liouville problem

(2.6) \hspace{1cm} X''(x) - \beta X(x) = 0, \quad X(0) = X(p) = 0,

$T$ is a solution of the matrix differential equation

(2.7) \hspace{1cm} T'(t) = (B + \beta A)T(t),

then $u(x, t)$ defined by (2.5) satisfies

$$u_t(x, t) - Au_{xx}(x, t) - Bu(x, t) = [T'(t) - B + \beta A]T(t)]X(x) = 0;$$

$$u(0, t) = u(p, t) = 0.$$

Solving the Sturm-Liouville problem (2.6), one obtains a sequence of eigenvalues $\beta_n = -(n\pi/p)^2$ and a sequence of eigenfunctions

(2.8) \hspace{1cm} X_n(x) = \sin(n\pi x/p)d_n, \quad d_n \in C^m, \quad n \geq 1, \quad d_n \neq 0.

Solving (2.7) for the values $\beta_n = -(n\pi/p)^2$ one gets $T_n(t) = \exp(t(B - (n\pi/p)^2A))$ and from (2.5) we obtain a sequence of solutions of (1.1)–(1.2) of the type

(2.9) \hspace{1cm} u_n(x, t) = T_n(t)X_n(x) = e^{t(B-(n\pi/p)^2A))} \sin(n\pi x/p)d_n, \quad d_n \in C^m, \quad n \geq 1.

By superposition we propose a series solution of the problem (1.1)–(1.3) of the form

(2.10) \hspace{1cm} u(x, t) = \sum_{n \geq 1} u_n(x, t) = \sum_{n \geq 1} e^{t(B-(n\pi/p)^2A))} \sin(n\pi x/p)d_n.
Let us suppose that \( F = (f_1, \ldots, f_m)^T \) is a function such that

\[
F(0) = F(p) = 0,
\]

and that each of its components \( f_i, 1 \leq i \leq m \), is continuous in \([0, p]\) and satisfies some of the following conditions:

(2.11) \( f_i(x) \) is locally of bounded variation in a neighbourhood of \( x \), [6, p. 57],

(2.12) \( f_i(x) \) admits one-side derivatives \( (f_i)_R(x) \) and \( (f_i)_L(x) \), see Corollary 1 of [6, p. 57].

Then, by virtue of the Reimann–Lebesgue theorem there exists a positive constant \( M \) such that

\[
c_n = (2/p) \int_0^p F(x) \sin(n\pi x/p) \, dx, \quad \|c_n\| \leq M, n \geq 1,
\]

the Fourier sine series of \( F(x) \) is continuous at \( x = 0 \) and \( x = p \) and converges to \( F(x) \) for every \( x \) in \([0, p]\). In particular, \( u(x, 0) = F(x) \).

To prove that the series (2.10) is a solution of the problem (1.1)–(1.3), we show that (2.10) is convergent and admits twice termwise partial differentiation with respect to \( x \) and once with respect to the variable \( t \). We use a local argument. Let \( t_0 \) and \( t_1 \) be positive numbers with \( t_1 > t_0 > 0 \) and let us consider the rectangle

\[
R(t_0, t_1) = [0, p] \times [t_0, t_1].
\]

Note that taking termwise partial differentiation in (2.10) with \( d_n = c_n \), one gets

(2.16) \( u_{xx}(x, t) = -\sum_{n \geq 1} \frac{(n\pi/p)^2 e^{t(B-(n\pi/p)^2 A)}}{\sin(n\pi x/p)} c_n, \)

(2.17) \( u_t(x, t) = \sum_{n \geq 1} (B - (n\pi/p)^2 A) e^{t(B-(n\pi/p)^2 A)} \frac{\sin(n\pi x/p)}{c_n}. \)

Now, let us consider the Schur decomposition of the matrix \( B - (n\pi/p)^2 A \),

\[
Q_n H (B - (n\pi/p)^2 A) Q_n = D_n + N_n,
\]

with

\[
\|N_n\| \leq \|N_n\|_F \leq \|B\|_F + (n\pi/p)^2 \|A\|_F = \varrho + n^2 \theta; \quad \varrho = \|B\|_F, \quad \theta = (\pi/p)^2 \|A\|_F.
\]

395
Let $\alpha_n(A, B)$ be the real number defined by

$$\alpha_n(A, B) = \max \{ \Re z; z \in \sigma(B - (n\pi/p)^2 A) \}, \quad n \geq 1.$$  

Under the hypothesis (1.4), Theorem 3 implies that

$$\alpha_n(A, B) \leq r\left( \frac{1}{2}(B + B^H) \right) - (n\pi/p)^2 \gamma(A), \quad n \geq 1.$$  

From (1.5), (1.6) and (2.18), if $(x, t) \in R(t_0, t_1)$, it follows that

$$\|e^{t(B-(n\pi/p)^2 A))}\| \leq e^{t\left(r(\frac{1}{2}(B+B^H)) - (n\pi/p)^2 \gamma(A)\right)} \sum_{k=0}^{m-1} \frac{(t_1)^k}{k!} \|N_n\|^k \leq e^{t_1\left(r(\frac{1}{2}(B+B^H))\right)} e^{-(n\pi/p)^2 t_0 \gamma(A)} \sum_{k=0}^{m-1} \frac{(t_1)^k}{k!} (q + n^2 \theta)^k.$$  

Since for $0 \leq k \leq m - 1$ each of the numerical series

$$S_k = \sum_{n \geq 1} e^{-(n\pi/p)^2 t_0 \gamma(A)} (q + n^2 \theta)^k n^{2s}, \quad s = 0, 1,$$

is convergent, the derivation theorem for functional series [1, Theorem 9.14] and the expressions (2.10), (2.14), (2.16), (2.17) and (2.21) imply that the series

$$u(x, t) = \sum_{n \geq 1} e^{t(B-(n\pi/p)^2 A))} \sin(n\pi x/p)c_n,$$

where $c_n$ is defined by (2.14), is a well defined solution of problem (1.1)–(1.3). Note that the local argument is applicable to any positive number $t_0$, thus the following result has been established.

**Theorem 4.** Let $F(x)$ be a continuous function in $[0, p]$ satisfying (2.11) and such that each of its components satisfies some of the conditions (2.12) or (2.13). Let $B$ be any matrix in $C_{m \times m}$ and let $A$ be a matrix in $C_{m \times m}$ satisfying the spectral condition (1.4). Then, if $c_n$ is defined by (2.14), a solution of problem (1.1)–(1.3) is given by (2.22).

**Remark 1.** Note that from the condition (1.4) and the inequality (2.1) of Theorem 1, one gets that for any eigenvalue of the matrix $A$, its real part is positive. Thus Theorem 1 of [11] is a direct consequence of Theorem 4. The condition (2.11) is required to guarantee that at $x = 0$ and $x = p$, the series (2.22) converges to $F(x)$.  

396
From the practical points of view the series solution \( u(x, t) \) provided by Theorem 4 has two drawbacks. First of all \( u(t, x) \) is an infinite series, and secondly, its general term involves computation of matrix exponentials which is not an easy task \([17]\). To avoid these difficulties, in the next section we truncate the infinite series and replace the matrix exponentials by truncations of their Taylor series expansions.

3. Analytic-numerical solutions of a prescribed accuracy

In this section we are interested in the following question: Given \( t_0, t_1 \) such that \( 0 < t_0 < t_1 \), how to construct an approximate solution of the problem (1.1)–(1.3), whose error is smaller than \( \varepsilon \) uniformly for \( (x, t) \in R(t_0, t_1) = [0, p] \times [t_0, t_1] \). Let \( \varphi \) and \( \theta \) be defined by (2.18) and let \( \psi \) be the real number defined by

\[
\psi = t_0(\pi/p)^2 \gamma(A).
\]

Let us consider the scalar function \( g_k(t) \) defined by

\[
g_k(t) = k \ln(e + t^2 \theta) - \psi t^2, \quad 0 \leq k \leq m - 1, \quad t > 0,
\]
and note that \( g'_k(t) = 2t[\theta k(e + t^2 \theta)^{-1} - \psi] < 0 \) and \( 1 + g'_k(t) < 0 \), if \( t \) satisfies

\[
\frac{\theta k}{e + t^2 \theta} + \frac{1}{2t} < \psi.
\]

The general term of the numerical series

\[
\sum_{n \geq 1} (e + n^2 \theta)^k e^{-(n\pi/p)^2 \gamma(A)t_0}
\]
may be written as \( \exp(g_k(n)) \). We are interested in the determination of an integer \( n_* \) such that

\[
g_k(n) < -n \quad \text{for} \quad 0 \leq k \leq m - 1, \quad n \geq n_*.
\]

The function \( g_k(t) \) may be written in the form

\[
g_k(t) = k \ln e + k \ln(1 + t^2 \theta e^{-1}) - t^2 \psi.
\]

Since \( \lim_{t \to \infty} t^{-2} \ln(1 + t^2 \theta e^{-1}) = 0 \), we can choose a real number \( t_k \) such that

\[
2k \ln(1 + t^2 \theta e^{-1}) < \psi t^2 \quad \text{and} \quad \frac{1}{2} t(-2 + t \psi) > k \ln e
\]

for \( t \geq t_k, \quad 0 \leq k \leq m - 1 \).
If \( n_* \) is an integer such that

\[(3.8) \quad n_* \geq t_k \quad \text{and} \quad \frac{\theta k}{\varrho + \theta n^2} + \frac{1}{2n^2} < \psi, \quad \text{for} \quad 0 \leq k \leq m - 1,
\]

then, by virtue of (3.3), (3.6), (3.7) and (3.8) the sequences \( \{g_k(n)\} \) and \( \{g_k(n) + n\} \) are decreasing and negative, i.e.,

\[(3.9) \quad g_k(n + 1) < g_k(n) \quad \text{and} \quad g_k(n + 1) + n + 1 < g_k(n) + n < 0,
\]

\[n \geq n_*, \quad 0 \leq k \leq m - 1,
\]

and it follows that

\[(3.10) \quad \sum_{n \geq n_*} \exp(g_k(n)) \leq \sum_{n \geq n_*} \exp(-n) = \frac{e^{-n_*}}{1 - e^{-1}}.
\]

From (2.14), (2.21) and (3.10), if \((x,t) \in R(t_0, t_1)\), it follows that

\[(3.11) \quad \sum_{n \geq n_*} \|e^{t(B-(n\pi/p)^2)A} \sin(n\pi x/p)c_n\| \leq M(1 - e^{-1})^{-1}e^{-n_*} e^{t_1r(\frac{1}{2}(B+B^*))} \sum_{k=0}^{m-1} \frac{(t_1)^k}{k!}.
\]

Let \( \varphi \) be the real number defined by

\[(3.12) \quad \varphi = M(1 - e^{-1})^{-1} e^{t_1r(\frac{1}{2}(B+B^*))} \sum_{k=0}^{m-1} \frac{(t_1)^k}{k!},
\]

and let \( n_{**} \geq n_* \) be such that

\[(3.13) \quad n_{**} > \ln(2\varphi/\varepsilon).
\]

Then, by (3.11)–(3.13), if \((x,t) \in R(t_0, t_1)\), it follows that

\[(3.14) \quad \sum_{n \geq n_{**}} \|e^{t(B-(n\pi/p)^2)A} \sin(n\pi x/p)c_n\| \leq \varepsilon/2.
\]

From Theorem 11.2.4 of [8, p. 390], if \((x,t) \in R(t_0, t_1)\) and \(q \) is a positive integer, it follows that

\[(3.15) \quad \|e^{t(B-(n\pi/p)^2)A} - \sum_{j=0}^{q} t^j (B - (n\pi)^2 A)^j / j! \| \leq \frac{m}{(q + 1)!} \max_{0 \leq s \leq t_1} \|e^{s(B-(n\pi/p)^2)A}\|.
\]
Under the hypothesis (1.4), from (1.5), (1.6) and (2.18) we have

\[
\max_{0 \leq s \leq t_1} \| e^{s(B-(n\pi/p)^2A)} \| \leq e^{t_1r(\frac{1}{2}(B+B^H))} \sum_{k=0}^{m-1} \frac{(t_1)^k}{k!} (\varrho + n^2\theta)^k/k!.
\]

Let $\Delta$ be the positive number defined by

\[
\Delta = mM(n_* - 1)e^{t_1r(\frac{1}{2}(B+B^H))} \sum_{k=0}^{m-1} \frac{(t_1)^k}{k!} (\varrho + (n_* - 1)^2\theta)^k.
\]

Then, by (3.15)–(3.17), if we denote by $W_{n_*,q}(x,t)$ the function

\[
W_{n_*,q}(x,t) = \sum_{n=1}^{n_*-1} \sum_{j=0}^{q} t^j (B - (n\pi)^2A)^j \sin(n\pi x/p)c_n/j!
\]

and take the first positive integer $q$ satisfying

\[
(q + 1)! \geq 2\Delta/\varepsilon,
\]

then, from (3.15)–(3.19), for $(x,t) \in R(t_0,t_1)$ it follows that the difference between the exact solution $u(x,t)$ of problem (1.1)–(1.3) and the approximation $W_{n_*,q}(x,t)$, satisfies

\[
\|u(x,t) - W_{n_*,q}(x,t)\|
\leq \left\| \sum_{n \geq n_*} e^{t(B-(n\pi/p)^2A)} \sin(n\pi x/p)c_n \right\|
+ \left\| \sum_{n=1}^{n_*-1} \left\{ e^{t(B-(n\pi/p)^2A)} - \sum_{j=0}^{q} t^j (B - (n\pi/p)^2A)^{j/j!} \right\} \sin(n\pi x/p)c_n \right\| \leq \varepsilon.
\]

By the above comments the following result has been established:

**Theorem 5.** Let us consider the hypotheses and the notation of Theorem 3, let $\varepsilon > 0$, and let $R(t_0,t_1) = [0,p] \times [t_0,t_1]$ with $t_1 > t_0 > 0$. Let $\psi$ and $\varphi$ be real numbers defined by (3.1) and (3.12), respectively, and let $n_*$ be the positive integer defined by (3.13). If $q$ is the first positive integer satisfying (3.19) with $\Delta$ given by (3.17), then $W_{n_*,q}(x,t)$ defined by (3.18) is an approximate solution of problem (1.1)–(1.3), whose error with respect to the exact solution $u(x,t)$ given in Theorem 3, is uniformly upper bounded by $\varepsilon$ in $R(t_0,t_1)$. 

399
References


Authors' address: Lucas Jódar, Enrique Navarro, M.V. Ferrer, Department of Applied Mathematics, Universidad Politécnica, Plaza Mestre Ripoll 10, Valencia, Spain 46022.