Time-accurate solution of stabilized convection-diffusion-reaction equations: II. Accuracy analysis and examples

Antonio Huerta¹,* Bernardino Roig² and Jean Donea³

¹ Departament de Matemàtica Aplicada III, E.T.S. de Ingenieros de Caminos, Universitat Politècnica de Catalunya, Jordi Girona 1, E-08034 Barcelona, Spain.
² Departament de Matemàtica Aplicada, Escola Politècnica Superior de Enginyeria, Universitat Politècnica de València, Carretera Natural-Oliva s/n, E-46071 Oliva, Spain.
³ LTAS, Thermo-Mécanique, Université de Liège, Campus du Sart-Tilman, Bât. 52, B-4000 Liège, Belgium.

SUMMARY

The paper addresses the development of time accurate methods for solving transient convection-diffusion-reaction problems using finite elements. The accuracy characteristics of the spatially stabilized implicit multi-stage time stepping schemes developed in a companion paper (Part I of this work) are analyzed and compared here. This is done by means of a Fourier analysis. An important improvement is observed when the order of the method is increased. Moreover, the stabilization techniques proposed (SUPG, GLS, SGS and least-squares) do not degrade the phase accuracy. Finally, some examples are presented to show the applicability of these schemes.

KEY WORDS: convection-diffusion-reaction, time stepping schemes, stabilization, least-squares, finite element method

1. INTRODUCTION

In a companion paper [9] implicit multi-stage time stepping schemes of high accuracy are combined with spatial stabilization techniques. A general presentation of the time marching schemes can be found in [7, 8]. The stabilization techniques proposed are well-known, see for instance [2, 14, 3]. In truly transient convection-diffusion-reaction problems, which are of interest here, and by contrast with the steady state case, the truncation error in the discretization of the linear one-dimensional equation cannot be expressed in the form of a diffusion operator. In transient problems, the overall truncation error depends upon the particular time-stepping method used in combination with finite elements [12, 4, 6, 5] and it generally involves both even and odd spatial derivatives of the unknown, thus simultaneously affecting the dissipative and the dispersive properties of the numerical schemes. Here a Fourier...
(von Neumann) analysis is performed in order to study the accuracy of the proposed schemes, see [1, 11, 16]. In particular, the influence of the stabilization schemes on the phase and damping responses is studied. The proposed time stepping schemes are based on diagonal Padé approximations [8], which are A-stable. They are combined with spatial stabilization techniques and numerical stability is guaranteed. This is not the case for the corresponding implicit Runge-Kutta methods of the Lobatto IIIA family. In the final section three examples are presented to validate the proposed methods.

2. STABILIZED IMPLICIT MULTI-STAGE TIME STEPPING SCHEMES

2.1. The partial differential equation

We are concerned with a time accurate finite element solution of the following convection-diffusion-reaction equation

\[ u_t + \mathcal{L}(u) = s, \quad \text{with } \mathcal{L} := a \cdot \nabla - \nu \nabla^2 + \sigma. \tag{1} \]

The coefficients in \( \mathcal{L} \) are assumed constant for the accuracy analysis.

2.2. The numerical schemes

As proposed in [9] the consistently stabilized weak form that must be solved is

\[ (v, \frac{\Delta u}{\Delta t})_\Omega - (v, W \Delta u^n)_\Omega - \sum_c (\mathcal{R}(\Delta u), \mathcal{R}(\Delta u))_{\Omega,c} = (v, W s^n)_\Omega, \tag{2} \]

where \( \nu \in [H^1_0]^{\text{Div}} \) and \( \mathcal{R}(\Delta u) \) is a residual, which characterizes the time marching scheme,

\[ \mathcal{R}(\Delta u) = \frac{\Delta u}{\Delta t} + W \mathcal{L}(\Delta u) - W [s^n - \mathcal{L}(u^n)] - W \Delta s. \tag{3} \]

Different second order Padé approximations and fourth order Padé approximations. For the former, see [11] and for the latter, see [8]. Here, results are shown only for the following:

**Second order Padé approximation: \( R_{11} \) (Crank-Nicolson)**

\[ \Delta u = u^{n+1} - u^n, \quad \Delta s = s^{n+1} - s^n, \]

\[ W = \frac{1}{2}, \quad w = 1. \tag{4} \]

Note that in this case \( n_{\text{stages}} = 1 \) and, consequently, the vectors and matrix are scalars.

**Fourth order Padé approximation: \( R_{22} \)**

\[ \Delta u = \left\{ \begin{array}{l} u^{n+1/2} - u^n \\ u^{n+1} - u^{n+1/2} \end{array} \right\}, \quad \Delta s = \left\{ \begin{array}{l} s^{n+1/2} - s^n \\ s^{n+1} - s^{n+1/2} \end{array} \right\}, \]

\[ W = \frac{1}{24} \begin{bmatrix} 7 & -1 \\ 13 & 5 \end{bmatrix}, \quad w = \frac{1}{2} \left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}. \tag{5} \]
Note the similarity between the above two-stage scheme and the implicit Lobatto IIIA Runge-Kutta method of fourth-order accuracy. While the present scheme is based on non-overlapping step increments, the corresponding Runge-Kutta method employs total increments such as

$$\Delta u^k = \begin{cases} u^{n+1/2} - u^n \\ u^{n+1} - u^n \end{cases}.$$  

This minor difference (a linear transformation relates both schemes) does not affect linear problems with a Galerkin formulation. However, as shown in the next section, it has major consequences as regards numerical stability when implemented in connection with stabilized spatial formulations.

### 3. ACCURACY ANALYSIS

The classical von Neumann analysis, see [1, 11, 16], assumes periodic boundary conditions and studies a homogeneous linear differential equation with constant coefficients, i.e. Eq. (1) with $s = 0$. The analytical solution for one Fourier mode in $n_\text{rd}$ spatial dimensions reads

$$u(x, t) = \prod_{j=1}^{n_\text{rd}} e^{-\left((\mathbf{a} \cdot \mathbf{k})^2 + \sigma \right) t} e^{i k_j (x_j - a_j t)},$$  \hfill (6)

where $a_j$ is the $j$ component of the convective velocity $\mathbf{a} = (a_1, a_2, \ldots, a_{n_\text{rd}})^T$ and $k_j$ is the $j$ component of the wave vector $\mathbf{k} = (k_1, k_2, \ldots, k_{n_\text{rd}})^T$, which defines the generalized Fourier component

$$e^{ik \mathbf{x}} \equiv e^{i k_1 x_1} e^{i k_2 x_2} \cdots e^{i k_{n_\text{rd}} x_{n_\text{rd}}}.$$  \hfill (7)

The exact amplification factor $G_{\text{ex}}$, which relates the value of the unknown $u$ at two consecutive instants $t^n$ and $t^{n+1}$, and at each point in $\Omega$, is then determined

$$u^{n+1} = G_{\text{ex}} u^n,$$  \hfill (8)

where $G_{\text{ex}} = e^{-(\delta u + \omega_{\text{ex}}) \Delta t}$.

To determine the numerical amplification factor for each time stepping scheme, taking into account the spatial discretization, it is necessary to determine the influence of each spatial operator on a Fourier mode. Thus discrete operators associated to the continuous spatial operators in (2), must be determined. In Table I these relations are shown for uniform linear 1D elements of size $h$ and using the well known difference operators, see [16] for details,

- Second order centered first difference: $\delta u(x, t) = u(x + h, t) - u(x - h, t)$
- Second order centered second difference: $\delta^2 u(x, t) = u(x - h, t) - 2u(x, t) + u(x + h, t)$

To determine the influence of the numerical scheme on a Fourier component, see (7), one introduces the dimensionless wave vector ("number" in the 1D case), $\xi = h \mathbf{k}$. Moreover, using $\Delta t$ and $h$ the parameters in the differential equation, (1), are scaled. The Courant number is $c = ||\mathbf{c}||$ with $\mathbf{c} = a \Delta t / h$, the diffusion number is $d = (\nu \Delta t) / h^2$, and the dimensionless reaction is $r = \sigma \Delta t$. The exact amplification factor, (8), is rewritten as

$$G_{\text{ex}}(\xi, c, d, r) = e^{-(\delta_{\text{ex}} + \omega_{\text{ex}}) \Delta t} = e^{-(d \xi + r + ic \xi)}.$$  \hfill (9)
<table>
<thead>
<tr>
<th>spatial operator</th>
<th>discrete</th>
<th>mode transf.</th>
<th>expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>((N_i, N_j)/h^2)</td>
<td>(1 + \frac{1}{6}\delta^2)</td>
<td>(\mathcal{M}(\xi))</td>
<td>(1 + (\cos \xi - 1)/3)</td>
</tr>
<tr>
<td>(\Delta t (N_i, a \cdot \nabla N_j)/h^2)</td>
<td>(\frac{1}{2}c\delta)</td>
<td>(\mathcal{A}(\xi, c))</td>
<td>(ic \sin \xi)</td>
</tr>
<tr>
<td>(-\Delta t \nu (\nabla N_i, \nabla N_j)/h^2)</td>
<td>(d\delta^2)</td>
<td>(\mathcal{K}(\xi, d))</td>
<td>(2d(\cos \xi - 1))</td>
</tr>
<tr>
<td>(\Delta t (a \cdot \nabla N_i, N_j)/h^2)</td>
<td>(-\frac{1}{2}c\delta)</td>
<td>(-\mathcal{A}(\xi, c))</td>
<td>(-ic \sin \xi)</td>
</tr>
<tr>
<td>(-\Delta t^2 (a \cdot \nabla N_i, a \cdot \nabla N_j)/h^2)</td>
<td>(c^2\delta^2)</td>
<td>(\mathcal{D}(\xi, c))</td>
<td>(2c^2(\cos \xi - 1))</td>
</tr>
</tbody>
</table>

Table I. Spatial, discrete and mode transformation operators for 1D.

In Table I the discrete operators are related with the transformations for each Fourier mode and the expressions are given for linear finite elements.

For example, to determine the numerical amplification factor for the proposed time schemes and a Galerkin formulation, corresponding to \(\tau = 0\) in (2), the following steps are necessary:

1. From (2) find the equation that modifies each Fourier component using Table I,

\[
\{\mathcal{M}(\xi)I - [\mathcal{K}(\xi, d) - \mathcal{A}(\xi, c) - r\mathcal{M}(\xi)]W\}u^n = [\mathcal{K}(\xi, d) - \mathcal{A}(\xi, c) - r\mathcal{M}(\xi)]u^n w.
\]

2. Solve the linear system of equations for \(z = (z_1, z_2, \ldots, z_{\text{range}})^T\):

\[
\{\mathcal{M}(\xi)I - [\mathcal{K}(\xi, d) - \mathcal{A}(\xi, c) - r\mathcal{M}(\xi)]W\}z = [\mathcal{K}(\xi, d) - \mathcal{A}(\xi, c) - r\mathcal{M}(\xi)]w.
\]

3. And finally determine the amplification factor as

\[
G_\tau(\xi, c, d, r) = 1 + z_1 + z_2 + \cdots + z_{\text{range}}.
\]

The same steps are repeated for the stabilized schemes. Recall that the operator \(P\) is different in each case, see [9]. The equations associated to the stabilization techniques for a typical Fourier mode corresponding to a wave vector \(\xi\) are as follows (note the nonlinear dependence on \(W\)).

**Streamline Upwind Petrov-Galerkin**:

\[
\mathcal{P}(v) = \frac{dv}{dt} + W \mathcal{L}(v),
\]

\[
\{M(\xi)I - [K(\xi, d) - A(\xi, c) - rM(\xi)]W = -\frac{T}{\Delta t} \left[ (D(\xi, c) + rA(\xi, c))W^T W + A(\xi, c)W^T \right] u^n \text{ w.}
\]

**Galerkin Least-Squares**:

\[
\{M(\xi)I - [K(\xi, d) - A(\xi, c) - rM(\xi)]W = -\frac{T}{\Delta t} \left[ M(\xi)I + A(\xi, c)(W^TW - 2W + 2rM(\xi)W^TW) + (D(\xi, c) - r^2M(\xi))W^TW \right] u^n \text{ w.}
\]
Least-squares: \( \tau \mathcal{P}(v) = \Delta t \mathbf{W} \mathcal{L}(v) \),

\[
\begin{align*}
\{ & M(\xi) 1 - [K(\xi,d) - A(\xi,c) - rM(\xi)] \mathbf{W} \\
& - [A(\xi,c) - rM(\xi)]W^T + (D(\xi,c) - r^2 M(\xi))W^T \mathbf{W} \} \Delta u \\
& = \left\{ [K(\xi,d) - A(\xi,c) - rM(\xi)] + [(D(\xi,c) - r^2 M(\xi))W^T] \right\} u^n \mathbf{w}. \quad (12)
\end{align*}
\]

The fourth order (in space and for 1D) accurate formula for the stabilization parameter proposed in [15] is used \( \tau = \left[ \left( \frac{2 a}{h} \right)^2 + 9 \left( \frac{4 \sigma}{h} \right)^2 + \sigma^2 \right]^{-1/2} \). Figures 1 and 2 show the relative errors for the numerical damping, \( |G(\xi,c,d,r)|/|G_{ex}(\xi,c,d,r)| \), and the phase error,
arg(G(ξ, c, d, r)) / arg(G_{ex}(ξ, c, d, r)) for different values of the Péclet number, Pe = c/(2d). All the curves are plotted in polar coordinates, the angle is ξ and the radius is the relative error. From a precision point of view only the range 0 ≤ ξ ≤ π/4 is of interest. Thus, the limit of this sector is also depicted. Note that for Crank-Nicolson, Figure 1, the Courant numbers employed are 0.75, 1.5 and 3. For R_{22} large values of the Courant number (1.5, 3 and 6) are used to highlight its range of precision which must compensate its extra cost (double number of unknowns per node).

Only the GLS stabilization results are shown, the other stabilization techniques present little differences. The least-squares (LS) formulation introduces, as expected, more numerical diffusion for \( R_{22} \) but it is almost negligible for Courant numbers up to 3. Moreover, it is only noticeable in pure convection (Pe = 0) and this effect is beneficial in the presence of reaction. The phase accuracy is also affected by the LS stabilization but this only occurs for low Péclet numbers.

One can observe that, as expected, the stabilized methods introduce numerical damping for high frequencies (ξ close to π), which cannot be properly represented on the discrete mesh. Recall that Galerkin methods for pure convection do not introduce any numerical damping when combined with diagonal Padé approximations. Moreover, from a precision point of view (i.e. for dimensionless wave numbers such that 0 ≤ ξ ≤ π/4) the response (for different Péclet, Courant and reaction values) is quantitatively the same as in the Galerkin formulation. Thus, the stabilization introduced as proposed here puts the numerical diffusion where it is needed and does not compromise the accuracy.

Moreover, from the point of view of the phase errors, the stabilized methods show a behavior identical to the Galerkin formulation. That is, phase errors are drastically reduced when the order of the temporal approximation is increased.

The reaction only influences the precision of the numerical damping, only for very large values of the reaction the phase precision is moderately decreased as r increases.

Results are only shown for Crank-Nicolson and \( R_{22} \). Similar conclusions can be drawn for the sixth-order method \( R_{33} \), see [6]. That is, for SUFG, GLS and SGS numerical damping is only introduced for high frequencies and the phase accuracy is improved for Courant numbers up to 6. The LS stabilization, however, is over diffusive and phase errors become more noticeable. Moreover, such a formulation requires three times more unknowns than Crank-Nicolson.

An important point to be made in concluding the accuracy analysis of the proposed schemes is that the use of non-overlapping increments in the multi-stage time scheme preserves unconditional stability when the method is combined with stabilized spatial formulations. By contrast, Figure 3 shows that the classical fourth-order accurate Lobatto IIIA implicit Runge-Kutta method combined with a LS stabilization loses stability for moderate to low Péclet numbers. It also shows that the sixth-order accurate Lobatto IIIA implicit Runge-Kutta method loses stability in pure convection when combined with a GLS stabilization. This is not the case for the corresponding multi-stage schemes derived from Padé approximations.
4. NUMERICAL EXAMPLES

4.1. Nonlinear propagation of a step

In order to compare the response of \( R_{11} \) (Crank–Nicolson) and \( R_{22} \) with stabilized formulations, the first test solves Burgers equation

\[
u_t + uu_x = v u_{xx}
\]

on the domain \([0, 1]\) with the Dirichlet condition \( u(0, t) = 1 \) and a homogeneous Neumann condition of zero total flux at \( x = 1 \). The analytical solution given by Sardany [13] is

\[u(x, t) = \left(1 + \exp\left(\frac{x - 0.5t}{\nu}\right)\right)^{-1}\]

The initial condition is imposed at \( t = 0.02 \), and a (maximum) Péclet number of 1000 is considered. A non-uniform mesh of 1000 linear elements is used. The mesh size is \( h = 0.002 \) on \([0, 0.2] \cup [0.4, 1]\) and \( h/10 \) on \((0.2, 0.4)\). Figure 4, where the logarithm of the absolute error is depicted at time \( t = 1 \), shows the influence of the non-uniform mesh. The response obtained with Crank–Nicolson (\( c = 0.75 \)) is clearly worse than \( R_{22} (c = 3) \) for both formulations: GLS and LS. Since the spatial mesh size is very small, this is clearly due to the inferior phase accuracy of the second-order Crank–Nicolson method. Note that \( R_{22} \) runs with a time step four times larger than \( R_{11} \).

4.2. Rotating pulse

This example, see [10], illustrates the applicability of the previously analyzed schemes to problems with internal layers and with a steady state solution. Equation (4) is solved in \( \Omega = (-1, 1) \times (-1, 1) \) with homogeneous Dirichlet boundary conditions, and the following
The asymptotic steady state solution presents a clear pattern with $u \approx 1/2$ if $\rho \leq 1/2$ and $u \approx 0$ otherwise, the internal layer is $O(\sqrt{\rho})$ along the circle $\rho = 1/2$. A uniform mesh of $40 \times 40$ bilinear elements is used. Figure 5 shows the results for a Péclet number of 50000 and a least-squares stabilization. Results for other stabilized methods or for higher-order time stepping schemes ($R_{s3}$) present negligible differences.
4.3. Two-dimensional Burgers equation

The Burgers equation in 2D over the square domain \( \Omega = (-1,1) \times (-1,1) \) is solved here with a uniform mesh of 30 \times 30 bilinear elements. The initial and boundary conditions are those already used in [8] where the analytical solution is also shown. The Galerkin formulation used in [8] needed low Péclet numbers, \( \text{Pe} = 3.33 \). The stabilized methods studied here allow to obtain reasonable solutions for Péclet numbers as large as 5000. When convection dominates the nonlinear transport, the solution includes the formation of a sharp internal layer along the diagonal of the domain passing through the points \((0,1)\) and \((1,0)\), see Figure 6. The Galerkin solution is not shown because spurious oscillations cover the whole domain.

5. CONCLUSIONS

Time accurate and spatially stable schemes have been proposed for the finite element solution of unsteady convection-diffusion-reaction problems. The accuracy analysis of the schemes has shown that the use of high-order implicit time stepping algorithms provides phase characteristics vastly superior to those obtained with the classical Crank-Nicolson method. On the other hand, the spatial stabilization techniques used in combination with the high-order multi-stage time marching schemes do not degrade the amplitude response in the accuracy range of the wave number. Instead, they provide damping of the high frequency modes which are badly represented on the practical finite element meshes. Moreover, the proposed time schemes based on a multi-stage factorization of Padé approximations presents superior stability properties than the corresponding Lobatto IIIA implicit Runge-Kutta methods when combined with spatial stabilization techniques.
REFERENCES