PUBLICATIONS MATHÉMATIQUES ET INFORMATIQUES DE RENNES

NABIL R. NASSIF

# Pade-Galerkin Methods for Parabolic Partial Differential Equations

*Publications des séminaires de mathématiques et informatique de Rennes*, 1975, fascicule S3
« Journées « éléments finis » », , p. 1-27

<http://www.numdam.org/item?id=PSMIR_1975___S3_A4_0>

© Département de mathématiques et informatique, université de Rennes, 1975, tous droits réservés.

L'accès aux archives de la série « Publications mathématiques et informatiques de Rennes » implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

*Article numérisé dans le cadre du programme*
*Numérisation de documents anciens mathématiques*
http://www.numdam.org/

# PADE-GALERKIN METHODS FOR PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS

Nabil R. NASSIF

Mathematics Department, American University of Beirut, Lebanon, and
Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne, Suisse.

## ABSTRACT

The variational formulation of Padé methods for solving linear parabolic equations is studied. Optimal error estimates are derived. The efficiency of a time discretization with respect to the initial condition is then considered. A starting index is defined which shows considerable advantage in using underdiagonal Padé discretizations. Some practical aspects of the methods are also presented.

## I. INTRODUCTION

In [1], Descloux and Nassif have derived a class of one step methods for the numerical integration of the initial-value problem

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

The principal feature of this class of methods is that when applied to the linear case $f(t,y) = \lambda y$ , it reduces to the family of Padé approximations to $\exp(\lambda t)$. Precisely let

$$c_k \equiv c_k(\ell,r) = \frac{(\ell+r-k)!\ (\ell)!}{(\ell+r)!\ (k)!\ (\ell-k)!} , \quad 0 \leqq k \leqq \ell ,$$

and $d_k = c_k(r,\ell)$ . Then it is proved [ibid] , that if $y \in C^{\ell+r+1}$ ,

$$\text{(1)} \qquad \sum_{k=0}^{r} (-\tau)^k\, c_k\, y^{(k)}(t+\tau) - \sum_{k=0}^{\ell} (\tau)^k\, d_k\, y^{(k)} = \int_t^{t+\tau} (s-t)^{\ell}(t+\tau-s)^{r}\, y\,[t^{<\ell>},(t+\tau)^{<r>},s]\, ds ,$$

where $y[t^{<\ell>},(t+\tau)^{<r>},s]$ is the usual notation for the divided difference as available in [2]. In particular when $f(t,y) = \lambda y$ , one obtains from (1)

$$z(t+\tau) = R_{\ell,r}(\lambda\tau)\, z(t)$$

where

$$R_{\ell,r}(x) = \sum_{k=0}^{\ell} d_k\, x^k \Big/ \sum_{k=0}^{r} c_k\, (-x)^k ,$$

is the $(\ell,r)$ entry of the Padé table, and $z(t)$ the approximating solution to $y(t)$ .

We shall systematically apply (1) to obtain fully discretized schemes for the evolution parabolic problem. These schemes will be semi-variational, in the sense that the space discretization is of the Galerkin (or finite element) type.

We note that these schemes can be also derived by the usual Galerkin semi-discretization followed by the approximation of the system of ordinary differential equations that is obtained. See for example [2] and [3]. This new approach is however more natural and will directly yield the necessary error estimates. At first, we shall introduce the problem together with few notations. Let $V$ and $H$ be two Hilbert spaces such that $V \subset H$. We shall respectively denote by $\|\cdot\|$ and $|||\cdot|||$, the norms on $H$ and $V$ and we assume as it is currently done (for example in [4]) that the identity mapping from $V$ into $H$ is bounded i.e. that $\|v\| \leqq |||v|||$, $\forall\, v \in V$ ; $(\cdot,\cdot)$ and $((\cdot,\cdot))$ are respectively the inner products on $H$ and $V$.

$C^k([0,T]\ ;\ X)$, with $X = V$ or $H$ is the space of $k$ continuously differentiable functions from $[0,T]$ into $X$.

We are interested in obtaining $u(t) \in V$ such that

$$(2) \qquad \begin{cases} u'(t) + Lu(t) = f(t)\ , \\ u(0) = g\ , \end{cases}$$

where $f(t)$ and $g$ are given in $H$ and $L$ is a continuous linear operator from $V$ into $H$, <u>independent</u> of $t$ satisfying a strong $V$-ellipticity property, namely that

$$a(v,v) = (Lv,v) \geqq \gamma_0 |||v|||^2\ ,\ \forall\, v \in V \quad \gamma_0 > 0$$

Existence, uniqueness and regularity of the solution to (2) has been discussed by several authors. See for example [4]. Note then that (2) can be written in the equivalent semi-variational form,

$$(2') \qquad \begin{cases} (u'(t),v) + a(u(t),v) = (f(t),v),\ t > 0 \\ (u(0),v) = (g,v),\ \forall\, v \in V \end{cases}$$

Although the formulation is weak the solution $u(t)$ to (2) - (2') will be assumed sufficiently regular. In fact if $u \in C^{\ell+r+1}([0,T]\ ;\ V]$, then using (1), we may write

$$(3)\qquad \sum_{k=0}^{r}(-\tau)^k c_k u^{(k)}(t+\tau) - \sum_{k=0}^{\ell}\tau^k d_k u^{(k)}(t) = \int_t^{t+\tau}(s-t)^{\ell}(t+\tau-s)^r u[t^{<\ell>},(t+\tau)^{<r>},s]ds$$

Simultaneously, these regularity assumptions on $u$ and henceforth on $f$ allow us to write

$$(4)\qquad (u^{(k)}(t),v) + a(u^{(k-1)}(t),v) = (f^{(k+1)}(t),v),\ \forall\, v \in V,\ t\geq 0,\ 1\leq k\leq \max(\ell,r)$$

Now let $M = M_h \subset V$, be a finite-dimensional subspace of $V$, with a "good" approximation properties for elements of $H$ and $V$, as $h \to 0$. Let us also define a partition of the interval $[0,T]$,

$$(5)\qquad 0 = t_0 < t_1 < \ldots < t_n = T\ ,$$

with $t_{j+1} - t_j = \tau_j$. We shall write $u_j^k \equiv u^{(k)}(t_j)$, and $f_j^k = f^{(k)}(t_j)$. Define then the <u>Padé-Galerkin method</u> as follows. We construct, $U_j^k \in M$, $(U_j = U_j^0)$, $0\leq k\leq p$, $0\leq j\leq n$, such that

$$(6)\qquad \begin{cases} \sum_{k=0}^{r}(-\tau_j)^k c_k\,(U_{j+1}^k,v) - \sum_{k=0}^{\ell}\tau_j^k d_k\,(U_j^k,v) = 0\ ,\ 0\leq j\ ,\ \forall\, v\in M \\ (U^0,v) = (g,v)\ ,\quad \forall\, v\in M \\ (U_j^{k+1},v) + a(U_j^k,v) = (f_j^k,v)\quad \forall\, v\in M\ ,\ j\geq 0\ . \end{cases}$$

Of course this scheme can be defined without the introduction of the derivatives of $f$. We may clearly replace $f_j^k$ by some difference approximation that is consistent with the order of accuracy of (1).

Let us now give examples of this scheme, for the particular cases that have been so far used :

(i) $\ell = 1$, $r = 0$. The application of (6) gives

$$(U_{j+1},v) - (U_j,v) - \tau_j(U_j',v) = 0$$
$$(U^0,v) = (g,v)\ ,$$
$$(U_j^1,v) + a(U_j,v) = (f_j,v)\ ,$$

which reduces to

$$\left(\frac{U_{j+1} - U_j}{\tau_j}, v\right) + a(U_j, v) = (f_j, v) \quad \forall\, v \in M, \; j \geq 0$$

$$(U_0, v) = (g, V) \quad \forall\, v \in M .$$

Which is the classical Euler-Forward scheme. Similarly

(ii) $\ell = 0$, $r = 1$, gives the fully-implicit or Euler Backward scheme. Precisely,

$$\left(\frac{U_{j+1} - U_j}{\tau_j}, v\right) + a(U_{j+1}, v) = (f_{j+1}, v) \quad \forall\, v \in M, \quad j \geq 0$$

$$(U_0, v) = (g, v), \quad \forall\, v \in M .$$

(iii) $\ell = 1$, $r = 1$, is the well-known Crank-Nicolson introduced and analysed by Douglas and Dupont [5]. Similar calculation as above yields,

$$\left(\frac{U_{j+1} - U_j}{\tau_j}, v\right) + a\left(\frac{U_{j+1} + U_j}{2}, v\right) = \left(\frac{f_j + f_{j+1}}{2}, v\right) \quad \forall\, v \in M, \quad j \geq 0$$

$$(U_0, v) = (g, v) \quad \forall\, v \in M .$$

(iv) The case $\ell = r = 2$ has been used in [6] and [7]. Formula (6) gives

$$(U_{j+1}, v) - (\tau_j/2)\,(U^1_{j+1}, v) + \frac{\tau_j^2}{12}(U^2_{j+1}, v) = (U_{j-1}, v) + (\tau_j/2)\,(U^1_j, v) + \frac{\tau_j^2}{12}(U^2_j, v) \quad \forall\, v \in M$$

$$(U_0, v) = (g, v) \quad \forall\, v \in M$$

$$(U^1_j, v) + a(U_j, v) = (f_j, v), \text{ and } (U^2_j, v) + a(U^1_j, v) = (f^1_j, v), \; j \geq 0 \quad \forall\, v \in M$$

which can be rewritten in a simplified form as

$$\left(\frac{U_{j+1} - U_j}{\tau_j}, v\right) + a\left(\frac{U_{j+1} + U_j}{2}, v\right) + \tau_j\, a\left(\frac{U^1_j - U^1_{j+1}}{12}, v\right) = \left(\frac{f_j + f_{j+1}}{2}, v\right) + \tau_j\left(\frac{f^1_j - f^1_{j+1}}{12}, v\right), \; j \geq 1$$

$$(U_0,v) = (g,v) \quad \forall\, v \in M$$

$$(U_j^1,v) + a(U_j,v) = (f_j,v) \quad \forall\, v \in M\ ,\ \ j \geq 0\ .$$

(v) More recently, in [8], several advantageshave appeared from the use of (6) with $\ell = 0$, $r = 2$, and $\ell = 1$, $r = 2$. We shall directly write these schemes. They are respectively

$$(\frac{U_{j+1} - U_j}{\tau_j}, v) + a(U_{j+1} - \frac{\tau_j}{2} U_{j+1}^1, v) = (f_{j+1} - \frac{\tau_j}{2} f_{j+1}^1, v)\ ,\ \forall\, v \in M\ ,\ j \geq 0$$

and

$$(\frac{U_{j+1} - U_j}{\tau_j}, v) + a(\frac{2}{3} U_{j+1} + \frac{1}{3} U_j, v) - \frac{\tau_j}{6} a(U_{j+1}^1, v) = (\frac{2}{3} f_{j+1} + \frac{1}{3} f_j - \frac{\tau_j}{6} f_{j+1}^1, v)$$

$$\forall\, v \in M\ .$$

with $$(U_j^1,v) + a(U_j,v) = (f_j,v)\ ,\ \forall\, v \in M\ ,\ \ j \geq 0\ .$$

and $$(U_0,v) = (g,v)\ .$$

Let us now introduce the linear operator $L : M \to M$ defined by

$$(Lw,v) = a(w,v)\ ,\ \forall\, v \in M \tag{7}$$

Note then that

$$(Lv,v) \geq \gamma_0 \|v\|^2\ ,\quad \forall\, v \in M \tag{7'}$$

implying that $(\alpha+L)^{-1}$ exists for all $\alpha$, with $\mathrm{Re}(\alpha) \geq -\gamma_0$.

Let also for $\varphi \in H$, $\bar{\varphi}$ be its projection on $M$, i.e.,

$$(\varphi-\bar{\varphi},v) = 0 \quad \forall\, v \in M\ . \tag{8}$$

It is then possible to write the above schemes in the following form,

$$U_{j+1} = R_{\ell,r}(-\tau_j L)\, U_j + \tau_j \tilde{f}_j\ ,\ \ j \geq 0 \tag{9}$$

where $\tilde{f}_j$ corresponds to the discretization of the non-homogeneous term $f$. To illustrate, let us take the Crank-Nicolson and the Padé (2.2). For the first $R_{\ell,r}(z) = (2+z)/(2-z)$, and $f_j = (\bar{f}_j + \bar{f}_{j+1})/2$ and for the second $R_{\ell,r}(z) = (12+6z+z^2)/(12-6z+z^2)$ and

$$\tilde{f}_j = \frac{1}{2}(\bar{f}_j + \bar{f}_{j+1}) + \frac{j}{12}(\bar{f}_j^1 - \bar{f}_{j+1}^1 + L\bar{f}_{j+1} - L\bar{f}_j)$$

From (7'), it is clear that $(2+L)^{-1}$ exists and also $(3 \pm i\sqrt{3} + L)^{-1}$, where $3 \pm i\sqrt{3}$ are the roots of $z^2 + 6z + 12$.

As for the algorithmic implementation of (9) when $r \geqq 2$, complex factorization is necessary. We shall return to this point in the last section.

Our paper will be divided as follows. In the next section we shall derive estimates on $u_j - U_j$, in terms of those on $\tilde{U}_j - U_j$, where $\tilde{U}_j$ is the V-projection of $u_j$ on $M$, which is known to be optimal in most practical cases. We will restrict ourselves to those cases shown above, i.e. for $\ell \leqq r \leqq 2$. However, the modification of our arguments to higher order cases can be done and we shall avoid if for simplicity. We shall rely heavily on a concept of strong stability, thus avoiding the use of the A-stability test, as it is done by Crouzeix in [9]. In section 3, we consider the starting problem in parabolic equations, and we show by defining a starting index how the accuracy can be increased at the initial stage when we use Padé under diagonal schemes. They have indeed a particular feature, that seems to lack in most other schemes, in the sense that they provide a good approximation to exp(-x) both near zero and at infinity. In the last section, we consider the numerical inplementation of the Padé schemes. We shall see in particular that they are considerably efficient when we use complex arithmetic.

## 2. STABILITY AND ERROR ESTIMATES

### 2.1. Stability

From (9), we have the following easy lemma

LEMMA 1

$$U_{j+1} = \Big[\prod_{k=0}^{j} R_{\ell,r}(-\tau_k L)\Big]\, U_0 + \sum_{k=0}^{j} \tau_k \Big[\prod_{i=k+1}^{j} R_{\ell,r}(-\tau_i L)\Big]\, \tilde{f}_k$$

(with $\prod_{i=k+1}^{j} R_{\ell,r}(-\tau_i L) = 1$ , when $k = j$ ) .

DEFINITION

Let

$$g_i = \|R_{\ell,r}(-\tau_i L)\| \,,\ \tau = \max_i\{\tau_i\} \,,\ \underline{\tau} = \min_i\{\tau_i\} \,. \tag{10}$$

We shall assume the partition $\{\tau_i\}_i$ to be regular, i.e.,

$$\tau/\underline{\tau} \leqq \delta \quad \text{(constant)} \tag{11}$$

The scheme (10) is said to be stable, if there exists $\delta_1 > 0$ , $\delta_2 > 0$ , and $\mu_0$ real such that

$$\inf_{\substack{\tau\leqq\delta_1 \\ h\leqq\delta_2}} \{\min_i [(1-g_i)/\tau_i]\} \geqq \mu_0 \tag{12}$$

If

(i) $\mu_0 \leqq 0$ the scheme $U_{j+1} = R_{\ell,r}(-\tau_j L)\, U_j + \tau_j \tilde{f}_j$ , is said to be weakly stable

(ii) If $\mu_0 > 0$ , it is strongly stable.

REMARK

This definition of stability stems from the well-known Von-Neumann condition of stability and also from Kreiss definition of dissipative difference operators, [ 9 ] in the case of strong stability.

We can state the following theorems.

THEOREM 1

A necessary and sufficient condition for stability is that there exists a real number $\gamma_1$ , and positive numbers $\delta_1, \delta_2$, such that,

$$(13) \qquad \prod_{k=j}^{j+r-1} g_k \leqq \exp\{-\gamma_1 (t_{j+r} - t_j)\} \ , \ \forall\, j,r \ , \ \tau \leqq \delta_1 \ , \ h \leqq \delta_2$$

PROOF

Stability implies

$$g_i \leqq 1 - \mu_0 \tau_i \ , \ \tau \leqq \delta_1 \ , \ h \leqq \delta_2 \ ,$$

and therefore $g_i \leqq \exp(-\mu_0 \tau_i)$ , which gives immediately condition (13), with $\gamma_1 = \mu_0$ .

Conversely if we put in (13) $r = 1$ , then we have

$$g_j \leqq \exp(-\gamma_1 \tau_j) \ , \ \tau \leqq \delta_1 \ , \ h \leqq \delta_2$$

and therefore

$$1 - g_j \geqq 1 - \exp(-\gamma_1 \tau_j) \ .$$

Using the relation $\exp(-x) = 1 - x + \frac{x^2}{2} e^{-\theta x}$ , $0 < \theta < 1$ we have if $\gamma_1 > 0$ , and by taking $\tau_j \leqq \min(\delta_1, \frac{1}{\gamma_1})$

$$(1-g_j)/\tau_j \geqq \mu_0 = \gamma_1 - \frac{\gamma_1}{2} = \frac{\gamma_1}{2} > 0 \ .$$

For $\gamma_1 = 0$ , it is obvious that $g_j \leqq 1$ , for $\tau \leqq \delta_1$ , $h \leqq \delta_2$. Last, for $\gamma_1 < 0$ , taking $\tau_j \leqq \min(\delta_1, 1)$ , we find easily that

$$(1-g_j)/\tau_j \geqq \gamma_1 - [\gamma_1^2 \exp(-\gamma_1)/2] = \mu_0 \ .$$

We have next

THEOREM 2

(i) If the stability is strong then

$$\|U_j\| \leq \exp(-\mu_0 t_j)\,\|U_0\| + \frac{\delta}{\mu_0} \sup_{0\leq k\leq j-1} \|\tilde{f}_k\| \tag{14}$$

(ii) If the stability is weak then

$$\|U_j\| \leq \exp(\mu_1 t_j)\,[\,\|U_0\| + t_j \sup_{0\leq k\leq j-1} \|\tilde{f}_k\|\,] \quad (\mu_1 = -\mu_0) \tag{15}$$

PROOF

Lemma 1 implies

$$\|U_{j+1}\| \leq \Big(\prod_{k=0}^{j} g_k\Big)\,\|U_0\| + \sum_{k=0}^{j} \tau_k \Big[\prod_{i=k+1}^{j} g_i\Big]\,\|\tilde{f}_k\| \ . \tag{16}$$

First, let us suppose the stability to be strong i.e., $\mu_0 > 0$ . Then the first term in the right hand side of (16) is bounded by $\exp(-\mu_0 t_{j+1})\ \|U_0\|$ . In the second term, note that

$$\prod_{i=k+1}^{j} g_i \leq (1-\mu_0\underline{\tau})^{j-k} \ . \tag{17}$$

Therefore

$$\|U_{j+1}\| \leq \exp(-\mu_0 t_{j+1})\,\|U_0\| + \tau\,[\sum_{k=0}^{j} (1-\mu_0\underline{\tau})^{j-k}] \sup_k \|\tilde{f}_k\|$$

Clearly

$$\sum_{k=0}^{j} (1-\mu_0\underline{\tau})^{j-k} \leq \frac{1}{\mu_0\underline{\tau}}$$

and therefore

$$\|U_{j+1}\| \leq \exp(-\mu_0 t_{j+1})\,\|U_0\| + \frac{\delta}{\mu_0} \sup_k \|\tilde{f}_k\| \ ,$$

which proves the first part of the theorem.

The second part relies simply on theorem 1, since (16) implies

$$\|U_{j+1}\| \leq \left(\prod_{k=0}^{j} g_k\right) \|U_0\| + \left(\sum_{k=0}^{j} \tau_k\right) \sup_{0\leq k\leq j} \|\tilde{f}_k\| \ . \qquad \blacksquare$$

As a result of theorems 1 and 2, our task is reduced to the estimation of $g_i = \|R_{\ell,r}(\tau_i L)\|$ . As we shall see in the case when $L$ (and therefore $L$) is symmetric, one can arrive at a great deal of information about $g_i$ , for <u>all</u> the Padé table. Unfortunately when $L$ and $L$ are not symmetric, an important tool introduced by Crouzeix [8], that relies on a spectral theorem due to Von Neumann, can give in some cases the same strong results obtained for the self-adjoint problem.

We assume the following well-known facts

(A1) $$a(v,v) = (Lv,v) \geq \gamma_0 |||v|||^2 \geq \gamma_0 \|v\|^2 \ , \ \gamma_0 > 0$$

(A2) $L$ is a $2m^{th}$ order elliptic operator, i.e.

$$H_0^m(\Omega) \subset V \subset H^m(\Omega) \ , \ H = L^2(\Omega) \ , \ \Omega \subset \mathbb{R}^p \ ,$$

where $$H^m(\Omega) = \{ v \mid D^\alpha v \in L^2(\Omega) \ , \ |\alpha| \leq m \}$$

$$H_0^m(\Omega) = \{ v \in H^m(\Omega) \mid D^\alpha v = 0 \ \text{ on } \ \partial\Omega \ , \ |\alpha| \leq m-1 \}$$

with $$\|v\|_m^2 = \sum_{|\alpha|\leq m} \int_\Omega |D^\alpha v|^2 \, dx \ .$$

(A3) The largest eigenvalue of $L$ , $\lambda_N$ is $O(h^{-2m})$ , and the first $\lambda_1$ is $O(1)$ .

We can then state

<u>THEOREM 3</u>

Suppose $L$ is symmetric, i.e.,

$$a(v,w) = a(w,v) \quad \forall \ v,w \in V \ ,$$

and let $A_1$, $A_2$, $A_3$ be satisfied.

Then

(i) The Padé under-diagonal schemes ($\ell<r$) are unconditionally strongly stable.

(ii) The Padé diagonal schemes are unconditionally stable. Furthermore if $\tau = O(h^{m\alpha})$, $\alpha \geq 1$, this stability is strong.

(iii) The other Padé schemes ($\ell>r$) are stable for $\tau = O(h^{\alpha m})$, $\alpha \geq 2$. In some cases this condition entails strong stability.

PROOF

For the Padé under diagonal schemes, and for $\tau_i$ sufficiently small, we have $g_i = R_{\ell,r}(-\tau_i\lambda_1)$, where $\lambda_1$ is the first eigenvalue of $L$, with $\lambda_1 \geq \gamma_0$. Thus it can be easily proved that $g_i \leq 1$ and $1-g_i = O(\tau)$, which implies (i).

For $\ell = r$, we have $c_k = d_k$ and $R(-x) = \sum_{k=0}^{r} c_k(-x)^k \Big/ \sum_{k=0}^{r} c_k x^k$.

This clearly implies, that

$$g_i = \max \{R(-\tau_i\lambda_1), |R(-\tau_i\lambda_N)|\},$$

and

$$1-g_i = \min \{1 - R(-\tau_i\lambda_1), 1 - |R(-\tau_i\lambda_N)|\} .$$

Note then that

$$R(-x) = \sum_{k=0}^{r} c_k(-\frac{1}{x})^{-k+r} \Big/ \sum_{k=0}^{r} c_k(\frac{1}{x})^{-k+r} = Q(\frac{1}{x}) ,$$

with

$$Q(y) = \sum_{k=0}^{r} c_{r-k}(-y)^k \Big/ \sum_{k=0}^{r} c_{r-k}(y)^k .$$

It is then easily checked that $|1-Q(y)| = O(|y|)$ as $y \to 0$. We may then have three cases : (i) $\tau = O(h^{\alpha m})$, $\alpha \geq 2$, and then $|R(-\tau_i\lambda_N)| \leq \delta < 1$, $\delta$ independent of $\tau$ and $h$, and then $1-g_i = \{1-R(-\tau_i\lambda_1)\}$, for $\tau$ sufficiently small, implying strong stability since $1-g_i = O(\tau)$, or (ii) $\tau = O(h^{\alpha m})$, $1 \leq \alpha \leq 2$, and then $|1-R(-\tau_i\lambda_N)| = |1-Q(\frac{1}{\tau_i\lambda_N})| = O(h^{(2-\alpha)m})$, implying that

$$|1 - R(-\tau_i\lambda_N)| \Big/ \tau_i = O(h^{(2-2\alpha)m}) , \quad -2 \leq 2-2\alpha \leq 0 \tag{18}$$

thus $(1-g_j)/\tau_j = [1 - R(-\tau_j\lambda_1)]/\tau_j = O(1)$ , hence strong stability.

Finally for (iii) $\tau = O(h^{\alpha m})$ , $0 < \alpha \leqq 1$ , then (18) is still valid with $2-2\alpha \geqq 0$ , and thus

$$(1-g_j)/\tau_j = |1 - R(-\tau_j\lambda_N)| \ / \ \tau_j \to 0 \quad \text{as} \quad \tau(\text{or } h) \to 0 \ .$$

i.e. no strong stability. This completes the proof of the second part of the theorem.

For the last part : $\ell > r$ , let $(0,a)$ be the interval such that $|R_{\ell,r}(-x)| \leqq 1$ , for $x \in (0,a)$, then by choosing $\tau_j\lambda_N \leqq a$ , i.e. $\tau_j \leqq \frac{a}{\lambda_N}$ , we obtain the stability of the corresponding scheme. Suppose now that $\tau_j$ is choosen such that $\tau_j\lambda_N \leqq a_1 < a$ , and thus $R_{\ell,r}(-\tau_j\lambda_N) \leqq \delta < 1$ , then for $\tau_j$ sufficiently small we have

$$g_j = R_{\ell,r}(-\tau_j\lambda_1) \quad \text{and} \quad 1-g_j = 1 - R_{\ell,r}(-\tau_j\lambda_1) = O(\tau_j) \ , \ \text{i.e.}$$

strong stability. ∎

We assume now that $L$ and $L^*$ are not symmetric. Then it is no longer possible to write

$$g_j = \max_i |R_{\ell,r}(-\tau_j\lambda_i)| \ ;$$

However according to a theorem of Von-Neumann [8, pp. ], because

$$(Lv,v) \geqq \gamma_0 \|v\|^2 \quad \text{or} \quad (\text{Re}(Lv,v) \geqq \gamma_0\|v\|^2) \ ,$$

it is still possible to write that

$$g_j \leqq \sup_{\text{Re}(z) \geqq \gamma_0} |R_{\ell,r}(-\tau_j z)| . \tag{19}$$

This important contribution is however conservative, in the sense that it restricts the results to those methods where $|R_{\ell,r}(-z)| \leqq 1$ for $\text{Re}(z) \geqq 0$ , i.e. methods that are by definition A-stable.

For the Padé table, Ehle [9], has proved that when $\ell = r$ , $r-1$ , $r-2$ , then $|R_{\ell,r}(-z)| \leqq 1$ , for $Re(z) \geqq 0$ . Furthermore for $\ell < r-2$ , this property is not true in general.

Thus we can state

THEOREM 4

Suppose $L$ is not symmetric, i.e., $a(v,w) \neq a(w,v)$ for some $v,w \in M$ and let (A1), (A2), (A3)' be satisfied, then

(i) The Padé under diagonal schemes $\ell = r-1$ , $\ell = r-2$ , are strongly stable,

(ii) The Padé diagonal schemes $\ell = r$ are weakly stable.

PROOF

Using (19), and Ehle results, we have if $\ell = r-1$ , $r-2$ ,

$$g_j \leqq \sup_{Re(z) \geqq \gamma_0} |R_{\ell,r}(-\tau_j z)| \leqq 1$$

Following a proposition of Crouzeix [8, p. 31] , there exists a positive constant $\gamma_1$ such that

$$\sup_{Re(z) \geqq \gamma_0} |R_{\ell,r}(-\tau_j z)| \leqq \exp(-\gamma_i \tau_j)$$

Using theorem 1, we obtain therefore strong stability. This proves the first part of the theorem. For the second part, it is a well-known result of Birkhoff and Varga [10] that

$$\sup_{Re(z) \geqq 0} |R_{\ell,r}(-z)| \leqq 1 \quad \text{for} \quad \ell = r ,$$

with $$|R_{\ell,r}(-z)| = 1 , \quad \text{if} \quad z = iy .$$

This immediately yields the weak stability. ■

## 2.2. Application to error estimates

### 2.2.1. Preliminaries

We are now prepared to obtain error estimates for the methods described in §1 . We shall consider successively the cases $\ell \leqq r = 1$ and $\ell \leqq r = 2$ . The first case has been analyzed by many authors ([11], [12], for example) ; our analysis will be based on the results obtained in §2.1., and the V-projection of $u_j^k$ , $k \leqq 1$ , on $M$ . Specifically we define $\tilde{U}_j^k \in M$ , such that

$$a(\tilde{U}_j^k, v) = a(u_j^k, v) \quad \forall\, v \in M \,,\ 0 \leqq k \leqq \max(\ell, r) \tag{20}$$

It is known that $\tilde{U}_j^k$ satisfies best approximation properties. We shall then compare $e_j^k = \tilde{U}_j^k - U_j^k$ , with $\epsilon_j^k = \tilde{U}_j^k - u_j^k$ . From (1), we need to define

$$\sigma_j(t) \equiv \sigma_j^{\ell,r}(x,t) = u(x; t_j^{<\ell>}, t_{j+1}^{<r>}, t) \,.$$

It can be also easily proved that

$$\text{if } u \in C^{\ell+r+1}((0,T);H) \,, \text{ then } \sigma_j \in C((0,T);H) \,. \tag{21}$$

We shall also need the divided-difference

$$\Delta u_j = (u_{j+1} - u_j)/\tau_j \,. \tag{22}$$

Note from (20), that

$$a(\tilde{U}_{j+1} - \tilde{U}_j, v) = a(u_{j+1} - u_j, v), \ \forall\, v \in M$$

and hence

$$a(\Delta \tilde{U}_j, v) = a(\Delta u_j, v) \quad \forall\, v \in M \,. \tag{23}$$

If a depends on $t$ this property is no longer true.

2.2.2. The case $\ell = 0$ , $r = 1$

We have then

$$\left(\frac{\tilde{U}_{j+1} - \tilde{U}_j}{\tau_j}, v\right) + a(\tilde{U}_{j+1}, v) = (f_{j+1}, v) + (\Delta\tilde{U}_j - \Delta u_j, v) + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1} - s)(\sigma_j, v)ds \quad \forall\, v \in M$$

and therefore

$$\left(\frac{e_{j+1} - e_j}{\tau_j}, v\right) + a(e_{j+1}, v) = (\Delta\epsilon_j, v) + \int_{t_j}^{t_{j+1}} (t_{j+1} - s)(\sigma_j, v)ds \ , \ \forall\, v \in M$$

i.e.

$$(I + \tau_j L)e_{j+1} = e_j + \tau_j \cdot \tilde{f}_j$$

where

$$\tilde{f}_j = \Delta\bar{\epsilon}_j + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1} - s) \cdot \bar{\sigma}_j(s)ds \ , \text{ and } \|\tilde{f}_j\| \leq \|\Delta\epsilon_j\| + \tau_j \sup_j \|\sigma_j\| \quad \blacksquare$$

From theorems 3 and 4 and from (21), we can state

THEOREM 5

Under the assumptions of theorems 3 and 4, and if $u \in C^2\,[(0,T); H]$ then there exists $\mu_0 > 0$ such that for all $\tau$ , and $h$ ,

$$\|U_j - \tilde{U}_j\| \leq \exp(-\mu_0 t_j)\,\|U_0 - \tilde{U}_0\| + \frac{1}{\mu_0}\left\{\sup_j \|\Delta(u_j - \tilde{U}_j)\| + \tau_j \sup_j \|\sigma_j\|\right\}.$$

and therefore

$$(24)\quad \|u_j - U_j\| \leq \|u_j - \tilde{U}_j\| + \exp(-\mu_0 t_j)\,\|U_0 - \tilde{U}_0\| + \frac{1}{\mu_0}\sup_j\left\{\|\Delta(u_j - \tilde{U}_j)\| + \tau\,\|\sigma_j\|\right\}$$

2.2.3. The case $\ell = r = 1$ (Crank-Nicolson)

Similarly we have here

$$\left(\frac{\tilde{U}_{j+1} - \tilde{U}_j}{\tau_j}, v\right) + \frac{1}{2} a\,(\tilde{U}_{j+1} + \tilde{U}_j, v) =$$

$$= \frac{1}{2}\,(f_{j+1} + f_j, v) + (\Delta\tilde{U}_j - \Delta u_j, v) + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1} - s)(s - t_j)(\sigma_j, v)\,ds \quad \forall\, v \in M$$

and therefore

$$\left(\frac{e_{j+1} - e_j}{\tau_j}, v\right) + \frac{1}{2} a\,(e_j + e_{j+1}, v) = (\Delta\epsilon_j, v) + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1} - s)(s - t_j)(\sigma_j, v)\,ds\ .$$

yielding

$$(I + \frac{\tau_j}{2} L)\, e_{j+1} = (I - \frac{\tau_j}{2} L)\, e_j + \tau_j\, \tilde{f}_j\ ,$$

with $$\tilde{f}_j = \Delta\bar{\epsilon}_j + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1} - s)(s - t_j)\,\bar{\sigma}_j(s)\,ds\ ,$$

thus $$\|\tilde{f}_j\| \leq \|\Delta\epsilon_j\| + \tau_j^2 \sup_j \|\sigma_j\|\ . \quad \blacksquare$$

Theorems 3 and 4 allow us to state

THEOREM 6

Under the assumptions of theorems 3, 4 and if $u \in C^3[(0,T)\ ;\ H]$, then :

(i) if $\tau = O(h^{\alpha m})$, $\alpha \geq 1$, and $L$ is symmetric, there exists a positive constant $\mu_0$ such that

$$(25)\quad \|u_j - U_j\| \leq \|u_j - \tilde{U}_j\| + \exp(-\mu_0 t_j)\,\|U_0 - \tilde{U}_0\| + \frac{1}{\mu_0}\left\{\sup_j \|\Delta(u_j - \tilde{U}_j)\| + \tau^2 \sup_j \|\sigma_j\|\right\}.$$

(ii) Otherwise, if $\tau = O(h^{\alpha m})$, $\alpha < 1$ or $L$ is not symmetric, then

$$(26)\quad \|u_j - U_j\| \leq \|u_j - U_j\| + \|U_0 - \tilde{U}_0\| + t_j\left\{\sup_j \|\Delta(u_j - \tilde{U}_j)\| + \tau^2 \sup_j \|\sigma_j\|\right\}.$$

2.2.4. The case $\ell = 0$ , $r = 2$

We have here

$$\left(\frac{\tilde{U}_{j+1} - \tilde{U}_j}{\tau_j}, v\right) + a\left(\tilde{U}_{j+1} - \frac{\tau_j}{2}\tilde{U}^1_{j+1}, v\right) = \left(f_{j+1} - \frac{\tau_j}{2} f^1_{j+1}, v\right) + (\Delta\tilde{U}_j - \Delta u_j, v) +$$

$$+ \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1}-s)\ (\sigma_j, v)ds \quad \forall\, v \in M\ .$$

with

$$(\tilde{U}^1_j, v) + a(\tilde{U}_j, v) = (f_j, v) + (\epsilon^1_j, v) \quad \forall\, v \in M\ .$$

Thus we may write

$$\left(\frac{e_{j+1} - e_j}{\tau_j}, v\right) + a\left(e_{j+1} - \frac{\tau_j}{2} e^1_{j+1}, v\right) = (\Delta\tilde{U}_j - \Delta u_j, v) + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1}-s)^2 (\sigma_j, v)ds$$

and $$(e^1_j, v) + a(e_j, v) = (\epsilon^1_j, v)$$

so that

$$\frac{e_{j+1} - e_j}{\tau_j} + L\, e_{j+1} - \frac{\tau_j}{2} L\, e^1_{j+1} = \Delta\bar{\epsilon}_j + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1}-s)^2\, \bar{\sigma}_j\, ds$$

and $$e^1_j + L\, e_j = \bar{\epsilon}^1_j\ .$$

The last two equations give

$$\left(I + \tau_j L + \frac{\tau_j^2}{2} L^2\right) e_{j+1} = e_j + \tau_j\, \tilde{f}_j$$

with

$$\tilde{f}_j = \Delta\bar{\epsilon}_j + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1}-s)^2\, \bar{\sigma}_j\, ds + \frac{\tau_j}{2} L\, \bar{\epsilon}^1_{j+1}$$

and $$e_{j+1} = R_{0,2}\,(-\tau_j L)\, e_j + \tau_j\, \tilde{g}_j$$

with $\tilde{g}_j = R_{0,2}(-\tau_j L)\,[\Delta\bar{\epsilon}_j + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1}-s)^2\,\bar{\sigma}_j\,ds] + \frac{1}{2} R_{0,2}(-\tau_j)\,\tau_j\,L\,\bar{\epsilon}^1_{j+1}$

Let $$Q(z) = \frac{z}{2} R_{0,2}(-z) .$$

Then from maximum principle considerations it is easily verified that,

$$|Q(z)| \leqq 1 \quad \text{for} \quad \mathrm{Re}(z) \geqq 0 .$$

So that, $\|\tilde{g}_j\| \leqq \|\Delta u_j - \Delta U_j\| + \tau_j^2 \|\sigma_j\| + \|u^1_{j+1} - \tilde{U}^1_{j+1}\|$ . ■

Thus we may state

THEOREM 7

Under the assumptions of Theorems 3, 4 and 6 there exists $\mu_0 > 0$ such that for all $\tau$ and $h$ we have

$$\|U_j - \tilde{U}_j\| \leqq \exp(-\mu_0 t_j)\,\|U_0 - \tilde{U}_0\| + \frac{1}{\mu_0} \sup_j \left\{\|\Delta u_j - \Delta\tilde{U}_j\| + \tau^2 \|\sigma_j\| + \|u^1_j - \tilde{U}^1_j\|\right\}$$

and therefore,

$$\|u_j - U_j\| \leqq \|u_j - \tilde{U}_j\| + \exp(-\mu_0 t_j)\,\|U_j - \tilde{U}_j\| + \tag{27}$$

$$+ \frac{1}{\mu_0} \sup_j \left\{\|\Delta u_j - \Delta\tilde{U}_j\| + \tau^2 \|\sigma_j\|\right\} + \|u^1_j - \tilde{U}^1_j\| .$$

The same study can be made to the

2.2.5. The case $\ell = 1$ , $r = 2$

We write directly in this case

$$e_{j+1} = R_{1,2}(-\tau_j L)\,e_j + \tau_j\,\tilde{g}_j$$

with $\tilde{g}_j = (I + \frac{2\tau_j}{5} L + \frac{\tau_j^2}{6} L^2)^{-1} \left\{\Delta\bar{\epsilon}_j + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1}-s)^2 (s-t_j)\bar{\sigma}_j\,ds + \frac{1}{6}\tau_j\,L\,\bar{\epsilon}^1_{j+1}\right\}$

and $$\|\tilde{g}_j\| \leq \|\Delta u_j - \Delta\tilde{U}_j\| + \tau_j^3 \|\sigma_j\| + \|\epsilon_{j+1}^1\|$$

since it is also easily verified here that

$$\left|\frac{z/6}{1 + 2z/3 + z^2/6}\right| \leq 1 \quad \text{for} \quad \mathrm{Re}(z) \geq 0 ,$$

thus we can state :

THEOREM 8

Under the assumptions of theorems 3, 4 and if $u \in C^4[(0,T) ; H]$ , there exists $\mu_0 > 0$ , such that for all $\tau$ and $h$ we have

$$\|U_j - \tilde{U}_j\| \leq \exp(-\mu_0 t_j)\|U_0 - \tilde{U}_0\| + \frac{1}{\mu_0} \sup_j \left\{\|\Delta u_j - \Delta\tilde{U}_j\| + \tau^3\|\sigma_j\| + \|u_j^1 - \tilde{U}_j^1\|\right\}$$

and therefore

$$\|u_j - U_j\| \leq \|u_j - \tilde{U}_j\| + \exp(-\mu_0 t_j)\|U_j - \tilde{U}_j\| + \frac{1}{\mu_0} \sup_j \left\{\|\Delta u_j - \Delta\tilde{U}_j\| + \tau^3 \|\sigma_j\| + \|u_j^1 - \tilde{U}_j^1\|\right\} . \blacksquare$$

and finally we look at

2.2.6. The case $\ell = r = 2$

We have $$e_{j+1} = R_{2,2}(-\tau_j L)\, e_j + \tau_j\, \tilde{g}_j$$

where

$$\tilde{g}_j = (I + \frac{\tau_j}{2} L + \frac{\tau_j^2}{12} L^2)^{-1} \left\{\Delta\bar{\epsilon}_j + \frac{1}{\tau_j}\int_{t_j}^{t_{j+1}} (t_{j+1} - s)^2 (s - t_j)^2\, \bar{\sigma}_j\, ds + \frac{1}{12}\tau_j L \Delta\bar{\epsilon}_j^1\right\} ,$$

with $$\|\tilde{g}_j\| \leq \|\Delta\epsilon_j\| + \tau_j^4\|\sigma_j\| + \tau_j \|\Delta\epsilon_j^1\| ,$$

which allows us to state

THEOREM 9

Under the assumptions of theorems 3, 4 and if $u \in C^5[(0,T) ; H]$ , then if

(i) $L$ is a symmetric and $\tau = O(h^{\alpha m})$ , $\alpha \geq 1$ , there exists $\mu_0 > 0$ , such that

$$\|U_j - \tilde{U}_j\| \leq \exp(-\mu_0 t_j)\|U_0 - \tilde{U}_0\| + \frac{1}{\mu_0} \sup_j \left\{ \|\Delta u_j - \Delta\tilde{U}_j\| + \tau^4 \|\sigma_j\| + \tau\|\Delta u_j^1 - \Delta\tilde{U}_j^1\| \right\}$$

and

$$\|u_j - U_j\| \leq \|u_j - \tilde{U}_j\| + \exp(-\mu_0 t_j)\|U_0 - \tilde{U}_0\| + \frac{1}{\mu_0} \sup_j \left\{ \|\Delta u_j - \Delta\tilde{U}_j\| + \tau^4 \|\sigma_j\| + \tau\|\Delta u_j^1 - \Delta\tilde{U}_j^1\| \right\}$$

(ii) Otherwise,

$$\|U_j - \tilde{U}_j\| \leq \|U_0 - \tilde{U}_0\| + t_j \sup_j \left\{ \|\Delta u_j - \Delta\tilde{U}_j\| + \tau^4 \|\sigma_j\| + \tau\|\Delta u_j^1 - \Delta\tilde{U}_j^1\| \right\}$$

and

$$\|u_j - U_j\| \leq \|u_j - \tilde{U}_j\| + \|U_0 - \tilde{U}_0\| + t_j \sup_j \left\{ \|\Delta u_j - \Delta\tilde{U}_j\| + \tau^4 \|\sigma_j\| + \tau\|\Delta u_j^1 - \Delta\tilde{U}_j^1\| \right\}. \quad \square$$

REMARK

In theorems 5 - 8 , the estimates depend on $\Delta u_j - \Delta\tilde{U}_j$ , and in theorem 8 also on $\Delta u_j^1 - \Delta\tilde{U}_j^1$ . It is not difficult to estimate these.

REMARK

In theorems 5 - 8 , the estimates depend on $\Delta u_j - \Delta\tilde{U}_j$ , and $\Delta u_j^1 - \Delta\tilde{U}_j^1$ (as in theorem 9). To estimate these, we shall see that the regularity of $u(t)$ (with respect to the space variables) imply that of $\Delta u_j^k$ . We shall prove

LEMMA 2

Let $W$ be a closed and dense subspace in $H$ , with $\|\cdot\|_W$ , then $u \in C^{k+1}[(0,T) ; W]$ implies that $\Delta u_j^k \in W$ .

PROOF

It is sufficient to note that

$$\Delta u_j^k = \frac{1}{\tau_j} \int_{t_j}^{t_{j+1}} u^{(k+1)}(s)\, ds$$

and therefore

$$\|\Delta u_j^k\|_W \leq \sup_{t_j \leq s \leq t_{j+1}} \|u^{(k+1)}(s)\|_W ,$$

which yields our assertion. ∎

Let us then assume that $W = H^s(\Omega)$ , $M \equiv M(h,p) \subset W \subset V$ and for $v \in W \subset V$

$$\inf_{z \in M} \left\{ \|v-z\| + h|||v-z||| \right\} \leq Ch^s \|v\|_W , \quad s \leq p+1 ,$$

We shall prove

LEMMA 3

Assume $u \in C^2[(0,T) ; W]$ , then

$$\|\Delta u_j - \Delta \tilde{U}_j\| \leq Ch^s \sup_{0 \leq t \leq T} \|u'(t)\|_W , \; s \leq p+1$$

and

$$\|\Delta u_j^1 - \Delta \tilde{U}_j^1\| \leq Ch^s \sup_{0 \leq t \leq T} \|u''(t)\|_W , \; s \leq p+1$$

PROOF

The proof of this lemma follows easily from lemma 2, equation (23) and a similar one for $\Delta u_j$ , specifically,

$$a(\Delta U_j^1, v) = a(\Delta u_j^1, v) \quad \forall \; v \in M .$$

The estimates in the lemma become then striaghtforward.

Replacing these estimates in theorems 5,6,7,8 and 9, plus the (same) corresponding one for $u_j^k - \tilde{U}_j^k$ , $k = 0,1$, we get the expected order of convergence, namely $[O(h^s) + O(\tau^{\ell+r})]$ .

## 3. EFFICIENCY WITH RESPECT TO THE REGULARITY OF THE INITIAL CONDITION

In parabolic equations, the regularity of the initial condition can affect the accuracy of the method at the first stages of the computation. Even if the method is unconditionnaly stable and with a high order of convergence, the first and second stages of the computation may exhibit a loss in accuracy. This has been recently observed by Zlamal [12], with respect to firts order, one-step methods.

To simplify our study, let us consider the case when

(i) $L$ is symmetric
(ii) the homogeneous case $f = 0$

We consider the one-step discrete scheme that corresponds to the rational function $R(x)$, provided by a Padé scheme or by others (for example a Runge Kutta method). Let $\tau_j = \tau$, $\forall j$, and let us consider the exact and approximate solutions at the first time step. We write first

$$(28) \qquad \begin{cases} u_1 = \exp(-\tau L) u_0 \\ U_1 = R(-\tau L) U_0 \end{cases}$$

and the identity

$$(29) \quad u_1 - U_1 = [\exp(-\tau L) - R(-\tau L)]\, U_0 + \exp(-\tau L)\, [u_0 - U_0] + [\exp(-\tau L) - \exp(-\tau L)]\, U_0 .$$

Clearly the first term of the right hand side corresponds to the time discretization, while the second and the third correspond respectively to the discretization of the initial condition and to that of the operator $L$, and therefore independant of $R(x)$.

Thus for methods that have the same order of accuracy, and for any type of initial condition, there will be a gain in accuracy when

$$\| \exp(-\tau L) - R(-\tau L) \| \quad \text{is minimized.}$$

Clearly, under our above assumptions

$$(30) \qquad \|\exp(-\tau L) - R(-\tau L)\| = \max_{\lambda_i} |\exp(-\tau\lambda_i) - R(-\tau\lambda_i)|$$

where $\lambda_i$ is any eigenvalue of $L$. Therefore, if we put no restriction on $\tau$, a good estimate that follows from (30) is

$$(31) \qquad \|\exp(-\tau L) - R(-\tau L)\| \leq \sup_{0\leq x<\infty} |\exp(-x) - R(-x)| .$$

On the other hand if $\tau\lambda_N \leq M$, then

$$(32) \qquad \|\exp(-\tau L) - R(-\tau L)\| \leq \sup_{0\leq x<M} |\exp(-x) - R(-x)| .$$

Since (31) does not depend on $\tau$, and since we are considering unconditionally stable schemes, we shall use it to define an index that will indicate a gain or (loss) of accuracy at the initial stages of the computation.

DEFINITION

Let $R(x)$ be the rational function associated with a one-step discrete scheme for solving the evolution equation (2) ; the starting index, $S_R$, of this scheme, is defined by*

$$S_R = \sup_{0\leq x<\infty} |\exp(-x) - R(-x)| .$$

Hence, we shall look for high order methods with low starting indices. Clearly, the Crank-Nicolson and the Padé diagonal schemes cannot be "good starting" methods since for these $S_R = 1$. On the contrary the Padé underdiagonal schemes have substantially better indices. In table 1, we show $S_R$ for $\ell \leq r \leq 3$, and in table 2, the abcisse $x_M \in (0,\infty)$ at which $S_R$ is attained. These have been obtained numerically by Gerald Wanner at Geneva University.

---

* $S_R$ is also the Chebyshev constant. See Varga R.S., "Functional Analysis and Approximation Theory in Numerical Analysis", SIAM Publication 3.

| r \ ℓ | | | | |
|---|---|---|---|---|
| 0 | 1 | | | |
| 1 | 0.208 | 1 | | |
| 2 | 0.068 | 0.098 | 1 | |
| 3 | 0.027 | 0.024 | 0.064 | 1 |

Table 1 : Values of $S_R$ for $\ell \leq r \leq 3$

| r \ ℓ | | | | |
|---|---|---|---|---|
| 0 | ∞ | | | |
| 1 | 3 | ∞ | | |
| 2 | 3 | 8 | ∞ | |
| 3 | 3 | 7 | 18 | ∞ |

Table 2 : Values of $x_M \in (0,\infty)$, where $S_R$ is attained.

Next, we consider first order methods that correspond to $R(x) \equiv R_\theta(x) = \frac{1+\theta x}{1-(1-\theta)x}$.

Note that for $\theta = 0$ and $\theta = 1$, one obtains respectively the Padé (0,1) and (1,1).

Estimating numerically $|\exp(-x) - R_\theta(x)|$, for $0 \leq x \leq 100$, Wanner found this quantity minimized for $\theta = 0.122$, which is close to the $\theta = \frac{1}{6}$ given by Zlamal.

Hence, there appears to be a considerable advantage in using Padé underdiagonal schemes since they have low starting indices and thus will improve the accuracy at the first stages of the computation. Moreover as we shall see in the next section, they are quite efficient from the computational point of view.

## 4. EFFICIENCY WITH RESPECT TO COMPUTATION

This has been discussed already in [6] . We shall briefly review it here. As it is known already, the Padé (0,1) and (1,1) do not present any problem since they involve only linear factors. However the Padé (0,2), (1,2) and (2,2) involve the quadratic factors,

$(I+\tau L+\frac{\tau^2}{2}L^2)$ , $(I+\frac{2}{3}\tau L+\frac{\tau^2}{6}L^2)$ , and $(I+\frac{\tau}{2}L+\frac{\tau^2}{12}L^2)$ , (respectively).

Unfortunately, there have complex conjuguate roots, and necessitate the use of complex arithmetic if one is to avoid working with full matrices. Let us suppose we are to solve the system

$$(I + \tau L + \frac{\tau^2}{2} L^2) Z = Y \ , \tag{33}$$

as in the Padé (0,2). By considering the roots $\alpha$ and $\bar{\alpha}$ of the polynomial $(1+z+\frac{z^2}{2})$ , with $\alpha = 1+i$ , we can rewrite (33) as

$$\frac{1}{2}(\alpha+\tau L)(\bar{\alpha}+\tau L) Z = Y$$

which decomposes into

$$(\alpha + \tau L)\ Z_1 = Y$$

and

$$(\bar{\alpha} + \tau L) Z = 2Z_1 \ .$$

Noting that in the last equation $Z$ is real ; then by equating its real and imaginary parts we obtain

$$Z = \frac{2}{\mathrm{Im}(\bar{\alpha})}\ \mathrm{Im}(Z_1)$$

which reduces (33) into the solution of one complex system of equations, namely

$$(\alpha + \tau L)\, Z_1 = Y \quad \text{and} \quad Z = \frac{2}{\mathrm{Im}(\alpha)}\, \mathrm{Im}(Z_1) \ .$$

Thus, using this technique one proves that the Padé (0,2), (1,2) and (2,2) require the solution of one complex system of sparse linear equations at each time step. See [6], [7] for details and numerical results.

## REFERENCES


[1] Descloux J. and Nassif N.R.

"Non-Linear Padé Approximations and their Applications to Non-Linear Parabolic Equations"

(To appear)

[2] Isaacson E. and Keller H.B.

"Analysis of Numerical Methods", Wiley (1966)

[3] Nassif N.R., Thesis, Harvard University, 1972

[4] Lions J.L., "Equations Différentielles Opérationnelles", Springer (1961)

[5] Douglas J. and Dupont T.

"Galerkin Methods for parabolic partial differential equations"

SIAM J. Num. Anal., 7 (1970) 575-626

[6] Nassif N.R.

"Numerical solution of the heat equation by Galerkin generalized Crank-Nicolson"

Calcolo (To appear)

[7] Nassif N.R.

"On the discretization of the time variable in parabolic partial differential equations"

MAFELAP Brunel (1975) (To appear)

[8] Crouzeix M.

"Sur l'approximation des équations différentielles opérationnelles linéaires par des méthodes de Runge-Kutta"

Thesis, Université de Paris VI

[9] Richtmyer R.D. and Morton K.W.

"Difference Methods for Initial Value Problems"

Interscience, New-York (1967)

[10] Ehle B., Thesis, University of Waterloo (1969)

[11] Birkhoff G. and Varga R.S.

"Discretization errors for well-set Cauchy problems" , I.J. Math. Phys., 44 (1965) 1-23

[12] Zlamal M.

"Finite Element Methods for parabolic equations"

Math. Comp., 28 (1974) 393-404

[13] Zlamal M., Brunel (1975)