# On multistep approximation of semigroups in Banach spaces

Michel Crouzeix

INRIA

CENTRE DE RENNES
IRISA

Rapports de Recherche

N° 572

# ON MULTISTEP APPROXIMATION OF SEMIGROUPS IN BANACH SPACES

Michel CROUZEIX

Octobre 1986

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE
Tél. : (99) 36.20.00
Télex : UNIRISA 95 0473 F

# ON MULTISTEP APPROXIMATION OF SEMIGROUPS IN BANACH SPACES

# APPROXIMATION DE SEMI-GROUPES DANS DES ESPACES DE BANACH PAR DES METHODES MULTIPAS

Publication Interne n°310 - Septembre 1986
14 pages

Michel CROUZEIX
Mathématiques et Informatique
IRISA- Université de Rennes
Campus de Beaulieu
35042 RENNES CEDEX
(FRANCE)

Abstract : We consider a multistep rational approximation of a bounded, strongly continuous semigroup on a Banach space. We study the convergence when the time step converges to zero under a weak form of A-stability assumption.

Résumé : Etant donné un semi-groupe fortement continu dans un espace de Banach et son approximation par une méthode rationnelle multipas, nous étudions la convergence sous une hypothèse de A-stabilité affaiblie.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

# ON MULTISTEP APPROXIMATION OF SEMIGROUPS IN BANACH SPACES

by Michel Crouzeix

*Abstract* : *we consider a multistep rational approximation of a bounded strongly continuous semigroup on a Banach space. We study the convergence when the time step converges to zero under a weak form of A-stability assumption.*

## 1. Introduction.

Given a uniformly bounded, strongly continuous semigroup $e^{tA}$ on a Banach space $X$ :

$$\| e^{tA} \| \leq C_0 \quad , \tag{1}$$

and $u_0$ in $X$ , we approximate the value $u(t) = e^{tA} u_0$ at the time $t_n = n \Delta t$ by the solution $u_n$ of the q-step method

$$u_{n+1} = \sum_{j=0}^{q-1} r_j( \Delta t A) u_{n-j} \quad , n \geq q-1 , \tag{2}$$

starting from the procedure

$$u_j = d_j( \Delta t A) u_0 , \quad j=1,\ldots,q-1 \tag{3}$$

where $r_j(z), d_j(z)$ are rational functions uniformly bounded for $\operatorname{Re} z \leq 0$ and $\Delta t > 0$ denotes the time step.

We first consider the scalar case where $X = \mathbf{C}$ , $u_0 = 1$ and $A = \lambda \in \mathbf{C}$ with $\operatorname{Re} \lambda \leq 0$ . In order to obtain a global error $u(t_n) - u_n = O( \Delta t^p)$, it is natural to assume a local error in (2) and in (3) of order p+1, that is to say , setting $z = \lambda \Delta t$,

$$e^{qz} - \sum_{j=0}^{q-1} r_j(z) e^{(q-j-1)z} = O(z^{p+1}) \tag{4}$$

and

$$e^{jz} - d_j(z) = O(z^{p+1}) , \quad j=1,\ldots,q-1. \tag{5}$$

We introduce the polynomial

$$P(X ; z) = X^q - \sum_{j=0}^{q-1} r_j(z) X^{q-j-1} , \tag{6}$$

associated with the linear recursion formula (2) ; then Condition (4) may be rewritten as

$$P(e^z ; z) = O(z^{p+1}). \tag{7}$$

We shall say that Scheme (2) is A-stable with defect $k \geq 0$ if :
*for all* $z$ *with* $\mathrm{Re}\, z \leq 0$ , *the roots of* $P( . ; z)$ *lie in the unit disk and the mulplicities of the unimodular roots are less than* $k+1$.

The integer $k$ is chosen as small as possible. We remark that, from (7), 1 is a root of $P( . ; 0)$. The case $k = 0$ corresponds to the classical notion of A-stability.

**Remark** : With each monic polynomial $P( . ; z)$, whose coefficients are rational functions of $z$, we can associate a method of Form (2) ; so, we can construct high order schemes in a simple way :
- if the method associated with the polynomial $P( . ; z)$ is of order $p$ and A-stable with defect $k$, the method associated with $P( . ; z)^m$ is of order $mp+m-1$ and A-stable with defect $km+m-1$. Similarly, if another method, associated with $Q( . ; z)$, is of order $q$ and A-stable with defect $r$, the method associated with $P( . ; z)Q( . ; z)$ is of order $p+q+1$ and A-stable with defect $k+r+1$.

The idea to consider this kind of construction was born during some discussions with G.A. Baker, see also [1].

In this note we prove the following theorem :

**Theorem 1**. *If Scheme* (2) *is A-stable with defect* $k$ , *if Conditions* (5) *and* (7) *are satisfied and if* $u_0$ *belongs to* $D(A^{p+1})$, *then the following inequality holds*

$$\| u(t_n) - u_n \|_X \leq C_1\, C_0\, t_n^{k+1}\, \Delta t^{p-k}\, \|A^{p+1} u_0\|_X \quad ,$$

*where the constant* $C_1$ *depends only on the rational functions* $r_j$ *and* $d_j$ .

The proof of this theorem uses the mathematical framework described in the paper of Brenner and Thomée [4] and some technical lemmas ; the case of analytical semigroups is considered in Section 3 and the hilbertian case in Section 4. For related works, see [2], [3], [6], [7].

## 2. Proof of Theorem 1.

We introduce the Frobenius matrix $R(z)$ associated with $P( . ; z)$ and the corresponding linear operator $R(\Delta t A)$ in $\mathcal{L}(X^q,X^q)$ defined by

$$R(z) = \begin{pmatrix} r_0 & r_1 & \cdots & \cdots & r_{q-1} \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \qquad R(\Delta t A) = \begin{pmatrix} R_0 & R_1 & \cdots & \cdots & R_{q-1} \\ I & 0 & \cdots & \cdots & 0 \\ 0 & I & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & I & 0 \end{pmatrix}$$

where $r_j$ stands for $r_j(z)$ and $R_j$ for $r_j(\Delta t A)$ ; we also consider the vectors in $X^q$ of approximate and exact solutions :

$$U_n = \begin{pmatrix} u_{n+q-1} \\ u_{n+q-2} \\ \vdots \\ u_n \end{pmatrix}, V_n = \begin{pmatrix} u(t_{n+q-1}) \\ u(t_{n+q-2}) \\ \vdots \\ u(t_n) \end{pmatrix}, U_0 = \begin{pmatrix} d_{q-1}(\Delta tA)u_0 \\ \vdots \\ d_1(\Delta tA)u_0 \\ u_0 \end{pmatrix}, V_0 = \begin{pmatrix} e^{(q-1)\Delta tA}u_0 \\ \vdots \\ e^{\Delta tA}u_0 \\ u_0 \end{pmatrix}.$$

Clearly (2) implies

$$U_{n+1} = R(\Delta tA)\, U_n \ , \ V_{n+1} = e^{\Delta tA} V_n \,;$$

therefore

$$U_n = R(\Delta tA)^n\, U_0, \quad V_n = e^{n\Delta tA} V_0$$

and we have the error representation formula

$$U_n - V_n = H_n(\Delta tA)\, u_0 + G_n(\Delta tA)\, u_0 \,, \tag{8}$$

with

$$H_n(z) = (R(z)^n - e^{nz} I) \begin{pmatrix} e^{(q-1)z} \\ \vdots \\ e^z \\ 1 \end{pmatrix}, \ G_n(z) = R(z)^n \begin{pmatrix} d_{q-1}(z) - e^{(q-1)z} \\ \vdots \\ d_1(z) - e^z \\ 0 \end{pmatrix}. \tag{9}$$

**Lemma 2.** a) *If* $z_0$ *is not a pole of* $R(z)$, *there exist a neighborhood* $V_0$ *of* $z_0$ *in* $\overline{\mathbb{C}}$ *and a constant* $C$ *such that*

$$\forall\, n \geq k_0,\ \forall\, z \in V_0\,, \quad \|R(z)^n\| \leq C\, [\, \rho(R(z))^n + n^{k_0}\, \rho(R(z))^{n-k_0}] \,,$$

*where* $\rho(R(z))$ *denotes the spectral radius of* $R(z)$ *and* $(k_0+1)$ *the highest multiplicity of the eigenvalues of* $R(z_0)$ *having modulus* $\rho(R(z_0))$.

b) *If furthermore the eigenvalues of* $R(z_0)$ *which have a modulus equal to* $\rho(R(z_0))$ *are differentiable over* $V_0$, *there exists a constant* $C$ *such that*

$$\forall\, n \geq k_0,\ \forall\, z \in V_0\,, \ \left\| \frac{d}{dz}(R(z)^n)\right\| \leq C\, [\, n\, \rho(R(z))^{n-1} + n^{k_0+1}\, \rho(R(z))^{n-k_0-1}].$$

**Sketch of proof**. a) In a neighborhood of $z_0$, it is possible to find an analytic and invertible matrix $H(z)$ such that

$$H^{-1}(z)\, R(z)\, H(z) = \mathrm{diag}(\, R_s(z)\,) \tag{10}$$

is a block diagonal matrix where each square matrix $R_s(z)$ corresponds to one of the distinct eingenvalues of $R(z_0)$. Clearly $\|R(z)^n\| \leq \max_s \|R_s(z)^n\|$ ;

since the matrix $R(z)$ is not derogatory and the eigenvalues of $R(z_0)$ are continuous at the point $z_0$, we can find invertible matrices $P_s(z)$, continuous at the point $z_0$, such that

$$P_s(z)^{-1}\, R_s(z)\, P_s(z) = \begin{pmatrix} \lambda_{s0} & & & 0 \\ 1 & \lambda_{s1} & & \\ & 1 & \ddots & \\ 0 & & 1 & \lambda_{sk_s} \end{pmatrix} = J_s(z)\,,$$

where $\lambda_{sj}$ stands for $\lambda_{sj}(z)$. Part a) follows from inequalities

$$\| J_s(z)^n \| \leq \rho(J_s(z))^n + n\,\rho(J_s(z))^{n-1} + \dots\dots + C_n^{k_s}\ \rho(J_s(z))^{n-k_s}$$

and

$$\| R_s(z)^n \| \leq C\, \| J_s(z)^n \| .$$

b) From (10), we deduce

$$\| \frac{d}{dz}(R(z))^n \| \leq C \max_s (\|R_s(z)^n\| + \| \frac{d}{dz}(R_s(z))^n \|) \ ;$$

the only difficulty is to obtain an estimate of $\| \frac{d}{dz}(R_s(z))^n \|$ when

$\rho(R_s(z_0)) = \rho(R(z_0))$ . But, in this case the matrix $P_s(z)$ may be chosen differentiable and therefore

$$\| \frac{d}{dz}(R_s(z))^n \| \leq C \ (\|J_s(z)^n\| + \| \frac{d}{dz}(J_s(z))^n \|) \ ;$$

for the matrix $J_s(z)^n$ we have the following upper bound

$$\| \frac{d}{dz}(J_s(z))^n \|) \leq C\,[\, n\,\rho(J_s(z))^{n-1} + \dots + C_n^{k_s}\ (n - k_s - 1)\ \rho(J_s(z))^{n-k_s-1}\,] .$$

**Lemma 3.** *If Scheme* (2) *is A-stable with defect* k, *the unimodular eigenvalues of* $R(iy)$ , *when* y *is real, are twice differentiable near y.*

**Proof** : First, we remark that the eigenvalues of $R(z)$ are the roots of the polynomial $P( \,.\, ; z )$. Without loss of generality, we give the proof only for the case $y = 0$ and for the eigenvalues which converge to 1 as $z \to 0$. These eigenvalues can be written as Puiseux series

$$\lambda_j(z) = 1 + a_j z^{r_j} + b_j z^{s_j} + o(z^{s_j}) ,$$

where $0 < r_j < s_j$ are rational and $a_j \neq 0$. Since $|\lambda_j(z)| \leq 1$ when $\mathrm{Re}\, z \leq 0$ ,we cannot have $r_j > 1$ ; neither can $r_j = p_j / q_j$ be $< 1$, since each determination of $z^{1/q_j}$ has to be considered. Therefore we have $r_j = 1$ and $a_j > 0$ . Considering the case $z = iy$ ,we must have $0 \leq 1 - \mathrm{Re}\,\lambda_j(iy) \approx - \mathrm{Re}(b_j(iy)^{s_j})$ , which implies $s_j \geq 2$ .

Using the compactness of $\overline{\mathbf{R}}$ and the previous lemma, we obtain the following corollary.

**Corollary 4.** *If the scheme is A-stable with defect* k , *there exists a constant* C *such that*

$$\forall n \geq 1 ,\ \forall y \in \mathbf{R} ,\quad \| R(iy)^n \| \leq C\, n^k , \quad \| \frac{d}{dy} R(iy)^n \| \leq C\, n^{k-1} .$$

Now we resume the study of the convergence estimates . Since

$$H_1(z) = (R(z) - e^z I)(e^{(q-1)z},\dots\dots,e^z,1)^T = (P(e^z;z),0,\dots\dots,0)^T = O(z^{p+1}),$$

we obtain from (9) and (7)

$$H_n(z) = \sum_{j=0}^{n-1} e^{jz} R(z)^{n-j-1} H_1(z) = O(z^{p+1}).$$

Similarly, using (5) , we have $\quad G_n(z) = O(z^{p+1})$ .
Therefore the functions

$$\tilde{H}_n(z) = H_n(z)/z^{p+1} \quad , \quad \tilde{G}_n(z) = G_n(z)/z^{p+1} \tag{11}$$

are analytic and uniformly bounded in the halfplane $\mathrm{Re}z \le 0$ .
Using the background described in Brenner-Thomée, there exist two functions $\mathcal{H}_n$ and $\mathcal{G}_n$ in $L^1(\mathbf{R})^q$ such that, for $\mathrm{Re}z \le 0$ ,

$$\tilde{H}_n(z) = \int_0^\infty e^{tz}\, \mathcal{H}_n(t)\, dt\ , \quad \tilde{G}_n(z) = \int_0^\infty e^{tz}\, \mathcal{G}_n(t)\, dt\ , \tag{12}$$

and

$$\mathcal{H}_n(t) = \mathcal{G}_n(t) = 0 \text{ for } t < 0 . \tag{13}$$

We define the two operators in $\mathbf{L}(X^q,X^q)$

$$\tilde{H}_n(\Delta tA) = \int_0^\infty e^{t\Delta tA}\, \mathcal{H}_n(t)\, dt\ , \quad \tilde{G}_n(\Delta tA) = \int_0^\infty e^{t\Delta tA}\, \mathcal{G}_n(t)\, dt, \tag{14}$$

and we have from (8) and (11) , if $u_0$ is in $D(A^{p+1})$,

$$U_n - V_n = \Delta t^{p+1}\ \tilde{H}_n(\Delta tA)\, A^{p+1}u_0 + \Delta t^{p+1}\, \tilde{G}_n(\Delta tA)\, A^{p+1}u_0 .$$

Now, Theorem 1 is a simple consequence of the following lemma :

**Lemma 5** . *Under the hypotheses of Theorem 1, there exists a constant* C *such that*

$$\| \tilde{H}_n\, (\Delta tA) \| \le C\, C_0\, n^{k+1} \quad \text{and} \quad \| \tilde{G}_n(\Delta tA)\| \le C\, C_0\, n^{k+1/2} .$$

**Proof** . We have, from (14) and (1),

$$\| \tilde{H}_n(\Delta tA) \| \le C_0 \, \| \mathcal{H}_n \|_{L^1(\mathbf{R})^q} \text{ and } \| \tilde{G}_n(\Delta tA) \| \le C_0 \, \| \mathcal{G}_n \|_{L^1(\mathbf{R})^q} ,$$

therefore it is sufficient to prove that

$$\| \mathcal{H}_n \|_{L^1(\mathbf{R})^q} \le C\, n^{k+1} \quad \text{and} \quad \| \mathcal{G}_n \|_{L^1(\mathbf{R})^q} \le C\, n^{k+1/2} . \tag{15}$$

Let us introduce the functions : $\hat{\mathcal{H}}_n(y) = H_n(iy)$ , $\hat{\mathcal{G}}_n(y) = G_n(iy)$ ; from (12) and (13), $\hat{\mathcal{H}}_n$ and $\hat{\mathcal{G}}_n$ are the Fourier transforms of $\mathcal{H}_n$ and $\mathcal{G}_n$ . A Carlson 's inequality gives

$$\|\mathcal{G}_n\|_{L^1(\mathbf{R})^q} \leq C \,\|\hat{\mathcal{G}}_n\|^{1/2}_{L^2(\mathbf{R})^q} \,\|\hat{\mathcal{G}}_n{}'\|^{1/2}_{L^2(\mathbf{R})^q} \;;$$

from Corollary 4, we have

$$\|\hat{\mathcal{G}}_n(y)\| \leq C\, n^k \min\left(\frac{1}{|y|^{p+1}}, 1\right)$$

and

$$\|\hat{\mathcal{G}}_n{}'(y)\| \leq C\, n^{k+1} \min\left(\frac{1}{|y|^{p+1}}, 1\right)$$

therefore

$$\|\hat{\mathcal{G}}_n\|_{L^2(\mathbf{R})^q} \leq C\, n^k \;, \quad \|\hat{\mathcal{G}}_n{}'\|_{L^2(\mathbf{R})^q} \leq C\, n^{k+1} \;,$$

which shows the second inequality in (15).

In order to prove the first inequality, we introduce the notations

$$S(z) = \begin{pmatrix} s_0 & s_1 & \cdots & \cdots & s_{q-1} \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \qquad D(z) = \begin{pmatrix} e^{(q-1)z} & & & 0 \\ & \ddots & & \\ & & e^z & \\ 0 & & & 1 \end{pmatrix}$$

where $s_j$ stands for $s_j(z) = r_j(z)\, e^{-(j-1)z}$ . We have $R(z) = e^z D(z) S(z) D(z)^{-1}$ , therefore

$$\hat{\mathcal{H}}_n(y) = \frac{1}{(iy)^{p+1}}\, e^{niy} D(iy)\, (S(iy)^n - I)E = e^{niy} D(iy)\, \hat{T}_n(y) \tag{16}$$

where $E=(1,1,...,1)^T$ and

$$\hat{T}_n(y) = \frac{1}{(iy)^{p+1}} (S(iy)^n - I)\; E\,. \tag{17}$$

We remark that $\hat{T}_n$ is the Fourier transform of $T_n$ which, from (16), is related to $\mathcal{H}_n$ by the relation , (on the $j^{th}$ component),

$$\mathcal{H}_n(t)_j = T_n(t-(n+q-j))_j$$

therefore , using a Carlson's inequality , we obtain

$$\|\mathcal{H}_n\|_{L^1(\mathbf{R})^q} \leq C' \|T_n\|_{L^1(\mathbf{R})^q} \leq C \,\|\hat{T}_n\|^{1/2}_{L^2(\mathbf{R})^q} \,\|\hat{T}_n{}'\|^{1/2}_{L^2(\mathbf{R})^q} \,.$$

The following lemma completes the proof of the first inequality in (15).

**LEMMA 6** . *If the multiplicity of the root* 1 *of* P ( . ; 0 ) *is* k+1, *then we have*

$$\|\hat{T}_n\|_{L^2(\mathbf{R})} \le C\, n^{k+1-1/(2r+2)} \quad \text{and} \quad \|\hat{T}_n'\|_{L^2(\mathbf{R})} \le C\, n^{k+1+1/(2r+2)},$$

*where* r *is a rational (which will be defined in the proof ).*
*If the multiplicity is less than or equal to* k, *then*

$$\|\hat{T}_n\|_{L^2(\mathbf{R})} \le C\, n^{k} \quad \text{and} \quad \|\hat{T}_n'\|_{L^2(\mathbf{R})} \le C\, n^{k+1}.$$

**Proof** . We consider only the case where the multiplicity of the root 1 is k+1, the other case is easier . Relation (17) and Corollary 4 imply

$$\|\hat{T}_n(y)\| \le C\, n^k/\,|y|^{p+1} \quad \text{and} \quad \|\hat{T}_n'(y)\| \le C(n^k/\,|y|^{p+2} + n^{k+1}/|y|^{p+1}). \tag{18}$$

In the definition of S(z) , we remark that $\mu(z)$ is an eigenvalue of S(z) if and only if $\lambda(z) = e^z\, \mu(z)$ is an eigenvalue of R(z) , and that the algebraic multiplicity is preserved. From Lemma 2, we can find, in a neiborhood $\mathcal{V}$ of zero on the imaginary axis, an invertible and twice differentiable matrix H(z) such that

$$\forall\, z \in \mathcal{V},\ S(z) = H(z)^{-1} \begin{pmatrix} A(z) & 0 \\ 0 & B(z) \end{pmatrix} H(z)\ , \tag{19}$$

where the matrix B(z) − I is invertible and the (k+1)x(k+1) matrix A(z) has the form

$$A(z) = \begin{pmatrix} \mu_0(z) & & & 0 \\ 1 & \mu_1(z) & & \\ & 1 & \ddots & \\ 0 & & 1 & \mu_k(z) \end{pmatrix}, \quad \text{with} \quad \mu_j(0) = 1\ .$$

Now we introduce the twice differentiable vectors $A_n(z)$, $B_n(z)$ , a(z), b(z), $\beta(z)$ by

$$\begin{pmatrix} A_n(z) \\ B_n(z) \end{pmatrix} = H(z)\,\hat{T}_n(z)\ , \qquad \begin{pmatrix} a(z) \\ b(z) \end{pmatrix} = H(z)\, E\ , \quad \beta(z) = \frac{1}{z^{p+1}}\, b(z)\ ; \tag{20}$$

then

$$A_n(z) = \frac{1}{z^{p+1}}\ (A(z)^n - I\,)\, a(z)\ ,\ \ B_n(z) = (\, B(z)^n - I\,)\, \beta(z)\ . \tag{21}$$

Since $\beta(z) = (B(z) - I)^{-1} B_1(z)$ is twice differentiable, we obtain

$$\|B_n(iy)\| \le C\, n^k\ ,\ \|B_n'(iy)\| \le C\, n^{k+1}\ ,\ \text{for } iy \in \mathcal{V}\ . \tag{22}$$

We assume that the order of the eigenvalues $\mu_j(z)$ of $A(z)$ has been chosen in order to ensure

$$0 \le r = r_0 \le r_1 \le \ldots \le r_k ,$$

for the first exponent of the Puiseux expansion

$$\mu_j(z) = 1 + a_j z^{r_j+1} + \ldots\ldots , \quad r_j \in \mathbb{Q}, \quad a_j \neq 0 .$$

Noticing that

$$A_n(z) = \sum_{j=0}^{n-1} A(z)^j A_1(z) \quad \text{and} \quad \| A'(z) \| \le C \,|\, z \,|^r ,$$

we get

$$\| A_n(z) \| \le C\, n^{k+1} , \quad \| A_n'(z) \| \le C\, n^{k+1} (1 + n\, |z|^r) . \tag{23}$$

With the notations

$$a(z) = (a_0(z), \ldots, a_k(z))^T , \quad e_0 = (1, 0, \ldots, 0)^T$$

$$c(z) = (A(z) - I)(\, a(z) - a_0(z)\, e_0 \,) = (0, c_1(z), \ldots, c_k(z))^T , \tag{24}$$
$$(\, c(z) = 0 \quad \text{when } k = 0 \,),$$

we can also write

$$A_n(z) = \frac{1}{z^{p+1}} \left[\, a_0(z)\, (A(z)^n - I\,)\, e_0 + \sum_{j=0}^{n-1} A(z)^j\, c(z) \,\right] . \tag{25}$$

Since the first row and the first column of $A(z)$ are not used in $A(z)^j\, c(z)$, this term is bounded by $C\, j^{k-1} \|c(z)\|$. Then

$$\| A_n(z) \| \le C\, n^k \,(\, |a_0(z)| + \|c(z)\| \,) / |z|^{p+1} . \tag{26}$$

Similarly, by derivation of (25), we obtain

$$\| A_n'(z) \| \le (p+1)\, \|A_n(z)\| / |z| + C\, n^k \,(\, |a_0'(z)| + \|c'(z)\| \,) / |z|^{p+1}$$
$$+ C\, n^{k+1} \,|\, z \,|^r \,(\, |a_0(z)| + \|c(z)\| \,) / |z|^{p+1} . \tag{27}$$

From the relations

$$(\, A(z) - I \,)\, a(z) = z^{p+1} A_1(z) ,$$

$$(\, \mu_0(z) - 1)\, a_0(z) = z^{p+1} \,(A_1(z)\,)_0 ,$$

we get

$$a_0(z) = O(\, z^{p-r}) , \quad a_0'(z) = O(\, z^{p-r-1}) ,$$

$$c(z) = z^{p+1} A_1(z) - a_0(z)\, (A(z) - I\,)\, e_0 = O(z^{p-r}) ,$$

and

$$c'(z) = O(\, z^{p-r-1}) .$$

Together with (26), (27) and (23), it yields

$$\| A_n(z) \| \le C\, n^k / |z|^{r+1},\ \| A_n'(z) \| \le C\, n^{k+1} / |z| \text{ for } z \in \mathcal{V}.$$

Considering (18), (22) and (23), we have proved that, for all $y \in \mathbf{R}$,

$$\| \hat{T}_n(y) \| \le C\, n^k \min( n, 1/|y|^{r+1} ),$$

$$\| \hat{T}_n'(y) \| \le C\, n^{k+1} \min( 1+n\,|y|^r, 1/|y| ).$$

The lemma follows from a simple calculation.

**3. The case of holomorphic semigroups.**

In this section, we make the stronger assumption that A generates a holomorphic semigroup on X ; more precisely we also assume that the spectrum of A is included in the sector $S_\theta$ and

$$\forall z \in \mathbf{C}\backslash S_\theta,\ \ \|(zI - A)^{-1}\| \le C / | z | , \tag{28}$$

where $\theta \in ]0, \pi/2[$, $C$ is a constant and

$$S_\theta = \{ z \in \mathbf{C} ;\ \pi - \theta \le | \operatorname{Arg} z | \le \pi \text{ or } z = 0 \}.$$

We can make weaker assumptions on the scheme ; we assume that the rational functions $r_j$ and $d_j$ are uniformly bounded in $S_\theta$ and satisfy

$$P( e^z ; z) = O(z^{p+1}) \tag{7}$$

and

$$e^{jz} - d_j(z) = O(z^p),\ \ j=1,\dots,q-1 ; \tag{29}$$

the last requirement is weaker than (5). We suppose also that the method is A($\theta$)-stable with defect k, that is to say

*for all* $z$ *belonging to* $S_\theta$ *, the roots of* $P( . ; z)$ *lie in the unit disk and the mulplicities of the unimodular roots are less than* $k+1$.

We need also the following hypothesis : there exist $\eta > 0$ and $\mu$, $0<\mu<\cos\theta$, such that

*for all root* $\lambda_j(z)$ *of* $P( . ; z)$ *such that* $\lambda_j(0)$ *is unimodular and of multiplicity* $k+1$ *, we have*
$\forall z \in S_\theta$ *with* $| z | \le \eta$, $| \lambda_j(z) | \le e^{-\mu|z|}$. (30)

We have the following theorems :

**Theorem 7.** *If the scheme is $A(\theta)$-stable with defect* $k$ *, if conditions* (7) ,(28), (29) *and* (30) *are satisfied and if* $u_0$ *belongs to* $D(A^p)$ *, then the following inequality holds*

$$\| u(t_n) - u_n \|_X \leq C\, t_n^{\,k}\ \Delta t^{p-k} \| A^p u_0 \|_X \ .$$

**Theorem 8.** *If the same assumptions are satisfied but* $u_0$ *belongs to* $X$ *, if furthermore , for all* $z \in S_\theta - \{0\}$ *and for* $z = \infty$ *, the roots of* $P(\,.\,;z)$ *lie in the open unit disk , then the following inequality holds*

$$\| u(t_n) - u_n \|_X \leq C\, \Delta t^{p-k} / t_n^{\,p-k} \|u_0\|_X \ .$$

**Proof of Theorem 7.** We use the same representation formula (8)

$$U_n - V_n = H_n(\Delta t A)\, u_0 + G_n(\Delta t A)\, u_0 \ , \tag{8}$$

as for Theorem 1 , but instead of (11), we use the functions

$$\tilde{H}_n(z) = H_n(z)/z^p \quad , \quad \tilde{G}_n(z) = G_n(z)/z^p \ . \tag{31}$$

Then

$$U_n - V_n \ = \Delta t^p \,(\, \tilde{H}_n(\Delta t A) + \tilde{G}_n(\Delta t A)\,)\, A^p u_0$$

and

$$\| U_n - V_n \|_{Xq} \leq C\, \Delta t^p \,(\| \tilde{H}_n(\Delta t A)\| + \| \tilde{G}_n(\Delta t A)\|)\, \|A^p u_0 \|_X \tag{32}$$

Since $\| \tilde{H}_n(z) \| \leq C_n \min(\, |z| \,,\, |z|^{-p}\,)$ , we can use the Dunford-Taylor spectral representation

$$\tilde{H}_n(\Delta t A) \ = \frac{1}{2\pi i} \int_{\Gamma_\theta} (z I - \Delta t A)^{-1}\, \tilde{H}_n(\, z\,)\, dz \ , \tag{33}$$

where $\Gamma_\theta$ denotes the oriented boundary of $S_\theta$ .

From Lemma 2 and (30), we deduce that

$$\forall\ z \in \Gamma_\theta \ , \qquad \| R(z)^n \| \leq C\, n^k \ ,$$

and

$$\forall\ z \in \Gamma_\theta \text{ with } |z| \leq \eta \ , \quad \| R(z)^n \| \leq C\, n^k\, e^{-\mu n |z|} \ ;$$

therefore

$$\forall\ z \in \Gamma_\theta \text{ with } z \neq 0 \quad , \quad \| \tilde{H}_n(z) \| \leq C\, n^k\, |z|^{-p}$$

and, since from (9) , $\tilde{H}_n(z) \ = \sum_{j=0}^{n-1} e^{jz}\, R(z)^{n-j-1} \tilde{H}_1(z) \ = O(z^{p+1})$.

$$\forall\ z \in \Gamma_\theta \text{ with } |z| \leq \eta \ , \quad \| \tilde{H}_n(z) \| \leq C\, n^{k+1}\, e^{-\mu n |z|}\, |z| \tag{34}$$

Then, using (33) and (28),

$$\| \tilde{H}_n(\Delta t A) \| \le \frac{1}{\pi} \int_0^\infty \frac{C}{r} \tilde{H}_n(r\, e^{i\theta})\, d\, r\ ,$$

$$\le C \int_0^\eta n^{k+1}\, e^{-\mu n r}\, dr + C\, n^k \int_\eta^\infty r^{-p-1}\, dr\ ,$$

$$\le C\, n^k\ .$$

We cannot use directly the Dunford-Taylor spectral representation for $\tilde{G}_n(\Delta t A)$ when $\tilde{G}_n(0) \neq 0$ , but we can write

$$\tilde{G}_n(\Delta t A) = e^{n \Delta t A}\, \tilde{G}_n(0) + \frac{1}{2\pi i} \int_{\Gamma_\theta} (z\, I - \Delta t A)^{-1}\, (\tilde{G}_n(z) - e^{nz}\, \tilde{G}_n(0))\, dz\ ;$$

therefore

$$\| \tilde{G}_n(\Delta t A) \| \le C\, n^k + \frac{1}{\pi} \int_0^\infty \frac{C}{r} \| \tilde{G}_n(r\, e^{i\theta}) - \exp(n r e^{i\theta}) \tilde{G}_n(0) \|\, dr\ ,$$

and we obtain easily

$$\forall\ z \in \Gamma_\theta \text{ with } |z| \ge \eta\ , \qquad \| \tilde{G}_n(z) - e^{nz} \tilde{G}_n(0) \| \le C\, n^k\, |z|^{-p}\ ,$$

and

$$\forall\ z \in \Gamma_\theta \text{ with } |z| \le \eta\ , \qquad \| \tilde{G}_n(z) - e^{nz} \tilde{G}_n(0) \| \le C\, n^{k+1}\, e^{-\mu n |z|}\, |z|\ .$$

Thus, $\| \tilde{G}_n(\Delta t A) \| \le C\, n^k$ and Theorem 7 follows from (32).

**Proof of Theorem 8.** From the representation formula (8), we have

$$\| u(t_n) - u_n \|_X \le C\, (\| H_n(\Delta t A) \| + \| G_n(\Delta t A) \|)\, \|u_0\|_X\ .$$

Now we write

$$H_n(\Delta t A) = (I - e^{\Delta t A})\, H_n(\infty) + \frac{1}{2\pi i} \int_{\Gamma_\theta} (z\, I - \Delta t A)^{-1} (H_n(z) - (1 - e^z) H_n(\infty))\, dz.$$

From Lemma 2, there exists $\beta$ , $0 < \beta < 1$, such that

$$\forall\ z \in \Gamma_\theta \quad \text{with } |z| \ge \eta\ , \qquad \| R(z)^n \| \le C\, \beta^n\ ;$$

using also $R(z) - R(\infty) = O(z^{-1})$, (when $z \to \infty$), we obtain

$$\forall\ z \in \Gamma_\theta \text{ with } |z| \ge \eta\ , \qquad \| H_n(z) - (1 - e^z) H_n(\infty) \| \le C\, n^{k+1}\, \beta^n\, |z|^{-1}\ ,$$

and from (34)

$\forall\, z \in \Gamma_\theta$ with $|z| \leq \eta$, $\| H_n(z)-(1-e^z)H_n(\infty) \| \leq C(n^{k+1}e^{-\mu n|z|}|z|^{p+1} + \beta^n |z|)$.
Therefore

$$\| H_n(\Delta t A) \| \leq C\,( n^{k+1}\beta^n + \int_0^\eta n^{k+1}e^{-\mu n r}\, r^p\, dr ),$$

$$\leq C\,( n^{k+1}\beta^n + n^{k-p}\int_0^\infty e^{-\mu t}\, t^p\, dt ),$$

$$\leq C\, n^{k-p} = C\,( \Delta t/t_n )^{p-k}.$$

Similarly we get

$\forall\, z \in \Gamma_\theta$ with $|z| \geq \eta$, $\| G_n(z)-(1-e^z)G_n(\infty) \| \leq C\, n^{k+1}\beta^n |z|^{-1}$,

and

$\forall\, z \in \Gamma_\theta$ with $|z| \leq \eta$, $\| G_n(z)-(1-e^z)G_n(\infty) \| \leq C(n^k e^{-\mu n|z|}|z|^p + \beta^n |z|)$,

which yields

$$\| G_n(\Delta t A) \| \leq C\,( \Delta t/t_n )^{p-k}$$

and completes the proof of Theorem 8.

## 4. Remarks.

Theorem 1 is valid in particular when X is a Hilbert space but in this case it may be improved by taking Condition (29) in the place of Condition (5). Furthermore an easier proof can be given ; indeed, changing possibly the norm in X, we can assume that $e^{tA}$ is a semigroup of contraction on X , i.e. $\| e^{tA} \| \leq 1$ , ( $\forall\, t > 0$ ) ; then, from a theorem of von Neumann, we have

$$\| \tilde{H}_n(\Delta t A) \| \leq C \max_{\mathrm{Re}\, z = 0} \| \tilde{H}_n(z) \| \leq C\, n^{k+1},$$

similarly , $\| \tilde{G}_n(\Delta t A) \| \leq C\, n^k$ , which completes the proof.

Now, if we assume further that A is a (semidefinite negative) selfadjoint operator, Theorem 7 and Theorem 8 are valid with $\theta = 0$ . In this case, the proof becomes very simple since, from the spectral theory,

$$\| H_n(\Delta t A) \| \leq \sup_{x \leq 0} \| H_n(x) \| ,\quad \| G_n(\Delta t A) \| \leq \sup_{x \leq 0} \| G_n(x) \| .$$

## References.

[1] **G.A. Baker**, *On approximations of holomorphic semigroups,* Internal report 78007, Université Pierre et Marie Curie (Paris 6), (1978).

[2] **G.A. Baker, J.H. Bramble and V. Thomée,** *Single step Galerkin approximations for parabolic problems,* Math.Comput. 31, (1977), pp.818-847.

[3] **J. Blair,** *Approximate solution of elliptic and parabolic boundary value problems,* thesis, Univ. of California, Berkeley, (1970).

[4] **P. Brenner and V. Thomée,** *On rational approximations of semigroups,* SIAM J. NUMER. ANAL., 16 n°4 (1979), pp. 683-694.

[5] **F. Carlson,** *Une inégalité*, Ark. Mat., 25B (1935).

[6] **H. Fujita and A. Mizutani,** *On the finite element method for parabolic equations,* J.Math.Soc.Japan 28, (1976), pp. 749-771.

[7] **M. N. Le Roux,** *Semi-discretization in time for parabolic problems,* Math. Comput. 33, (1979), pp. 919-931.

Michel Crouzeix
Mathématiques et Informatique
IRISA Université de Rennes
Campus de Beaulieu
35042 RENNES Cédex **(France)**

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique