数理解析研究所講究録
第 199 巻 1974 年 98-114

Approximation of Exponential Function of a Matrix by Continued Fraction Expansion

Masatake MORI (Research Institute for Mathematical Sciences, Kyoto University)

§1 Introduction

The solution of an equation of evolution

$$\frac{du}{dt} = Au, \quad u(0) = u_0 \tag{1.1}$$

in which A is an N×N matrix is formally given

$$u(t) = e^{tA}u_0 \tag{1.2}$$

where the matrix exp tA is defined by

$$e^{tA} = I + \frac{t}{1!}A + \frac{t^2}{2!}A^2 + \ldots \tag{1.3}$$

Such a system of ordinary differential equation is often a result of discretization of space variables of a certain time-dependent linear partial differential equation. Varga [1] has shown the relation between various methods for numerical solution of parabolic partial differential equations from the standpoint of the Padé approximation of the exponential function exp tA, and proposed new methods based on the higher order approximation of exp tA.

The purpose of the present paper is to give a method based on the continued fraction expansion of exp tA, where A is an N×N matrix, as a device to solve an equation of the form (1.1). This method may be included in those proposed by Varga, but it has an advantage that it is reduced to an iterative method in a simple form owing to the recurrence relation which gives the continued fraction expansion of e^z. Moreover, as will be shown below, the approximant $H_k(z)$ of e^z always satisfies $|H_k(z)|\leq 1$ in Re $z\leq 0$ and hence the resulting method is applicable to a family of non-self adjoint problems and is unconditionally stable as long as every eigenvalue of A lies in the left half-plane.

In order to express a continued fraction

$$b_0+\cfrac{a_1}{b_1+\cfrac{a_2}{b_2+\ddots+\cfrac{a_n}{b_n+\ddots}}} \tag{1.4}$$

in a simpler form, we use the notation

$$b_0 + \frac{a_1}{b_1}\,{}_{+}\,\frac{a_2}{b_2}\,{}_{+}\,\frac{a_3}{b_3}\,{}_{+}\cdots{}_{+}\,\frac{a_n}{b_n}\,{}_{+}\cdots . \tag{1.5}$$

§2 Continued fraction expansion of e^z

It is well known that the exponential function e^z has a continued fraction expansion

$$e^z = \frac{1}{1} - \frac{z}{1} + \frac{z}{2} - \frac{z}{3} + \frac{z}{2} - \frac{z}{5} + \cdots + \frac{z}{2} - \frac{z}{(2j-1)} + \cdots \qquad (2.1)$$

and that the right hand side of (2.1) converges for any finite value of z in the complex z-plane. (See e.g. [2,p.348], [3,p.113].) If we define two sequences $\{F_k\}$ and $\{G_k\}$ by

$$F_0=1,\ F_1=1,\ G_0=0,\ G_1=1 \qquad (2.2)$$

$$\begin{cases} F_j = \begin{cases} (j-1)F_{j-1} - zF_{j-2} & ;\ j=2,4,6,\ldots \\ 2F_{j-1} + zF_{j-2} & ;\ j=3,5,7,\ldots \end{cases} & (2.3) \\ G_j = \begin{cases} (j-1)G_{j-1} - zG_{j-2} & ;\ j=2,4,6,\ldots \\ 2G_{j-1} + zG_{j-2} & ;\ j=3,5,7,\ldots \end{cases} & (2.4) \end{cases}$$

the quotient

$$H_n(z) \equiv G_n(z)/F_n(z) \qquad (2.5)$$

is identical to the n-th approximant of (2.1) [2,p.15], and converges uniformly in any finite domain of z [3,p.112]:

$$\lim_{n\to\infty} H_n(z) = e^z . \qquad (2.6)$$

By contraction [3,p.13] the expansion (2.1) is reduced to its odd part

$$e^z = 1 + \frac{2z}{2-z} + \frac{z^2}{6} + \frac{z^2}{10} + \cdots + \frac{z^2}{2(2j-1)} + \cdots , \qquad (2.7)$$

whose sequence of approximants is that of odd approximants $H_{2k+1}(z)=G_{2k+1}(z)/F_{2k+1}(z)$ of (2.1). The approximants H_{2k+1} can be generated by the recurrence relation

$$F_1=1,\ F_3=2-z,\ G_1=1,\ G_3=2+z \qquad (2.8)$$

$$\begin{cases} F_{2j+1}=2(2j-1)F_{2j-1}+z^2F_{2j-3}\ ;\ j=2,3,4,\ldots & (2.9) \\ G_{2j+1}=2(2j-1)G_{2j-1}+z^2G_{2j-3}\ ;\ j=2,3,4,\ldots & (2.10) \end{cases}$$

These odd approximants are found in the diagonal elements of the Pade table for e^z [4,s.16], and from (2.8) we see that $H_{2k+1}(z)$ satisfies

$$H_{2k+1}(-z) = \frac{1}{H_{2k+1}(z)} \qquad (2.11)$$

corresponding to $e^{-z}=1/e^z$.

In the similar way, we have the even approximants $H_{2k}(z)=G_{2k}(z)/F_{2k}(z)$ of (2.1) by the relation

$$F_0=1,\ F_2=1-z,\ G_0=0,\ G_2=1 \qquad (2.12)$$

$$\begin{cases} F_{2j}=\{2(2j-1)+\frac{2}{2j-3}z\}F_{2j-2}+\frac{2j-1}{2j-3}z^2F_{2j-4}\ ;\ j=2,3,4,\ldots & (2.13) \\ G_{2j}=\{2(2j-1)+\frac{2}{2j-3}z\}G_{2j-2}+\frac{2j-1}{2j-3}z^2G_{2j-4}\ ;\ j=2,3,4,\ldots & (2.14) \end{cases}$$

When $z\neq 0$, another expansion can be obtained by equivalence transformation [2,p.19]. If we multiply every odd terms of (2.1) by $s=1/z$, we have

$$e^{z} = \frac{1}{1} - \frac{1}{s} + \frac{1}{2} - \frac{1}{3s} + \frac{1}{2} - \cdots + \frac{1}{2} - \frac{1}{(2j-1)s} + \cdots \quad ; s=1/z , \tag{2.15}$$

the approximants $H_n(z)=G_n(z)/F_n(z)$ of which are generated by the recurrence relation

$$F_0=1,\ F_1=1,\ G_0=0,\ G_1=1 \tag{2.16}$$

$$F_j = \begin{cases} (j-1)sF_{j-1}-F_{j-2} & ; j=2,4,6,\ldots \\ 2F_{j-1}+F_{j-2} & ; j=3,5,7,\ldots \end{cases} \tag{2.17}$$

$$G_j = \begin{cases} (j-1)sG_{j-1}-G_{j-2} & ; j=2,4,6,\ldots \\ 2G_{j-1}+G_{j-2} & ; j=3,5,7,\ldots \end{cases} \tag{2.18}$$

The truncation error of the n-th approximant of the continued fraction (2.1) can be expressed in various forms. For example, if we write

$$e^{z} = \frac{1}{1} - \frac{z}{1} + \frac{z}{2} - \frac{z}{3} + \frac{z}{2} - \cdots - \frac{z}{2k-1} + \frac{z}{2} - \frac{z}{R_{2k+1}(z)} \tag{2.19}$$

and subtract

$$H_{2k+1}(z) = \frac{1}{1} - \frac{z}{1} + \frac{z}{2} - \frac{z}{3} + \frac{z}{2} - \cdots - \frac{z}{2k-1} + \frac{z}{2} \tag{2.20}$$

from (2.19), we have for the odd approximant

$$E_{2k+1}(z) = e^z - H_{2k+1}(z)$$

$$= \frac{(-1)^k z^{2k+1}}{F_{2k+1}(z)\{F_{2k+1}(z)R_{2k+1}(z)-zF_{2k}(z)\}} \tag{2.21}$$

$$= O(z^{2k+1}) , \tag{2.22}$$

where

$$R_{2k+1}(z) = (2k+1) + \frac{z}{2} - \frac{z}{2k+3} + \frac{z}{2} - \cdots . \tag{2.23}$$

For the even approximant we have

$$E_{2k}(z) = e^z - H_{2k}(z)$$

$$= \frac{(-1)^k z^{2k}}{F_{2k}(z)\{F_{2k}(z)R_{2k}(z)+zF_{2k-1}(z)\}} \tag{2.24}$$

$$= O(z^{2k}) , \tag{2.25}$$

where

$$R_{2k}(z) = 2 - \frac{z}{2k+1} + \frac{z}{2} - \frac{z}{2k+3} - \cdots . \tag{2.26}$$

As to the asymptotic behavior for large $|z|$, a difference is observed between those of $H_{2k+1}(z)$ and $H_{2k}(z)$. When n is odd, since the polynomials $F_{2k+1}(z)$ and $G_{2k+1}(z)$ are of the same order with the equal coefficients at the terms of the highest order, we have

$$\lim_{|z|\to\infty} H_{2k+1}(z) = 1 \tag{2.27}$$

so that for any $\varepsilon>0$

$$\lim_{|z|\to\infty} |E_{2k+1}(z)|=1, \text{ uniformly in } \frac{\pi}{2}+\varepsilon\leq\arg z\leq\frac{3\pi}{2}-\varepsilon \ . \qquad (2.28)$$

On the other hand, when n is even, $F_{2k}(z)$ is a polynomial of order 2k and $G_{2k}(z)$ is of order 2k-1, and hence

$$H_{2k}(z) = O(1/z) \quad (|z| \to \infty) \qquad (2.29)$$

so that

$$E_{2k}(z) = O(1/z) \quad (|z|\to\infty), \ \frac{\pi}{2} < \arg z < \frac{3\pi}{2} \ , \qquad (2.30)$$

or for any $\varepsilon>0$

$$\lim_{|z|\to\infty} |E_{2k}(z)|=0, \text{ uniformly in } \frac{\pi}{2}+\varepsilon\leq\arg z\leq\frac{3\pi}{2}-\varepsilon \ . \qquad (2.31)$$

As an example for the behavior of the error at intermediate value of z we show the map of $|E_n(z)|$, n=4 in Fig.1. The values of $|E_n(z)|$ on the negative real axis for various values of n are also shown in Fig.2. (See p.16.)

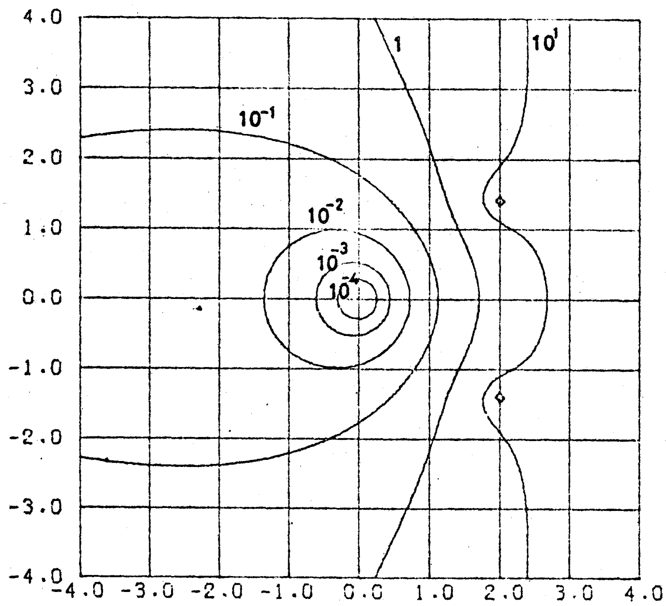

Fig.1. Contour map of $|E_4(z)|$

§3 Boundedness and regularity of the approximant in the left half-plane and on the imaginary axis

In this section we shall prove that the approximant $H_n(z)=G_n(z)/F_n(z)$ is bounded as $|H_n(z)|\leq 1$ in the left half-plane including the imaginary axis and hence is regular there. We use the following elementary property of the fractional linear transformation.

(a) If Re $s\leq 0$, the transformation

$$t = \frac{1}{(2j-1)s+w} \quad ; \; j=1,2,3,\ldots \tag{3.1}$$

maps the left half-plane Re $w\leq 1/2$ into all or a part of $|t-1|\geq 1$. In fact, from (3.1) $\mathrm{Re}[(2j-1)s+w]$ $=\mathrm{Re}[1/t]=(t+\bar{t})/(2t\bar{t})$, and since Re $s\leq 0$ and Re $w\leq 1/2$, we have $\mathrm{Re}[(2j-1)s+w]\leq 1/2$ so that $(t+\bar{t})/(t\bar{t})\leq 1$ or $|t-1|\geq 1$.

(b) The transformation

$$t = \frac{1}{2-w} \tag{3.2}$$

maps $|w-1|\geq 1$ onto Re $t\leq 1/2$. This would be evident from the relation $1\leq|1-w|^2=|1/t-1|^2=\{t\bar{t}-(t+\bar{t})+1\}/(t\bar{t})$.

Lemma If Re $z\leq 0$, $H_n(z)$ satisfies

$$|H_n(z)| \leq 1 \tag{3.3}$$

and is regular there.

Proof Since $s=1/z$ maps Re $z\leq 0$ onto itself, we take the expansion (2.15) instead of (2.1) and consider the images of Re $s\leq 0$. The continued fraction expansion (2.15) can be regarded to be given by composing the following fractional linear transformations:

$$w_0 = T_0[w_1] = \frac{1}{1-w_1} \tag{3.4}$$

$$\begin{cases} w_{2j-1} = T_{2j-1}[s;w_{2j}] = \dfrac{1}{(2j-1)s+w_{2j}} \; ; \; j=1,2,3,\ldots & (3.5) \\ w_{2j} = T_{2j}[w_{2j+1}] = \dfrac{1}{2-w_{2j+1}} \; ; \; j=1,2,3,\ldots & (3.6) \end{cases}$$

That is, the (2k+1)-th and the 2k-th approximants are given respectively

$$\begin{cases} H_{2k+1}(z) = T_0T_1T_2\cdots T_{2k}[0] \; ; \; k=1,2,3,\ldots & (3.7) \\ H_{2k}(z) = T_0T_1T_2\cdots T_{2k-1}[s;0] \; ; \; k=1,2,3,\ldots & (3.8) \end{cases}$$

First we take $H_{2k+1}(z)$ and consider the image of Re $s \leq 0$ by

$$w_{2k-1} = T_{2k-1}T_{2k}[0] = \frac{1}{(2k-1)s+1/2} \, . \tag{3.9}$$

From (a) we see that (3.9) maps Re $s \leq 0$ into a part of

$|w_{2k-1} - 1| \geq 1$. Then from (b)

$$w_{2k-2} = T_{2k-2}[w_{2k-1}] = \frac{1}{2-w_{2k-1}}$$

maps $|w_{2k-1} - 1| \geq 1$ into Re $w_{2k-2} \leq 1/2$. Successive and alternative use of (a) and (b) leads to $|w_1 - 1| \geq 1$, where $w_1 = T_1[w_2] = T_1T_2\cdots T_{2k}[0]$. Hence from (3.4) we finally have

$$|H_{2k+1}(z)| = |T_0 T_1 \cdots T_{2k}[0]| = |w_0| = 1/|1-w_1| \le 1.$$

Next we consider $H_{2k}(z)$. In this case Re $s \le 0$ is mapped by

$$w_{2k-1} = T_{2k-1}[s;0] = \frac{1}{(2k-1)s}$$

onto Re $w_{2k-1} \le 0$, and this is entirely included in the region $|w_{2k-1} - 1| \ge 1$. Then from the above proof for $H_{2k+1}(z)$, we can immediately conclude that $|H_{2k}(z)| = |w_0| \le 1$. Finally, since $H_n(z)$ is a rational function of z, $|H_n(z)| \le 1$ over Re $z \le 0$ implies the regularity of $H_n(z)$ over Re $z \le 0$.

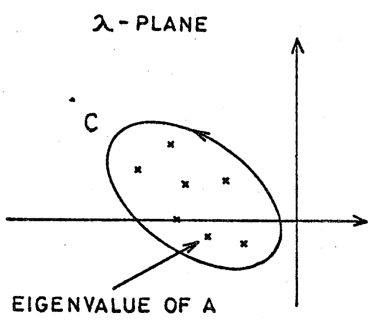

Fig.3. Path C of Dunford integral (4.6)

§4 High order iterative approximation for exp tA

For the approximation of exp tA where A is an N×N matrix, we are ready to make use of the recurrence relation (2.2)-(2.4). The replacement of z by the matrix tA leads formally to the following iterative procedure for the approximation of exp tA:

$$F_0=I,\ F_1=I,\ G_0=0,\ G_1=I \quad (I\text{:identity matrix}) \tag{4.1}$$

$$F_j=\begin{cases}(j-1)F_{j-1}-tAF_{j-2} & ;\ j=2,4,6,\ldots\\ 2F_{j-1}+tAF_{j-2} & ;\ j=3,5,7,\ldots\end{cases} \tag{4.2}$$

$$G_j=\begin{cases}(j-1)G_{j-1}-tAG_{j-2} & ;\ j=2,4,6,\ldots\\ 2G_{j-1}+tAG_{j-2} & ;\ j=3,5,7,\ldots\end{cases} \tag{4.3}$$

$$H_n(tA) = F_n^{-1}(tA)G_n(tA) \fallingdotseq \exp tA \tag{4.4}$$

We adopt a certain norm for N×N matrix. Then as to the convergence of $H_n(tA)$ to exp tA, we have

Theorem . If every eigenvalue λ_ℓ of a matrix A satisfies Re $\lambda_\ell \le 0$, then

$$\lim_{n\to\infty} H_n(tA) = \exp tA\ ,\ t\ge 0 \tag{4.5}$$

Proof Since $H_n(z)$ is regular over Re $z\le 0$ from Lemma, $E_n(t\lambda)=\exp t\lambda - H_n(t\lambda)$ is also regular over Re $\lambda\le 0$ when $t\ge 0$. Then the error $E_n(tA)$ can be expressed in terms of Dunford integral [5,p287]:

$$E_n(tA) = \exp tA - H_n(tA)$$

$$= \frac{1}{2\pi i} \oint_C \frac{1}{\lambda - A} E_n(t\lambda)\, d\lambda \quad . \tag{4.6}$$

The path C of the integral is a simple closed contour enclosing all of the eigen values λ_ℓ of A and not enclosing any singularity of $E_n(t\lambda)$ as shown in Fig.3. Taking the norm of (4.6) we have

$$\|E_n(tA)\| \le \frac{1}{2\pi} \oint_C \|(\lambda - A)^{-1}\| \, |E_n(t\lambda)| \, |d\lambda|$$

$$\le \frac{1}{2\pi} \{\max_C |E_n(t\lambda)|\} \oint_C \|(\lambda - A)^{-1}\| \, |d\lambda| \quad . \tag{4.7}$$

Since $\lambda - A$ is regular along C, the integral $\oint \|(\lambda - A)^{-1}\| \, |d\lambda|$ along C is bounded, and hence in view of the uniform convergence of $|E_n(t\lambda)|$ to zero as $n \to \infty$ as a scalar function over any finite domain in the λ-plane we have $\|E_n(tA)\| \to 0$ as $n \to \infty$, which completes the proof.

By making use of a vector

$$g_j = G_j u_0 \tag{4.8}$$

instead of the matrix G_j itself when calculating $(\exp tA)u_0$, we can reduce the product between two matrices into that between a matrix and a vector as follows:

$$F_0 = I, \; F_1 = I, \; g_0 = 0, \; g_1 = u_0 \tag{4.9}$$

$$F_j = \begin{cases} (j-1)F_{j-1} - tAF_{j-2} \; ; \; j = 2,4,6,\ldots \\ 2F_{j-1} + tAF_{j-2} \; ; \; j = 3,5,7,\ldots \end{cases} \tag{4.10}$$

110

$$g_j = \begin{cases} (j-1)g_{j-1} - tAg_{j-2} \ ; & j=2,4,6,\ldots \\ 2g_{j-1} + tAg_{j-2} \ ; & j=3,5,7,\ldots \end{cases} \tag{4.11}$$

$$u(t) = (\exp\ tA)u_0 \fallingdotseq F_n^{-1} g_n \ . \tag{4.12}$$

If A^2 is a priori calculated, we have another procedure from (2.8)-(2.10) that makes, though theoretically, double the rate of convergence of the above procedure:

$$F_1 = I,\ F_3 = 2I - tA,\ g_1 = u_0,\ g_3 = 2u_0 + tAu_0 \tag{4.13}$$

$$F_{2j+1} = 2(2j-1)F_{2j-1} + t^2A^2F_{2j-3} \ ; \ j=2,3,4,\ldots \tag{4.14}$$

$$g_{2j+1} = 2(2j-1)g_{2j-1} + t^2A^2g_{2j-3} \ ; \ j=2,3,4,\ldots \tag{4.15}$$

$$u(t) = (\exp\ tA)u_0 \fallingdotseq F_{2k+1}^{-1} g_{2k+1} \ . \tag{4.16}$$

When A^{-1} is obtainable, we may have other procedures by replacing s by $t^{-1}A^{-1}$ in (2.17) and (2.18), and, if preferable, by reducing it into contracted forms.

We assume that every eigenvalue λ_ℓ of N×N matrix A lies in the left half-plane, i.e. Re $\lambda_\ell<0;\ell=1,2,\ldots,N$. Then it can easily be seen from the proof of Lemma that the spectral radius ρ of $H_n(tA)$ satisfies $\rho(H_n(tA))<1$ for all $t>0$, and hence the matrix approximation $H_n(tA)$ under the above assumption is unconditionally stable for any n [6,p.265]. It would be clear that $H_n(tA)$ is a consistent approximation to exp tA in the sense of Lax and Richtmyer [7,p.271].

§5 Discussions

The present method has the advantages of a simple iterative procedure and of a high order stable approximation. It would yield a result with high precision even when it is applied with a fairly large time mesh t owing to the rapid convergence of the continued fraction expansion, and hence this situation is considered to recover the disadvantage of the method that it requires one matrix product for every one iteration.

It should be noted, however, that a serious situation may arise at the actual computation when the maximum $t|\lambda_M|$ of the absolute value of the eigenvalues of the matrix tA is too large compared with 1 while the minimum is less than 1 as in the case of a parabolic problem with fairly large t, since then the condition number of

$$F_n(tA) = \begin{cases} I - \cdots +(-1)^k \frac{(k-1)!}{(2k-1)!}(tA)^k & ; n=2k \\ \\ I - \cdots +(-1)^k \frac{k!}{(2k)!}(tA)^k & ; n=2k+1 \end{cases} \tag{5.1}$$

[1,p.223] becomes remarkably large as n is increased to an appropriate value for convergence, resulting in a seriously large error in the solution $F_n^{-1}(tA)g_n$. This drawback may be recovered if the eigenvalues of A are shifted to the left by multiplying $\exp(-\sigma t)$ $(\sigma \simeq |\lambda_M|)$ to $\exp(tA)$ so that the condition number of $A-\sigma$ may be reduced to the order of nearly unity, but then the convergence would turn out to be very slow. When A is a diagonal dominant sparse matrix as is obtained from a parabolic equation, the factorization of $F_n^{-1}(tA)$ into

$$F_n^{-1}(tA) = (tA-\mu_1)^{-1}(tA-\mu_2)^{-1}\cdots(tA-\mu_n)^{-1} \qquad (5.2)$$

will be very efficient.

The following procedure will generally be recommended. Divide t into equal and small n subintervals Δt, i.e. $t=n\Delta t$, and compute $F_k(\Delta tA)$ for fixed value of Δt to an appropriate order k. Then, using $F_k(\Delta tA)$, iterate

$$u(j\Delta t) = F_k^{-1}(\Delta tA)u((j-1)\Delta t), \ j=1,2,\ldots,n \qquad (5.3)$$

with the initial value $u_0=u(0)$. This method would be applicable with slight modification to obtain an approximate solution of

$$\frac{du}{dt} = A(t)u\ , \qquad (5.4)$$

where $A(t)$ depends on t moderately, if we use the matrix $A(j\Delta t)$ in the calculation at the subinterval $j\Delta t<t\le(j+1)\Delta t$.

The present analysis may be formally extended to the approximation of exp tA in a Banach space X in which A is such a closed linear operator on X into X that the spectrum lies in the left half-plane including the imaginary axis and that the Dunford integral representation holds in $E_n(tA)$. When A is a bounded operator the extension is immediate. When A is unbounded, however, some additional conditions must be satisfied. For example, such an operator that $\|(\lambda-A)^{-1}\| \le M(|\lambda|+1)^{-1}$ holds for λ in the resolvent set $\rho(A)$ in a sector $\pi/2+\varepsilon\le\arg\lambda\le 3\pi/2-\varepsilon, \varepsilon>0$ comes within this class of operators, if we use the even approximants $H_{2k}(tA)$ in view of (2.30).

Acknowledgement

The author wishes to express his sincere thanks to Professors Hidetosi Takahasi and Hiroshi Fujita for their valuable advices and discussions. Thanks are also due to Professors P.D.Lax and Teruo Ushijima for their valuable comments.

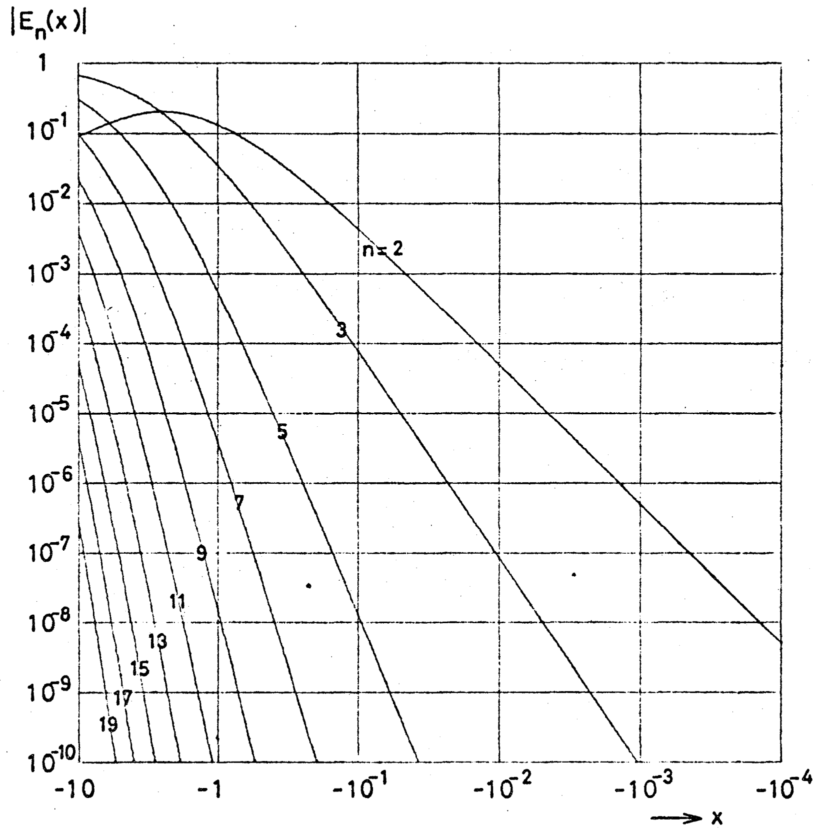

Fig.2. Absolute errors $|E_n(z)|$ of continued fraction expansion of exp z on the negative real axis

References

1. R.S.Varga, On higher order stable implicit methods for solving parabolic partial differential equations, J.Math. and Phys. 40 (1961) 220-231.

2. H.S.Wall, Analytic theory of continued fractions, Chelsea, 1948.

3. A.N.Khovanskii, The application of continued fractions and their generalizations to problems in approximation theory, Noordhoff, 1963.

4. M.H.Padé, Sur la représentation approchée d'une fonction par des fractions rationnelles, Ann.de l'Ec.Normale. 3[e] série, 9 (1892).

5. A.E.Taylor, Introduction to functional analysis, John-Wiley, 1958.

6. R.S.Varga, Matrix iterative analysis, Prentice-Hall, 1962.

7. P.D.Lax and R.D.Richtmyer, Survey of the stability of linear finite difference equation, Comm.pure appl.Math. 9 (1956) 267-293.