Approximation of Exponential Function of a Matrix by Continued Fraction Expansion

By

Masatake Mori

Abstract

A numerical method for high order approximation of $u(t) = \exp(tA)u_0$, where A is an $N \times N$ matrix and u_0 is an N dimensional vector, based on the continued fraction expansion of $\exp z$ is given. The approximants $H_k(z)$ of the continued fraction expansion of $\exp z$ are shown to satisfy $|H_k(z)| \le 1$ for Re $z \le 0$, which results in an unconditionally stable method when every eigenvalue of A lies in the left half-plane or on the imaginary axis.

§1. Introduction

The solution of an equation of evolution

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

in which A is an $N \times N$ matrix is formally given by

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

where the matrix $\exp tA$ is defined by

(1.3)
$$e^{tA} = I + \frac{t}{1!}A + \frac{t^2}{2!}A^2 + \cdots$$

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

Received December 18, 1973.

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 $N \times N$ 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 with 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 $|H_k(z)| \le 1$ in Re $z \le 0$ 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

in a simpler form, we use the notation

(1.5)
$$b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \frac{a_3}{b_3} + \dots + \frac{a_n}{b_n} + \dots$$

§2. Continued Fraction Expansion of e^{*}

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

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

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

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

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

(2.4)
$$\begin{cases} (j-1)G_{j-1} - zG_{j-2}; & j=2, 4, 6, \dots \\ 2G_{j-1} + zG_{j-2}; & j=3, 5, 7, \dots, \end{cases}$$

the quotient

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

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]:

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

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

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

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

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

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

(2.10)
$$G_{2j+1} = 2(2j-1)G_{2j-1} + z^2G_{2j-3}; \qquad j=2, 3, 4,...$$

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

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

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

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

(2.13)
$$\begin{cases} F_{2j} = \left\{ 2(2j-1) + \frac{2}{2j-3}z \right\} F_{2j-2} + \frac{2j-1}{2j-3}z^2 F_{2j-4}; \ j=2, 3, 4, \dots \end{cases}$$

(2.14)
$$\left(G_{2j} = \left\{ 2(2j-1) + \frac{2}{2j-3}z \right\} G_{2j-2} + \frac{2j-1}{2j-3}z^2 G_{2j-4}; \ j=2, 3, 4, \dots \right.$$

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

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

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

$$(2.16) F_0 = 1, F_1 = 1, G_0 = 0, G_1 = 1$$

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

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

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

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

and subtract

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

from (2.19), we have the error of the odd approximant:

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

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

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

where

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

For the even approximant we have

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

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

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

where

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

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

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

so that for any $\varepsilon > 0$

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

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

MASATAKE MORI

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

so that

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

or for any $\varepsilon > 0$



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

Approximation of Exponential Function of a Matrix

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

As an example for the behavior of the error at intermediate value of z we show the contour 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.

§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 in such a way that $|H_n(z)| \le 1$ in the left half-plane including the imaginary axis and hence is regular there. We use some elementary relations satisfied by the following fractional linear transformation.

(a) If
$$\text{Res} \leq 0$$
, the transformation

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

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

(b) The transformation

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

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

Theorem 1. If $\operatorname{Re} z \leq 0$, $H_n(z)$ satisfies

$$(3.3) |H_n(z)| \le 1$$

and is regular there.

Proof. Since s=1/z maps $\operatorname{Re} z \leq 0$ onto itself, we take the examplion (2.15) instead of (2.1) and consider the images of $\operatorname{Re} s \leq 0$. The con-

tinued fraction expansion (2.15) can be regarded to be given by composing the following fractional linear transformations:

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

(3.5)
$$\int w_{2j-1} = T_{2j-1}[s; w_{2j}] = \frac{1}{(2j-1)s + w_{2j}}; \quad j = 1, 2, 3, \dots$$

(3.6)
$$w_{2j} = T_{2j} [w_{2j+1}] = \frac{1}{2 - w_{2j+1}}; \quad j = 1, 2, 3, \dots$$

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

(3.7)
$$(H_{2k+1}(z) = T_0 T_1 T_2 \dots T_{2k}[0]; \qquad k = 1, 2, 3, \dots$$

(3.8)
$$H_{2k}(z) = T_0 T_1 T_2 \dots T_{2k-1}[s; 0]; \quad k = 1, 2, 3, \dots$$

First we take $H_{2k+1}(z)$ and consider the image of Res ≤ 0 by

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

From (a) we see that (3.9) maps $\operatorname{Re} s \le 0$ into a part of $|w_{2k-1}-1| \ge 1$. Then from (b)

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

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

$$|H_{2k+1}(z)| = |T_0T_1...T_{2k}[0]| = |w_0| = 1/|1-w_1| \le 1.$$

Next we consider $H_{2k}(z)$. In this case Res ≤ 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 $z \le 0$.

§4. High Order Iterative Approximation for $\exp tA$

For the approximation of $\exp tA$ where A is an $N \times N$ 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 $\exp tA$.

(4.1) $F_0 = I$, $F_1 = I$, $G_0 = 0$, $G_1 = I$ (*I*: identity matrix)

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

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

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

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

Theorem 2. Let A be a square matrix of finite dimension. Then for any finite t

(4.5)
$$\lim_{n \to \infty} H_n(tA) = \exp tA$$

Proof. Since $H_n(z)$ converges uniformly to e^z over any finite domain in the z-plane, $H_n(z)$ is regular on any finite domain for sufficiently large *n*, so that the error $E_n(tA)$ can be expressed in terms of Dunford integral [5, p. 287]:

(4.6)
$$E_n(tA) = \exp tA - H_n(tA)$$
$$= \frac{1}{2\pi i} \oint_C \frac{1}{\lambda - A} E_n(t\lambda) d\lambda.$$

The path C of the integral is a simple closed contour enclosing all

of the eigen values λ_i 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



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

(4.7)
$$\|E_n(tA)\| \leq \frac{1}{2\pi} \oint_C \|(\lambda - A)^{-1}\| \|E_n(t\lambda)\| \|d\lambda\|$$
$$\leq \frac{1}{2\pi} \{\max_C \|E_n(t\lambda)\| \} \oint_C \|(\lambda - A)^{-1}\| \|d\lambda\|.$$

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

$$(4.8) g_j = G_j u_0$$

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:

(4.9)
$$F_0 = I, \quad F_1 = I, \quad g_0 = 0, \quad g_1 = u_0$$

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

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

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

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:

(4.13)
$$F_1 = I, \quad F_3 = 2I - tA, \quad g_1 = u_0, \quad g_3 = 2u_0 + tAu_0$$

(4.14)
$$\begin{cases} F_{2j+1} = 2(2j-1)F_{2j-1} + t^2 A^2 F_{2j-3}; & j=2, 3, 4, \dots \end{cases}$$

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

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

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 $N \times N$ matrix A lies in the left half-plane, i.e. Re $\lambda_l < 0$; l=1, 2, ..., N. Then it can easily be seen from the proof of Theorem 1 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 regular and 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 MASATAKE MORI

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

[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

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

may be 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

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

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

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

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 professor Hidetosi Takahasi and to professor Hiroshi Fujita for their valuable advices and discussions. Thanks are also due to Professor Teruo Ushijima for his valuable comments.

References

- Varga, R.S., On higher order stable implicit methods for solving parabolic partial differential equations, J. Math. and Phys. 40 (1961) 220-231.
- [2] Wall, H.S., Analytic theory of continued fractions, van Nostrand, 1948, 2nd ed. Chelsea, 1967.
- [3] Khovanskii, A.N., The application of continued fractions and their generalizations to problems in approximation theory, (English translation) Noordhoff, 1963.
- [4] Pabé, M. H., Sur la représentation approchée d'une fonction par des fractions rationnelles, Ann. de l'Ec. Normale. 3^e série, **9** (1892).
- [5] Taylor, A.E., Introduction to functional analysis, John-Wiley, 1958.
- [6] Varga, R.S., Matrix iterative analysis, Prentice-Hall, 1962.
- [7] Lax, P.D. and Richtmyer, R.D., Survey of the stability of linear finite difference equation, Comm. pure appl. Math. 9 (1956) 267-293.