

ATTAINABLE ORDER OF RATIONAL APPROXIMATIONS TO THE EXPONENTIAL FUNCTION WITH ONLY REAL POLES

SYVERT P. NØRSETT and ARNE WOLFBRANDT

Abstract.

Rational approximations of the form $\sum_{i=0}^m a_i q^i / \prod_{i=1}^n (1 + \gamma_i q)$ to $\exp(-q)$, $q \in C$, are studied with respect to order and error constant. It is shown that the maximum obtainable order is $m+1$ and that the approximation of order $m+1$ with least absolute value of the error constant has $\gamma_1 = \gamma_2 = \dots = \gamma_n$. As an application it is shown that the order of a ν -stage semi-implicit Runge-Kutta method cannot exceed $\nu+1$.

1. Introduction.

Different approximations to the exponential function play an important role in connection with the numerical solution of stiff systems of ordinary differential equations. For example, entries in the Padé table for $\exp(-q)$ are connected with implicit Runge-Kutta methods of optimal order, Ehle [1], Chipman [2]. Further multiple Padé approximations¹ are, Nørsett [3], related to the special semi-implicit Runge-Kutta methods of Nørsett [4] and the special one-step Hermite methods of Nørsett [5].

The Padé approximations $R_n^m(q)$ to $\exp(-q)$, $q \in C$, are such that the coefficients of the denominator and the numerator are chosen to give optimal order $n+m$. However, the zeros of the denominator are all complex except for one when n is an odd number. In connection with, for example, semi-implicit Runge-Kutta methods, the relevant rational approximations to $\exp(q)$ are such that the denominator has only real zeros. The natural question to be discussed is therefore, what is the optimal order of such rational approximations? This problem was treated in Nørsett [3] for the case where the denominator was of one of the forms $(1 + \gamma q)^n$ and $(1 + \gamma q)^{n-1}(1 + \delta q)$ and the numerator a polynomial of degree $m \leq n$. The case of n different real zeros in the denominator and a polynomial of degree $n-1$ in the numerator is discussed in Wolf-

Received Oct. 22, 1976.

¹ In Siemieniuch [6] called Nørsett approximations.

brandt [8]. In this paper the general case of n distinct real factors in the denominator and a polynomial of degree m in the numerator is discussed thereby extending the results of Nørsett [3] and Wolfbrandt [8]. The surprising result is that the maximum reachable order is $m + 1$. Hence the error is of the form $Cq^{m+2} + O(q^{m+3})$ where C , the error constant, is a function of $n - 1$ real variables. By minimising C with respect to these variables the next main result is obtained, saying that the least value of C is obtained when all the zeros of the denominator are equal, which means that the optimal approximations in the sense of minimising C are the multiple Padé approximations of Nørsett [3].

2. N -approximations and C -polynomials.

Let $p \in \pi_n = \{\text{polynomials of degree } \leq n, p^{(n)}(x) = 1\}$. Then, (1.2) in Nørsett [7] (q is replaced by $-q$),

$$(2.1) \quad R_n^m(p; q) = \frac{\sum_{i=0}^m g_{i-n}(0)q^i}{\sum_{i=0}^n g_{i-n}(1)q^i} = e^{-q} + O(q^{m+1}),$$

where we have defined

$$(2.2) \quad \begin{cases} g_0(x) = p(x), \\ g_{i+1}(x) = \int_1^x g_i(t) dt, & i \geq 0 \\ g_{-i}(x) = g_0^{(i)}(x), & i \geq 0. \end{cases}$$

p is called the C -polynomial for the rational approximation $R_n^m(p; q)$. Following Nørsett [7], $R_n^m(p; q)$ is an approximation to $\exp(-q)$ of order $s > m$ iff

$$(2.3) \quad \begin{cases} g_{i-n+1}(0) = 0 & \text{for } m \leq i \leq s-1 \\ g_{s-n+1}(0) \neq 0. \end{cases}$$

DEFINITION 2.1 A rational approximation to $\exp(-q)$ whose denominator has only real zeros is called an N -approximation to $\exp(-q)$.

LEMMA 2.1. *Let the denominator of a rational approximation to $\exp(-q)$ be of the form*

$$(2.4) \quad D_n(q) = \hat{D}_{n-k}(q)(1 + \gamma q)^k,$$

where

$$(2.5) \quad \hat{D}_{n-k}(q) = \sum_{i=0}^{n-k} \hat{p}^{(n-k-i)}(1)q^i, \quad \hat{p} \in \pi_{n-k}.$$

Then

$$(2.6) \quad D_n(q) = \sum_{i=0}^n p^{(n-i)}(1)q^i$$

with

$$(2.7) \quad p(t) = \sum_{j=0}^k \binom{k}{j} \gamma^{k-j} g_j(t),$$

where g_j is defined by (2.2) and $g_0 = \hat{p}$.

PROOF. By induction on k . Let $k=1$, then

$$(2.8) \quad \begin{aligned} D_n(q) &= [\sum_{i=0}^{n-1} \hat{p}^{(n-i-1)}(1)q^i](1+\gamma q) \\ &= \hat{p}^{(n-1)}(1) + \sum_{i=1}^{n-1} (\hat{p}^{(n-i-1)}(1) + \gamma \hat{p}^{(n-i)}(1))q^i + \gamma \hat{p}(1)q^n. \end{aligned}$$

Using p from (2.7) with $k=1$,

$$(2.9) \quad p(t) = \gamma \hat{p}(t) + \int_1^t \hat{p}(x) dx = \gamma g_0(t) + g_1(t)$$

which gives (2.6) from (2.8).

Suppose that the lemma is true for $k \leq k_0 - 1$. For $k = k_0$,

$$(2.10) \quad D_n(q) = \hat{D}_{n-k_0}(q)(1+\gamma q)^{k_0-1}(1+\gamma q) = \check{D}_{n-k_0}(q)(1+\gamma q),$$

where by the induction hypothesis

$$(2.11) \quad \check{D}_{n-k_0}(q) = \sum_{i=0}^{n-1} \tilde{p}^{(n-1-i)}(1)q^i$$

and

$$\tilde{p}(t) = \sum_{j=0}^{k_0-1} \binom{k_0-1}{j} \gamma^{k_0-1-j} g_j(t).$$

From (2.9), (2.10) and (2.11)

$$\begin{aligned} p(t) &= \gamma \tilde{p}(t) + \int_1^t \tilde{p}(x) dx \\ &= \sum_{j=0}^{k_0-1} \binom{k_0-1}{j} \gamma^{k_0-j} g_j(t) + \sum_{j=0}^{k_0-1} \binom{k_0-1}{j} \gamma^{k_0-1-j} g_{j+1}(t) \\ &= \sum_{j=0}^{k_0} \binom{k_0}{j} \gamma^{k_0-j} g_j(t). \end{aligned}$$

Let $R_n^m(p, q)$ be an N -approximation to $\exp(-q)$ of the form

$$(2.12) \quad R_n^m(p, q) = Q_m(q) / \prod_{j=1}^n (1 + \gamma_j q), \quad \gamma_j \geq 0, \quad Q_m \in \pi_m.$$

The corresponding C -polynomial $p = p_{n,m}$ is then a function of $\gamma_1, \dots, \gamma_n$, $p_{n,m}(t) = p_{n,m}(\gamma_1, \dots, \gamma_n; t)$ where $p_{n,m}$ is a symmetric function in its variables $\gamma_1, \dots, \gamma_n$. Let $g_k(\gamma_1, \dots, \gamma_n; t)$ be defined as in (2.2) with $p = p_{n,m}$. Using Lemma 2.1 we obtain,

$$(2.13) \quad \begin{aligned} g_k(\gamma_1, \dots, \gamma_n; t) &= \gamma_i g_k(\gamma_1, \dots, \gamma_{i-1}, \gamma_{i+1}, \dots, \gamma_n; t) \\ &\quad + g_{k+1}(\gamma_1, \dots, \gamma_{i-1}, \gamma_{i+1}, \dots, \gamma_n; t), \quad 1 \leq i \leq n \end{aligned}$$

for $k \geq -n+1$. According to (2.3) R_n^m in (2.12) has order $m+2$ if and only if

$$(2.14) \quad \begin{cases} g_{m-n+1}(\gamma_1, \dots, \gamma_n; 0) = 0 \\ g_{m-n+2}(\gamma_1, \dots, \gamma_n; 0) = 0. \end{cases}$$

Define $h_k(z_1, \dots, z_n; t)$ by

$$(2.15) \quad h_k(z_1, \dots, z_n; t) = g_k(\gamma_1, \dots, \gamma_n; t) / (\prod_{i=1}^n \gamma_i), \quad z_i = \gamma_i^{-1}.$$

From (2.13),

$$(2.16) \quad h_k(z_1, \dots, z_n; t) = h_k(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n; t) + z_i h_{k+1}(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n; t), \quad 1 \leq i \leq n.$$

Obviously h_k is a symmetric function of its z -arguments. Order $m+2$ is then equivalent with (by (2.14))

$$(2.17) \quad \begin{cases} h_{m-n+1}(z_1, \dots, z_n; 0) = 0 \\ h_{m-n+2}(z_1, \dots, z_n; 0) = 0. \end{cases}$$

LEMMA 2.2. For $n \geq 1$ the following identity holds ($k \geq -n+1$):

$$(2.18) \quad \sum_{j=1}^n (\partial/\partial z_j) h_k(z_1, \dots, z_n; t) = (n+k+1)h_{k+1}(z_1, \dots, z_n; t) - (t-1)h_k(z_1, \dots, z_n; t).$$

PROOF. By induction. By definition of C -polynomials,

$$g_0(\gamma_1; t) = \gamma_1 + (t-1).$$

Hence, using (2.2),

$$g_{-1} = 1, \quad g_k = \gamma_1(t-1)^k/k! + (t-1)^{k+1}/(k+1)!, \quad k \geq 0,$$

and therefore,

$$(2.19) \quad h_{-1}(z_1; t) = z_1. \\ h_k(z_1, t) = (t-1)^k/k! + z_1(t-1)^{k+1}/(k+1)!, \quad k \geq 0.$$

(2.18) then follows for $n=1$.

Suppose that (2.18) is satisfied for $n \leq m$. Using (2.16) and (2.18) for $n=m$ we have

$$\begin{aligned} \sum_{j=1}^{m+1} (\partial/\partial z_j) h_k(z_1, \dots, z_{m+1}; t) &= h_{k+1}(z_1, \dots, z_m; t) \\ &+ \sum_{j=1}^m (\partial/\partial z_j) h_k(z_1, \dots, z_m; t) + z_{m+1} \sum_{j=1}^m (\partial/\partial z_j) h_{k+1}(z_1, \dots, z_m; t) \\ &= h_{k+1}(z_1, \dots, z_m; t) + (m+k+1)h_{k+1}(z_1, \dots, z_m; t) \\ &- (t-1)h_k(z_1, \dots, z_m; t) \\ &+ z_{m+1}(m+k+2)h_{k+2}(z_1, \dots, z_m; t) - z_{m+1}(t-1)h_{k+1}(z_1, \dots, z_m; t) \\ &= (m+k+2)[h_{k+1}(z_1, \dots, z_m; t) + z_{m+1}h_{k+2}(z_1, \dots, z_m; t)] \\ &- (t-1)[h_k(z_1, \dots, z_m; t) + z_{m+1}h_{k+1}(z_1, \dots, z_m; t)] \\ &= (m+k+2)h_{k+1}(z_1, \dots, z_{m+1}; t) - (t-1)h_k(z_1, \dots, z_{m+1}; t). \quad \blacksquare \end{aligned}$$

LEMMA 2.3. The following two statements are equivalent,

$$(2.20) \quad (\partial/\partial z_j) h_k(z_1, \dots, z_n; 0) = 0, \quad h_k(z_1, \dots, z_n; 0) = 0$$

and

$$(2.21) \quad \begin{cases} h_k(z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n; 0) = 0 \\ h_{k+1}(z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n; 0) = 0 \end{cases} \quad 1 \leq j \leq n.$$

PROOF: By observing that

$$(2.22) \quad (\partial/\partial z_j)h_k(z_1, \dots, z_n; t) = h_{k+1}(z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_n; t),$$

the lemma is an easy consequence of (2.16). ■

The main result is then,

THEOREM 2.1. *The maximum attainable order of an N -approximation to $\exp(-q)$ of the form (2.12) is $m + 1$.*

(This result has been obtained when $m = n - 1$ by Wolfbrandt [8]).

PROOF. By induction on n we show that order $m + 2$ is not attainable.

(i) Let $n = 1$. From (2.17) and (2.19) for order $m + 2$,

$$\begin{aligned} (-1)^m/m! + z_1(-1)^{m+1}/(m+1)! &= 0 \\ (-1)^{m+1}/(m+1)! + z_1(-1)^{m+2}/(m+2)! &= 0 \end{aligned}$$

or $z_1 = m + 1$ and $z_1 = m + 2$ and it is impossible to satisfy both equations at the same time.

(ii) Suppose that the theorem is true for $n \leq r - 1$.

(iii) Let $n = r$. The equations to be satisfied for order $m + 2$ are by (2.17) and lemma 2.2 given by

$$(2.23a) \quad h_k(z_1, \dots, z_r; 0) = 0, \quad k = m - r + 1$$

$$(2.23b) \quad \sum_{j=1}^r (\partial/\partial z_j)h_k(z_1, \dots, z_r; 0) = 0.$$

With (ii) in mind, (2.23a) gives

$$(2.24) \quad (\partial/\partial z_j)h_k(z_1, \dots, z_r; 0) \neq 0, \quad j = 1(1)r.$$

The proof is completed if we can verify that the quantities in (2.24) all have the same sign under the condition (2.23a). In this connection we first observe that the domain

$$\Omega = \{(z_1, \dots, z_r) \in R^r; h_k(z_1, \dots, z_r; 0) = 0\}$$

can be split up in exactly $\hat{r} = \begin{cases} r & \text{if } k \geq -1 \\ r+k+1 & \text{if } -r+1 \leq k < -1 \end{cases}$

subdomains such that to each subdomain belongs a point $\underbrace{(z, \dots, z)}_{r\text{-times}}$, where z is a solution to

$$(2.25) \quad \begin{cases} L_{r+k}^{(k)}(z) = 0 & \text{for } k \geq 0 \\ z^k L_r^{(-k)}(z) = 0 & \text{for } -r+1 \leq k \leq 0 \end{cases}$$

and L_n is the Laguerre polynomial of degree n . This observation follows by induction from the fact that $(z_1, \dots, z_r) \in \Omega$ implies

$$z_r = -h_k(z_1, \dots, z_{r-1}; 0)/h_{k+1}(z_1, \dots, z_{r-1}; 0),$$

and (by (ii)) it is impossible to have $h_k(z_1, \dots, z_{r-1}; 0) = h_{k+1}(z_1, \dots, z_{r-1}; 0) = 0$. Further, when $z_1 = \dots = z_{r-1} = z > 0$,

$$z_r = \begin{cases} -(r+k)L_{r+k-1}^{(k)}(z)/L_{r+k}^{(k+1)}(z), & k \geq 0 \\ -zL_{r-1}^{(-k)}(z)/L_{r-1}^{(-k-1)}(z), & -r+2 \leq k \leq -1 \end{cases}$$

which shows that the number of subdomains is at least \hat{r} (the zeros of the nominator and denominator separate each other).

For each such subdomain of Ω ,

- a) $(\partial/\partial z_j)h_k(z_1, \dots, z_r; 0) \neq 0$ (from (2.24))
 - b) $(\partial/\partial z_j)h_k(z_1, \dots, z_r; 0)$ is continuous.
 - c) $(z_1, \dots, z_r) \in \Omega$ with $(\partial/\partial z_j)h_k(z_1, \dots, z_r; 0)$, $j = 1(1)r$, of the same sign.
- r -times

It is then obvious that the quantities in (2.24) all have the same sign. ■

3. The error constant.

When we are dealing with rational approximations $R_n^m(q)$ of order s to $\exp(-q)$ for $|q| \ll 1$, the size of the *error constant*, $C_{n,s+1}^m$, is of interest. Asymptotically we have

$$(3.1) \quad R_n^m(q) = \exp(-q) + C_{n,s+1}^m q^{s+1} + O(q^{s+2}).$$

If the order is at least m , then

$$(3.2) \quad C_{n,m+2}^m = g_{m-n+2}(0),$$

with $g_0(t) = p(t)$, p the C -polynomial for R_n^m . Having established in theorem 2.1 that the highest order obtainable for the N -approximation (2.12) to $\exp(-q)$ is $m+1$, the natural point to consider is the problem of minimizing $|C_{n,m+2}^m| = |g_{m-n+2}(0)|$. In this respect the unexpected result is,

THEOREM 3.1. *Let the order of the N -approximation R_n^m in (2.12) be $m+1$. The minimum value of $|g_{m-n+2}(0)|$ is then obtained with $\gamma_1 = \gamma_2 = \dots = \gamma_n$.*

PROOF. The polynomial $w_r \in \pi_r$ is defined by

$$(3.3) \quad \prod_{i=1}^r (1 + \gamma_i q) = \sum_{i=0}^r w_r^{(r-i)}(1) q^i.$$

From (2.7) in Lemma 2.1 (with $k=1$)

(3.4)

$$g_0^{r+1}(t) \equiv w_{r+1}(t) = \gamma_{r+1}w_r(t) + \int_1^t w_r(x)dx = \gamma_{r+1}g_0^r(t) + g_1^r(t), \quad r \geq 0.$$

Hence, $p(t) = g_0^n(t)$ and order $m+1$ is obtained when

$$(3.5) \quad \gamma_n g_{m-n+1}^{n-1}(0) + g_{m-n+2}^{n-1}(0) = 0.$$

We may obviously assume that $g_{m-n+1}^{n-1}(0) \neq 0$. If not, order $m+2$ is obtained when n is replaced by $n-1$ in (2.12), and this is not possible from theorem 2.1.

Now g_r^{n-1} is a function of $\gamma_1, \dots, \gamma_{n-1}$. Let

$$(3.6) \quad G(\gamma_1, \dots, \gamma_{n-1}) = g_{m-n+2}^n(0) = \gamma_n g_{m-n+2}^{n-1}(0) + g_{m-n+3}^{n-1}(0)$$

and

$$(3.7) \quad g_r^{n-1}(t) = \delta g_r^{n-2}(t) + g_{r+1}^{n-2}(t), \quad \delta = \gamma_{n-1}.$$

Minimum of G is obtained when $\partial G / \partial \gamma_i = 0$, $i = 1(1)n-1$. In particular

$$(3.8) \quad \partial G / \partial \delta = \partial G / \partial \gamma_{n-1} = (\partial \gamma_n / \partial \delta) g_{m-n+2}^{n-1}(0) + \gamma_n (\partial / \partial \delta) g_{m-n+2}^{n-1}(0) + (\partial / \partial \delta) g_{m-n+3}^{n-1}(0) = 0.$$

But, from (3.5),

$$(3.9) \quad (\partial \gamma_n / \partial \delta) g_{m-n+1}^{n-1}(0) + \gamma_n (\partial / \partial \delta) g_{m-n+1}^{n-1}(0) + (\partial / \partial \delta) g_{m-n+2}^{n-1}(0) = 0.$$

Combining (3.8) and (3.9), using (3.7),

$$(3.10) \quad (\gamma_n - \delta) \{g_{m-n+1}^{n-2}(0) g_{m-n+3}^{n-2}(0) - [g_{m-n+2}^{n-2}(0)]^2\} = 0.$$

Assume that $\gamma_n \neq \delta$. Then we are asking for order m for the approximation

$$Q_{m-2}(q) / \prod_{i=1}^{n-2} (1 + \gamma_i q), \quad Q_{m-2} \in \pi_{m-2},$$

which from theorem 2.1 is impossible. Hence $\gamma_{n-1} = \delta = \gamma_n$. In the same way we may show

$$\gamma_1 = \gamma_2 \dots = \gamma_{n-2} = \gamma_n. \quad \blacksquare$$

The error constant with $\gamma_1 = \dots = \gamma_n = \gamma$ is given by

$$(3.11) \quad E_n^m(\gamma) = G(\gamma, \dots, \gamma) = \begin{cases} (-1)^{m+n} \gamma^{m+2} L_n^{(n-m-2)}(1/\gamma) & \text{for } m \leq n-2 \\ (n! / (m+2)!) \gamma^n L_{m+2}^{(m-n+2)}(1/\gamma) & \text{for } m > n-2, \end{cases}$$

where γ is a solution to

$$(3.12) \quad \begin{cases} L_n^{(n-m-1)}(1/\gamma) = 0 & \text{for } m \leq n-1 \\ L_{m+1}^{(m-n+1)}(1/\gamma) = 0 & \text{for } m > n-1. \end{cases}$$

Values of γ and $E_n^m(\gamma)$ can be found in Nørsett [3] and Wolfbrandt [8].

4. Application.

The ν -stage Runge-Kutta methods for solving $y' = f(y)$, $y(a) = y_0$, $y : [a, b] \rightarrow R^s$, numerically are defined by

$$Y_i = y_0 + h \sum_{j=1}^{\nu} a_{ij} f(Y_j), \quad i = 1(1)\nu + 1,$$

$$y_1 = Y_{\nu+1}.$$

It is well known that when the methods are explicit, $a_{ij} = 0$ for $j \geq i$, the maximum attainable order is ν . For a fully implicit method the highest obtainable order is 2ν .

When $a_{ij} = 0$ for $j > i$, the method is called semi-implicit. So far, only the case $a_{ii} = \gamma$, $i = 1(1)\nu$, has been considered by Nørsett [4] for the question of reachable order, with the result that order $\nu + 2$ is not obtained by any method. But by using the results from the preceding section we have,

THEOREM 4.1. *The highest attainable order of a ν -stage semi-implicit Runge-Kutta method is $\nu + 1$.*

PROOF. Applying the ν -stage semi-implicit Runge-Kutta method on $y' = -\lambda y$, $\lambda \in C$, and using $q = \lambda h$, results in an N -approximation to $\exp(-q)$ of the form (2.12) with $n := \nu$ and $m := \nu$. The theorem then follows from theorem 2.1. ■

With $a_{ii} = \gamma$, $i = 1(1)\nu$, methods of order $\nu + 1$ have been constructed for $\nu = 1, 2, 3$ by Nørsett [4]. For $\nu = 4$ Nørsett [4] shows that no method of order 5 exists. Hence, to get 4-stage semi-implicit Runge-Kutta methods of order 5, at least two different a_{ii} 's must be used in the method.

REFERENCES

1. Ehle, B. L., *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems*. Thesis, Univ. of Waterloo, Canada, 1969.
2. Chipman, F. H., *Numerical solution of initial value problems using A-stable Runge-Kutta processes*. Thesis, Univ. of Waterloo, Canada, 1971.
3. Nørsett, S. P., *Multiple Padé approximations to the exponential function*. Math. and Comp. 4 (1974). Dept. of Math., Univ. of Trondheim, Norway.
4. Nørsett, S. P., *Semi-explicit Runge-Kutta methods*, Math. and Comp. 6 (1974). Dept. of Math., Univ. of Trondheim, Norway.
5. Nørsett, S. P., *One-step methods of Hermite type for numerical integration of stiff systems*. BIT 14 (1974) 63-77.
6. Siemieniuch, J. L., *Properties of certain rational approximations to $\exp(-z)$* . BIT 16 (1976) 172-191.

7. S. P. Nørsett, *C-polynomials for rational approximation to the exponential function.* Numer. Math. 25 (1975) 39–56.
8. A. Wolfbrandt, *A study of Rosenbrock processes with respect to order conditions and stiff stability.* Thesis, 1977, Chalmers Univ. of Techn., Göteborg, Sweden.

INSTITUTT FOR NUMERISK MATEMATIKK
UNIVERSITETET I TRONDHEIM - NTH
N-7034 TRONDHEIM, NORWAY

DEPT. OF COMPUTER SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
S-402 20 GÖTEBORG 5, SWEDEN