Rational Interpolation to exp(-x) with Application to Certain Stiff Systems

Arieh Iserles

SIAM Journal on Numerical Analysis, Vol. 18, No. 1. (Feb., 1981), pp. 1-12.

Stable URL:
http://links.jstor.org/sici?sici=0036-1429%28198102%2918%3A1%3C1%3ARITWAT%3E2.0.CO%3B2-W

SIAM Journal on Numerical Analysis is currently published by Society for Industrial and Applied Mathematics.

SIAM J. NUMER. ANAL.
Vol. 18, No. 1, February 1981

0036-1429/81/1801-0001 $01.00/0

RATIONAL INTERPOLATION TO $\exp(-x)$ WITH APPLICATION TO CERTAIN STIFF SYSTEMS*

ARIEH ISERLES†

Abstract. Rational functions which interpolate the exponential function along an equispaced mesh are investigated. The C-polynomial theory of Nørsett is extended to the interpolatory case, demonstrating the connection between exponential interpolations and the usual exponential approximations. Explicit expressions for interpolatory type analogues to Padé and N-approximations are derived. The paper is supplemented by a numerical example.

1. Introduction. Exponential approximations, mainly the Padé approximations [1], [2], [10], N-approximations [6] and exponentially fitted approximations [3], [4], have attracted great interest during recent years, because they are connected with methods for the numerical solution of ordinary differential systems.

The common feature of these approximations is that they mainly fit to the exponential and to its derivatives at the origin. Hence, they are, in fact, Hermitian interpolations rather than approximations. The high degree of interpolation at the origin is of great benefit if (in the case of linear or mildly nonlinear autonomous differential systems) the behavior of the solution is determined by a single principal eigenvalue of the Jacobian matrix.

From the computational point of view, this is not the case if parabolic or hyperbolic partial differential systems are solved by the method of lines: If $u = u(t; x_1, \cdots, x_N)$ is in an appropriate Banach space and satisfies a partial equation

$$u_t = Au + f \quad \text{or} \quad u_{tt} = Au + f,$$

together with appropriate initial and boundary conditions, where A is a differential operator which is dependent on the spatial variables $x_1, \cdots, x_N$, then the method of lines consists of replacing A by a finite difference approximation A_h. Thus the partial equation is replaced by the ordinary differential system $u_t = A_h u + f$ (or $u_{tt} = A_h u + f$). Since the operator A is usually either linear (for a heat conduction equation, wave equation, etc.) or mildly nonlinear, the numerical solution of the ordinary differential system to a very large extent depends on the approximation to $\exp(gA_h)$, where g is the time step. Even a simple case of a heat conduction equation in one spatial variable exemplifies the difficulty which is encountered when one uses the conventional exponential approximation for this purpose. Applying the approximation

$$\frac{\partial^2}{\partial x^2} u(t, x) \approx \frac{1}{h^2} (u(t, x+h) - 2u(t, x) + u(t, x-h)),$$

the equation $u_t = (\partial^2/\partial x^2)u$ is approximated by $\mathbf{V}' = B\mathbf{V}$, where $\mathbf{V}$ is the space discretization of u and the prime denotes differentiation with respect to t and

$$B = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & \ddots & \ddots & & 0 \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & 0 & & 1 & -2 \end{pmatrix}.$$

* Received by the editors April 16, 1979, and in final revised form March 3, 1980.

† King's College, University of Cambridge, Cambridge, CB2 1ST, England.

If there are n mesh points, then the eigenvalues of $\exp(gB)$ are

$$\lambda_k = \exp\left(-2\frac{g}{h^2}\left(1+\cos\frac{k\pi}{n+1}\right)\right), \qquad 1 \leqq k \leqq n.$$

If $\exp(gB)$ is approximated by a rational or polynomial function of gB, then it is often necessary to use an approximation method that would give good accuracy for each of the terms $\exp(g\lambda_k I)$, $1 \leqq k \leqq n$, separately. Hence, one can expect rather poor results from an approximation which centers on the origin.

In general, the spectrum of the Jacobian matrix of A_h is unknown, but sometimes a useful practical assumption is that the eigenvalues are more or less spread uniformly in an interval $[-T, 0]$. Out of these eigenvalues, a smaller subset, contained in $[-T_0, 0]$, $0 < T_0 \ll T$, may influence strongly the numerical solution of the system. There is a need for exponential approximations that take into account this particular behavior. In order to meet this need, this paper presents rational approximations which interpolate the exponential function along an equispaced mesh in $[T_0, 0]$, instead of high-degree interpolation at the origin.

It should be mentioned that an alternative approach exists, namely to use rational Chebyshev approximations to $\exp(-x)$ over $[0, \infty)$. These approximations were extensively studied by Cody, Meinardus and Varga [13] and were generalized by Ehle [3].

In § 2 the C-polynomial theory of Nørsett [7] is extended to the interpolatory case. It is proved that a one-to-one correspondence exists between functions of a certain form and interpolations to the exponential. In § 3 interpolatory type analogues to Padé approximations and N-approximations are obtained. In § 4 upper bounds on the mesh size are found which yield A_0-acceptability. Furthermore, it is shown that sometimes stricter bounds prevent oscillations of the error curve. In § 5 numerical results regarding the heat conduction equation are given. It is shown that, using the same amount of computation, the interpolation method of this paper provides better accuracy than the usual Padé approximation method.

It is important to mention that according to the maximal interpolation theorem [4], if $R = P/Q$ is a rational function, $\deg P = m$, $\deg Q = n$, then the equation $R(x) - \exp(-x) = 0$ has at most $n+m+1$ zeros, where each zero is counted according to its multiplicity. This result is relevant because it provides a bound on the number of interpolation points.

2. *C*-function theory. Let $\tilde{R} = P/Q$ be a rational function, $\deg P = m$, $\deg Q = n \geqq m$, such that $\tilde{R}(x) - \exp(-x) = O(x^{n+1})$. Then, according to Nørsett [7], and Nørsett and Wanner [12], a unique polynomial $\tilde{p}$ exists, $\deg \tilde{p} = n$, $(d^n/dx^n)\tilde{p}(x) \equiv 1$, such that

$$\tilde{R}(x) = \frac{\sum_{k=0}^{n} (-1)^k \tilde{p}^{(n-k)}(1) x^k}{\sum_{k=0}^{n} (-1)^k \tilde{p}^{(n-k)}(0) x^k}, \tag{2.1}$$

where $\tilde{p}^{(k)}(x)$ denotes $(d^k/dx^k)\tilde{p}(x)$. Conversely, if $\tilde{R}$ is of the form (2.1), where $\tilde{p}$ is any polynomial of degree n, $\tilde{p}^{(n)}(x) \equiv 1$, then $\tilde{R}(x) - \exp(-x) = O(x^{n+1})$. This polynomial is called the *C-polynomial* of the approximation $\tilde{R}$.

Let $c > 0$ be the mesh size and

$$p(x, c) = \sum_{k=0}^{n} \frac{1}{k!} p_k(c) x^k, \qquad p_n(c) \equiv 1,$$

where $p_0(c), \cdots, p_{n-1}(c)$ are arbitrary functions of c. We denote $D_x^k p(x, c) = (\partial^k/\partial x^k) p(x, c)$.

THEOREM 1. *If*

$$P(x, c, p) = \sum_{k=0}^{n} D_x^{n-k} p(1, c)(1-e^{-c})^k (-x/c)_k$$

and

$$Q(x, c, p) = \sum_{k=0}^{n} D_x^{n-k} p(0, c)(e^c - 1)^k (-x/c)_k,$$

where $(-y)_k$ *is the factorial function* $(-y)_0 = 1$, $(-y)_k = (-y)(-y+1)\cdots(-y+k-1)$ *for* $k \geqq 1$, *then*

$$P(qc, c, p) = \exp(-qc) Q(qc, c, p)$$

for every integer q *such that* $0 \leqq q \leqq n$. *If* $D_x^k p(-1/(e^c-1), c) \neq 0$ *for every* $0 \leqq k \leqq n$, *then* Q *does not vanish at the mesh points* $\{kc\}_{k=0}^{n}$, *and if*

$$R(x, c, p) = \frac{P(x, c, p)}{Q(x, c, p)}, \tag{2.2}$$

then $R(qc, c, p) = \exp(-qc)$, $0 \leqq q \leqq n$.

Proof. If $x = qc$, $0 \leqq q \leqq n$, then

$$(-x/c)_k = (-q)_k = \begin{cases} (-1)^k q!/(q-k)!, & 0 \leqq k \leqq q, \\ 0, & q+1 \leqq k, \end{cases}$$

and

$$P(qc, c, p)/q! = \sum_{k=0}^{q} (-1)^k D_x^{n-k} p(1, c)(1-e^{-c})^k/(q-k)!,$$

$$Q(qc, c, p)/q! = \sum_{k=0}^{q} (-1)^k D_x^{n-k} p(0, c)(e^c-1)^k/(q-k)!.$$

But $D_x^{n-k} p(0, c) = p_{n-k}(c)$, $D_x^{n-k} p(1, c) = \sum_{i=0}^{k} p_{n-i}(c)/(k-i)!$; hence by changing summation sequence and interchanging i and k,

$$\begin{aligned} P(qc, c, p)/q! &= \sum_{k=0}^{q} (-1)^k \Big(\sum_{i=0}^{k} p_{n-1}(c)/(k-i)! \Big)(1-e^{-c})^k/(q-k)! \\ &= \sum_{k=0}^{q} (-1)^k \Big(\sum_{i=0}^{q-k} (-1)^i \binom{q-k}{i} (1-e^{-c})^i \Big)(1-e^{-c})^k p_{n-k}(c)/(q-k)! \\ &= \sum_{k=0}^{q} (-1)^k e^{-(q-k)c}(1-e^{-c})^k p_{n-k}(c)/(q-k)! = e^{-qc} Q(qc, c, p)/q!. \end{aligned}$$

It remains to prove that if $D_x^k p(-1/(e^c-1), c) \neq 0$, $0 \leqq k \leqq n$, then Q does not vanish at the mesh points. But

$$\begin{aligned} Q(qc, c, p)/q! &= (1-e^c)^q \sum_{k=0}^{q} \frac{1}{(q-k)!} D_x^{n-k} p(0, c) \frac{1}{(1-e^c)^{q-k}} \\ &= (1-e^c)^q D_x^{n-q} p\Big(-\frac{1}{e^c-1}, c \Big) \end{aligned}$$

for every q, $0 \leqq q \leqq n$, and the proof follows. Q.E.D.

DEFINITION. The function $p(x, c)$ is called the *C-function* of the rational interpolation (2.2).

THEOREM 2. *If* $R = P/Q$, $\deg P = m$, $\deg Q = n \geqq m$ and $R(sc) = \exp(-sc)$, s *integer,* $0 \leqq s \leqq n$, $c \neq 0$, *then a unique C-function p exists and* $R(x) = R(x, c, p)$ (cf. (2.2)).

Proof. The set $\{(-x/c)_0, \cdots, (-x/c)_n\}$ forms a basis of $\pi_n[x]$, the vector space of polynomials of degree n at most. Therefore $p_0(c), \cdots, p_n(c), q_0(c), \cdots, q_n(c)$ exist such that

$$P(x, c) = \sum_{k=0}^{n} q_{n-k}(c)(1-e^{-c})^k(-x/c)_k,$$

$$Q(x, c) = \sum_{k=0}^{n} p_{n-k}(c)(1-e^{c})^k(-x/c)_k. \tag{2.3}$$

Set $p(x, c) = \sum_{k=0}^{n} (1/k!)p_k(c)x^k$. Then $R(sc, c) = \exp(-sc)$, s integer between zero and n, implies

$$\sum_{k=0}^{s} (-1)^k q_{n-k}(c)(1-e^{-c})^k/(s-k)! = e^{-sc} \sum_{k=0}^{s} (-1)^k p_{n-k}(c)(e^c-1)^k/(s-k)!. \tag{2.4}$$

Using (2.4), induction on k easily gives

$$q_{n-k}(c) = \sum_{i=0}^{k} \frac{1}{(k-i)!} p_{k-i}(c) = D_x^{n-k} p(1-c),$$

and the proof follows. Q.E.D.

DEFINITION. If $p(x, c)$ is defined for every $0 < c \leqq c_0$ and $\lim_{c \to 0+} p(x, c)$ exists, the C-function is said to be *smooth.*

COROLLARY 1. *If p is a smooth C-function, let* $\tilde{p}(x) = \lim_{c \to 0+} p(x, c)$, *and let* $\tilde{R}(x, q)$ *be the exponential approximation* (2.1) *which is generated by the C-polynomial* $\tilde{p}$. *Then, if for all sufficiently small c, and for a positive integer s,* $R(kc, c, p) = \exp(-kc)$, $0 \leqq k \leqq n+s$, *then* $\tilde{R}(x, q) = \exp(-x) + O(x^{n+s+1})$.

Proof. By letting $c > 0$ tend to zero in (2.2), the form (2.1) (with $\tilde{p}$) is obtained. Let us suppose that $R(qc, c, p) = \exp(-qc)$ for every integer q between $n+1$ and $n+s$. Straightforward algebra, using the proof of Theorem 2, shows that

$$\begin{aligned} &P(qc, c, p) - e^{-qc} Q(qc, c, p) \\ &\quad = -q! \sum_{i=0}^{n} \frac{(-i)^i}{(q-i)!} \left[\sum_{k=n+1-i}^{q-i} (-i)^k \binom{q-i}{k} (1-e^{-c})^k \right] (1-e^{-c})^i p_{n-i}(c) \\ &\quad = (-1)^{q+1} q! \sum_{k=0}^{q-n-1} \frac{(-1)^k}{k!} (1-e^{-c})^{q-k} \left[\sum_{i=0}^{n} \frac{1}{(q-n-k+i)!} p_i(c) \right]. \end{aligned}$$

Hence the interpolation order is $n+s$ if and only if $P_k^*(1, c) = 0$ for every integer k between one and s, where

$$P_0^*(x, c) = p(x, c), \qquad P_{k+1}^*(x, c) = \int_0^x P_k^*(t, c)\, dt.$$

Hence, in the limit $P_k^*(1, 0) = 0$, $1 \leqq k \leqq s$, and the equality of interpolation and approximation orders follows by [7]. Q.E.D.

COROLLARY 2. *Let p and* $\tilde{p}$ *be as in Corollary* 1 *and R and* $\tilde{R}$ *be as in* (2.2) *and* (2.1) *respectively. Let* $\tilde{R}$ *be strongly* A_0*-acceptable and p be smooth. Then R is* A_0*-acceptable for* $c > 0$ *small enough.*

Proof. A function r is said to be A_0-acceptable if $|r(x)|<1$ for every $x>0$. It is strongly A_0-acceptable if, additionally, $\lim_{x\to\infty}|r(x)|<1$. The C-function p is smooth. Hence, by (2.1, 2), Corollary 1 and the C-polynomial theory for $c>0$ small enough S exists such that $R(x, c, p)=\tilde{R}(x, \tilde{p})+S(x)c$, and $|S|$ is uniformly bounded for $0\leqq x$. Let

$$m_r=\max_{r\leqq x}|S(x)|<\infty \quad \text{and} \quad M_r=\min_{r\leqq x}(1-|\tilde{R}(x, \tilde{p})|)$$

for $r>0$. Strong A_0-acceptability of $\tilde{R}$ implies that $M_r>0$. Then, if $0<c\leqq M_r/m_r$,

$$|R(x, c, p)|<1 \quad \text{for every } x\geqq r>0.$$

We cannot let r tend to zero, because $\lim_{r\to 0} M_r=0$. However, we can show that if $x, c>0$ are in the neighborhood of the origin then R is A_0-acceptable: the formula (2.2) implies that if $R=P/Q$ then

$$P(x, c)=1-xD_x^{n-1}\tilde{p}(1)+O(xc)+O(x^2),$$

$$Q(x, c)=1-xD_x^{n-1}\tilde{p}(0)+O(xc)+O(x^2).$$

The acceptability for $x, c>0$ is equivalent to $P(x, c)<Q(x, c)$, hence to $D_x^{n-1}\tilde{p}(1)>D_x^{n-1}\tilde{p}(0)$. But $D_x^{n-1}\tilde{p}(1)=D_x^{n-1}\tilde{p}(0)+D_x^n\tilde{p}(0)=D_x^{n-1}\tilde{p}(0)+1$, and the corollary is established. Q.E.D.

3. Padé interpolations and *N*-interpolations. The N-approximations [6] are of the form

$$R_n(x)=P_n(x)/(1+ax)^n,$$

where, in general, $\deg P_n\leqq n$. One of the most interesting cases is when $\deg P_n=n$ and a is such that

$$R_n(x)-e^{-x}=O(x^{n+2}).$$

DEFINITION. The *N-interpolation* to the exponential function is the rational function $R_n(x, c)=P_n(x, c)/(1+ax)^n$, such that $\deg P_n=n$ and $R_n(qc, c)=\exp(-qc)$, q an integer between zero and n.

THEOREM 3. *If*

$$R_n(x, c, a)=\sum_{k=0}^{n}\frac{1}{k!}\Big(\sum_{i=0}^{k}(-1)^i\binom{k}{i}(1+cai)^n e^{-ci}\Big)(-x/c)_k/(1+ax)^n, \tag{3.1}$$

then $R_n(qc, c, a)=\exp(-qc)$, $0\leqq q\leqq n$ *and the C-function is*

$$P(x, c, a)=\frac{(1-e^{-c})^{-n}}{n!}\sum_{k=0}^{n}\binom{n}{k}\Big(\sum_{i=0}^{n-k}(-1)^i\binom{n-k}{i}(1+cai)^n e^{-ic}\Big)((x-1)(1-e^{-c}))^k.$$

In addition, if a_0 *is a zero of the polynomial* $\sum_{k=0}^{n+1}(-1)^k\binom{n+1}{k}(1+kca)^n e^{-ck}$, *then* $R_n((n+1)c, c, a_0)=\exp(-(n+1)c)$.

Proof. Let $P_n(x)=\sum_{k=0}^{n}p_{n,k}(-x/c)_k$ and $R_n(x, c, a)=p_n(x)/(1+ax)^n$. Then

$$R_n(qc, c, a)=\exp(-qc), \qquad 0\leqq q\leqq n$$

is equivalent to

$$P_n(qc)=(1+aqc)^n e^{-qc}, \qquad 0\leqq q\leqq n. \tag{3.2}$$

Hence P_n is the Lagrange interpolation polynomial to $(1+axc)^n e^{-xc}$ at $x_q=q$,

$0 \leqq q \leqq n$. But

$$P_n(qc) = \sum_{k=0}^{q} (-1)^k \frac{q!}{(q-k)!} P_{n,k}.$$

Hence

$$\sum_{q=0}^{m} (-1)^q \binom{m}{q} p_n(qc) = \sum_{q=0}^{m} (-1)^q \binom{m}{q} \sum_{k=0}^{q} (-1)^k \frac{q!}{(q-k)!} p_{n,k}$$

$$= \sum_{k=0}^{m} \frac{m!}{(m-k)!} \left(\sum_{q=0}^{m-k} (-1)^q \binom{m-k}{q} \right) p_{n,k} = m! P_{n,m}.$$

But

$$\sum_{q=0}^{m} (-1)^q \binom{m}{q} p_n(qc) = \sum_{q=0}^{m} (-1)^q \binom{m}{q} (1+aqc)^n e^{-qc}.$$

Therefore

$$p_{n,m} = \frac{1}{m!} \sum_{q=0}^{m} (-1)^q \binom{m}{q} (1+aqc)^n e^{-qc},$$

and (3.1) is established. The expression for the C-function is readily established, using Theorem 1.

If, in addition, $R_n((n+1)c, c, a) = \exp(-(n+1)c)$ then, together with (3.2),

$$P_n((n+1)c) = (1+a(n+1)c)^n e^{-(n+1)c} \tag{3.3}$$

must be satisfied. Equations (3.2) and (3.3) form a set of $n+2$ equations in the $n+1$ variables $P_{n,0}, \cdots, p_{n,n}$. The consistency implies that these equations are dependent; i.e., multipliers $y_0, \cdots, y_{n+1}$ exist such that

$$\sum_{q=0}^{n+1} y_q p_n(qc) = 0 \quad \text{and} \quad \sum_{q=0}^{n+1} y_q^2 > 0.$$

By proceeding as in the first part of the theorem it is easy to verify that $y_q = (-1)^q \binom{n+1}{q}$. Therefore, if a_0 is such that

$$\sum_{q=0}^{n+1} (-1)^q \binom{n+1}{q} (1+qca_0)^n e^{-cq} = 0,$$

then (3.3) is satisfied and the rational function (3.1) interpolates $\exp(-x)$ at $x_q = qc$, $0 \leqq q \leqq n+1$. Q.E.D.

The analogues of Padé approximations are no less interesting than the N-approximations.

DEFINITION. If $R_{n,m}(x, c) = P_{n,m}(x, c)/Q_{n,m}(x, c)$ satisfies $R_{n,m}(qc, c) = \exp(-qc)$, $0 \leqq q \leqq n+m$, it is said to be a *Padé interpolation.*

Another technique will be used in the derivation of explicit expressions of the Padé interpolations. This technique, which was described by the present author in [5], relies on certain properties of the hypergeometric functions.

LEMMA. *If $a+b$ is neither zero nor a negative integer, then for every x and $0 < c < \ln 2$,*

$$ {}_2F_1(a, -x/c; a+b; 1-e^{-c}) = e^{-x} {}_2F_1(b, -x/c; a+b; 1-e^{c}). \tag{3.4}$$

Proof. If $0<c<\ln 2$, then $-1<1-e^{c}$, $1-e^{-c}<1$, and the hypergeometric functions in (3.4) exist and converge. Furthermore,

$$e^{-x}\,{}_2F_1(b,-x/c;a+b;1-e^{c})=e^{-x}\sum_{n=0}^{\infty}\frac{(b)_n(-x/c)_n}{n!(a+b)_n}(1-e^{c})^n$$

$$=\sum_{n=0}^{\infty}(-1)^n\frac{(b)_n(-x/c)_n}{n!(a+b)_n}e^{-(x-nc)}(1-e^{-c})^n.$$

But, according to [8, p. 74],

$$e^{-(x-nc)}=(1-(1-e^{-c}))^{x/c-n}={}_1F_0(-x/c+n;1-e^{-c})$$

$$=\sum_{m=0}^{\infty}\frac{(n-x/c)_m}{m!}(1-e^{-c})^m.$$

Therefore

$$e^{-x}\,{}_2F_1(b,-x/c;a+b;1-e^{c})$$

$$=\sum_{n=0}^{\infty}\frac{(b)_n(-x/c)_n}{n!(a+b)_n}\sum_{m=0}^{\infty}\frac{(n-x/c)_m}{m!}(-1)^n(1-e^{-c})^{n+m}$$

$$=\sum_{m=0}^{\infty}\frac{1}{m!}\left(\sum_{n=0}^{m}(-1)^n\binom{n}{m}\frac{(b)_n(-x/c)_n(n-x/c)_{m-n}}{(a+b)_n}\right)(1-e^{-c})^m.$$

Let us define

$$Z_m(t)=\sum_{n=0}^{m}(-1)^n\binom{n}{m}\frac{(b)_n(-t)_n(n-t)_{m-n}}{(a+b)_n}.$$

$Z_m(t)$ is a polynomial, $\deg Z_m=m$. But for every integer p, $0\leqq p\leqq m-1$,

$$0\leqq n\leqq p \quad \text{implies} \quad (n-p)_{m-n}=0,$$

$$p+1\leqq n\leqq m \quad \text{implies} \quad (-p)_n=0.$$

Hence $Z_m(p)=0$, $0\leqq p\leqq m-1$. Therefore $Z_m(t)=d(-t)_m$. The constant d can be computed by collecting the coefficients of $(-t)^m$:

$$d=\sum_{n=0}^{m}(-1)^n\binom{n}{m}\frac{(b)_n}{(a+b)_n}={}_2F_1(-m,b;a+b;1),$$

and, according to the Vandermonde theorem [8, p. 69]

$$d=\frac{(a)_m}{(a+b)_m}.$$

Hence

$$e^{-x}\,{}_2F_1(b,-x/c;a+b;1-e^{c})=\sum_{m=0}^{\infty}\frac{1}{m!}\frac{(a)_m(-x/c)_m}{(a+b)_m}(1-e^{-c})^m$$

$$={}_2F_1(a,-x/c;a+b;1-e^{-c}). \qquad \text{Q.E.D.}$$

THEOREM 4. *The explicit form of Padé interpolations is*

$$R_{n,m}(x,c)=P_{n,m}(x,c)/Q_{n,m}(x,c),$$

$$P_{n,m}(x,c)=\sum_{k=0}^{m}\frac{(n+m-k)!m!}{(n+m)!k!(m-k)!}(1-e^{-c})^k(-x/c)_k, \tag{3.5}$$

$$Q_{n,m}(x,c)=\sum_{k=0}^{n}\frac{(n+m-k)!n!}{(n+m)!k!(n-k)!}(1-e^{c})^k(-x/c)_k,$$

and the expansion of $G_{n,m}(x,c)=P_{n,m}(x,c)-e^{-x}Q_{n,m}(x,c)$ *is*

$$G_{n,m}(x,c)=(-1)^n\sum_{k=n+m+1}^{\infty}\frac{m!(k-m-1)!}{k!(n+m)!(k-n-m-1)!}(1-e^{-c})^k(-x/c)_k.$$

Proof. It is sufficient to establish the theorem for $0<c<\ln 2$. The result for $\ln 2\leqq c$ follows by analytic continuation.

By proceeding as in [5], let m and n be nonnegative integers, let $0<\varepsilon<1$ and $a=m+\varepsilon$, $b=n$ in the formula (3.4). Hence

$$\sum_{k=0}^{\infty}\frac{(-m-\varepsilon)_k(-x/c)_k}{k!(-n-m-\varepsilon)_k}(1-e^{-c})^k=e^{-x}\sum_{k=0}^{n}\frac{(-n)_k(-x/c)_k}{k!(-n-m-\varepsilon)_k}(1-e^{c})^k.$$

We split the sum on the left into the three parts

$$\sum_{k=0}^{m}\frac{(-m-\varepsilon)_k(-x/c)_k}{k!(-n-m-\varepsilon)_k}(1-e^{-c})^k+\sum_{k=m+1}^{n+m}\frac{(-m-\varepsilon)_k(-x/c)_k}{k!(-n-m-\varepsilon)_k}(1-e^{-c})^k$$
$$+\sum_{k=n+m+1}^{\infty}\frac{(-m-\varepsilon)_k(-x/c)_k}{k!(-n-m-\varepsilon)_k}(1-e^{-c})^k=I_{1,\varepsilon}+I_{2,\varepsilon}+I_{3,\varepsilon}.$$

We now consider the effect of letting ε tend to zero:

$$\lim_{\varepsilon\to 0+}(-m-\varepsilon)_k=\begin{cases}(-1)^k\dfrac{m!}{(m-k)!}, & 0\leqq k\leqq m,\\[2ex] 0, & m+1\leqq k,\end{cases}$$

and

$$\lim_{\varepsilon\to 0+}(-n-m-\varepsilon)_k=(-1)^k\frac{(n+m)!}{(n+m-k)!},\qquad 0\leqq k\leqq n+m.$$

Hence

$$\lim_{\varepsilon\to 0+}I_{1,\varepsilon}=\sum_{k=0}^{m}\frac{(n+m-k)!m!(-x/c)_k}{(n+m)!k!(m-k)!}(1-e^{-c})^k=P_{n,m}(x,c),$$

$$\lim_{\varepsilon\to 0+}I_{2,\varepsilon}=0,$$

and

$$\lim_{\varepsilon\to 0+}\sum_{k=0}^{n}\frac{(-n)_k(-x/c)_k}{k!(-n-m-\varepsilon)_k}(1-e^{c})^k=\sum_{k=0}^{n}\frac{(n+m-k)!n!(-x/c)_k}{(n+m)!k!(n-k)!}(1-e^{c})^k.$$

According to [5], for $k\geqq n+m+1$

$$\lim_{\varepsilon\to 0+}\frac{(-m-\varepsilon)_k}{(-n-m-\varepsilon)_k}=(-1)^n\frac{m!(k-m-1)!}{(n+m)!k!(k-n-m-1)!},$$

implying

$$\lim_{\varepsilon\to 0+}I_{3,\varepsilon}=(-1)^n\sum_{k=n+m+1}^{\infty}\frac{m!(k-m-1)!(-x/c)_k}{(n+m)!k!(k-n-m-1)!}(1-e^{-c})^k.$$

Therefore,

$$G_{n,m}(x,c)=P_{n,m}(x,c)-e^{-x}Q_{n,m}(x,c)$$
$$=(-1)^n\sum_{k=n+m+1}^{\infty}\frac{m!(k-m-1)!(-x/c)_k}{(n+m)!k!(k-n-m-1)!}(1-e^{-c})^k.$$

For every integer q, $0 \leqq q \leqq n+m$, $G_{n,m}(qc, c) = 0$. Hence $R_{n,m}$ as given in (3.5) is indeed the Padé interpolation. Q.E.D.

Remark. It is possible to deduce Theorem 4 from the C-function theory and the order conditions which appear in the proof of Corollary 1. Still, the lemma provides a result which is of some interest in the theory of special functions, being a generalization of the first Kummer formula [8].

COROLLARY. *If $c > 0$ is small enough and $n \geqq m+1$ then the m/n Padé interpolation $R_{n,m}$ is A_0-acceptable.*

Proof. If $n \geqq m+1$, then the strong A_0-acceptability of the respective Padé approximations [9], together with Corollary 1 to Theorem 2, implies the A_0-acceptability of $R_{n,m}(x, c)$ for $c > 0$ small enough. Q.E.D.

Remark. In a forthcoming paper [11] the author proves that no Padé interpolation can be A-acceptable, with the exception of $R_{1,0}$ and $R_{1,1}$.

4. A_0-acceptability and computational considerations. As most of the parabolic systems give rise to either symmetric or almost symmetric Jacobian matrices when solved by the method of lines, the A_0-acceptability of the exponential approximation is important, in order to provide stability.

The conventional Padé approximations $R_{n,m}$ are A_0-acceptable if and only if $n \geqq m$ [9]. If $n \geqq m+1$ this feature is preserved in respect to Padé interpolations if the mesh size $c > 0$ is small enough by the corollary to Theorem 4. Table 1 gives the upper bounds for c for $1 \leqq n \leqq 4$, $0 \leqq m \leqq n$.

TABLE 1.
Upper bounds for c for A_0-acceptability for Padé interpolations.

n \ m	0	1	2	3	4
1	A_0-stable	A_0-stable			
2	1.09856	1.51760	1.76272		
3	0.67392	1.10576	1.51728	1.65264	
4	0.57232	0.88496	1.12416	1.31408	1.46832

Consideration of the error curve for small values of n and m shows that the A_0-unacceptability for growing c is due to oscillation of the interpolation error in the interval $[0, (n+m)c]$. So, even if the interpolation is A_0-acceptable, such oscillations can prohibit its application. Hence, sometimes stricter bounds on the mesh size must be imposed.

For every $d > 0$ an upper bound c_d can be computed, such that for every $0 < c < c_d$ the oscillations of the error curve do not exceed d in the interval $[0, (n+m)c]$:

$$c_d = \max\{c > 0: |R_{n,m}(x, c') - e^{-x}| \leqq d,\ 0 \leqq x \leqq (n+m)c',\ 0 < c' \leqq c\}.$$

These upper bounds are listed in Table 2 for $d = .05, .10, .15, .25$ and $1 \leqq n \leqq 4$, $0 \leqq m \leqq n$.

It is evident from Tables 1 and 2 that, while the avoidance of oscillations of the error curve does not place restrictions on the use of Padé interpolations with $n = 3$ or 4, it apparently does so for $n = 1$ and 2.

TABLE 2.
c_d *for* $d = .05, .10, .15, .25$ *for Padé interpolations.*

n \ m	0	1	2	3	4
1	0.74830	1.40854			
	1.14011	1.89371			
	1.48086	2.28119			
	2.10620	2.94750			
2	0.85383	1.32328	1.65064		
	1.03008	1.52665	1.86429		
	1.13799	1.64729	1.98942		
	1.27263	1.79419	2.14022		
3	0.90167	1.28218	1.55869	1.77657	
	1.05259	1.45362	1.73955	1.96289	
	1.15685	1.57018	1.86174	2.08835	
	1.31703	1.74713	2.04639	2.27754	
4	0.87567	1.19119	1.42839	1.61991	1.78084
	0.96701	1.29338	1.53585	1.73048	1.89345
	1.02010	1.35198	1.59706	1.79318	1.95711
	1.08432	1.42208	1.66983	1.86744	2.03228

5. Numerical results. In this section the performances of the Padé approximations and Padé interpolations are compared when applied to the heat conduction equation with a single spatial variable.

To remove any influence of initial and boundary conditions, the approximation error $\|R(B)-\exp(-B)\|$ was computed for the matrix

$$B=\frac{g}{h^2}\begin{pmatrix} -2 & 1 & & & 0 \\ 1 & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & 1 \\ 0 & & & 1 & -2 \end{pmatrix}.$$

Two Hermitian L_2-norms were used, corresponding to two typical situations. The norm $\|\cdot\|_1$, induced by the Euclidean scalar product $(f, g)_1 = f^T g$, corresponds to the error along the boundary layer, where the contribution of the very small negative eigenvalues (i.e., large in absolute value) cannot be neglected. The norm $\|\cdot\|_2$, induced by the scalar product $(f, g)_2 = f^T \exp(B)g$, corresponds to the error along a "smooth" segment, emphasizing the error due to the larger eigenvalues. Observe that $\exp(B)$ is symmetric and positive definite, and so $\|\cdot\|_2$ is indeed a norm.

The results were computed for a 100×100 matrix (corresponding to spatial discretization with 100 mesh points) and $g/h^2 = 100$.

Table 3 gives errors of the approximation $R_{n,m}(B, c)$ to e^{-B} in the first norm versus the error for the Padé approximations. For the sake of comparison, the least-squares approximation error (in respect to this norm) for the n-by-n rational functions was computed; it appears in the last column. The last approximation is A_0-acceptable for $n = 2, 4$ only $(1 \leqq n \leqq 4)$.

Table 4 gives the same data for the second norm. Also, in this case, the A_0-acceptability of the best n-by-n rational approximations to the exponential is attained for $n = 2, 4$ only.

TABLE 3.

Columns 0–4 give the error of Padé aoproximations (first line) and Padé interpolations (second line) in the first norm. The third lines gives the mesh size c which yields the smallest error. The last column gives the error of the least-squares n-by-n rational approximations.

n \ m	0	1	2	3	4	Least-squares rational approximation
1	4.7_{-1}	9.2_{0}				2.1_{-1}
	2.5_{-1}	5.5_{-1}				
	(1.276)	(3.303)				
2	1.3_{-1}	3.2_{-1}	8.6_{0}			2.5_{-2}
	6.1_{-2}	5.7_{-2}	2.3_{-1}			
	(0.714)	(1.156)	(1.908)			
3	4.8_{-2}	5.7_{-2}	2.6_{-1}	7.9_{0}		2.9_{-3}
	1.8_{-2}	9.9_{-3}	1.5_{-2}	6.4_{-2}		
	(0.601)	(0.841)	(1.159)	(1.678)		
4	1.9_{-2}	1.5_{-2}	3.3_{-2}	2.3_{-1}	7.3_{0}	3.1_{-4}
	5.9_{-3}	2.4_{-3}	2.2_{-3}	4.3_{-3}	2.0_{-2}	
	(0.541)	(0.705)	(0.895)	(1.151)	(1.526)	

Inspection of Tables 3 and 4 shows that Padé interpolations are better than the corresponding Padé approximations, when applied to the heat conduction equation in the above-mentioned conditions, both in the boundary layer and along "smooth" segments of the solution. It seems that there is little to be gained by using the least-squares n-by-n rational approximations, even in the cases when they are A_0-acceptable.

There is one further advantage in Padé interpolations: The results for the diagonal approximations are better in the absence of transient components of the numerical solution, when the order matters most, while in the boundary layer the subdiagonal L_0-acceptable approximations give better results. This behavior is consistent both for

TABLE 4.

The same information as Table 3, for the second norm.

n \ m	0	1	2	3	4	Least-squares rational approximation
1	1.4_{-1}	9.9_{-2}				1.8_{-2}
	8.0_{-2}	2.6_{-2}				
	(0.711)	(0.903)				
2	3.7_{-2}	1.9_{-2}	1.4_{-2}			6.2_{-4}
	1.7_{-2}	4.1_{-3}	1.5_{-3}			
	(0.511)	(0.647)	(0.742)			
3	1.2_{-2}	4.4_{-3}	2.6_{-3}	2.1_{-3}		1.8_{-5}
	4.1_{-3}	7.8_{-4}	2.4_{-4}	9.0_{-5}		
	(0.449)	(0.543)	(0.618)	(0.680)		
4	4.2_{-3}	1.2_{-3}	5.7_{-4}	3.8_{-4}	3.1_{-4}	4.1_{-7}
	1.1_{-3}	1.7_{-4}	4.3_{-5}	1.4_{-5}	5.6_{-6}	
	(0.412)	(0.485)	(0.545)	(0.598)	(0.645)	

Padé approximations and for Padé interpolations. However, this distinction is less acute for the Padé interpolations: in the boundary layer the [1/4] Padé approximation gives an error which is smaller than the error for the [4/4] approximation by a factor of 487. The corresponding factor for the respective interpolations is only 8 (cf. Table 3). Hence, if Padé approximations are used, the best strategy is to employ different approximations for different segments of the solution, while the diagonal Padé interpolations are effective both inside and outside the boundary layer.

Acknowledgment. The author wishes to thank Professor M. J. D. Powell for his very kind remarks and the significant improvements of this paper. The proof of Theorem 3 is due to Professor Powell; it replaces a different proof which was rather tedious and less elegant.

REFERENCES

[1] G. BIRKHOFF AND R. S. VARGA, *Discretization errors for well-set Cauchy problems*, I. J. Math. Phys., 44 (1965), pp. 1–23.

[2] B. L. EHLE, *A-stable methods and Padé approximations to the exponential*, SIAM J. Math. Anal., 4 (1973), pp. 671–680.

[3] ———, *On certain order constrained Chebyshev rational approximations*, J. Approx. Th., 17 (1976), pp. 297–306.

[4] A. ISERLES, *On the generalized Padé approximations to the exponential function*, this Journal, 16 (1979), pp. 631–636.

[5] ———, *A note on Padé approximations and generalized hypergeometric functions*, BIT, 19 (1979), pp. 543–545.

[6] S. P. NØRSETT, *Restricted Padé approximations to the exponential function*, this Journal, 15 (1978), pp. 1008–1029.

[7] ———, *C-polynomials for rational approximation to the exponential function*, Numer. Math., 25 (1975), pp. 39–56.

[8] E. D. RAINVILLE, *Special Functions*, Macmillan, New York, 1967.

[9] R. S. VARGA, *On higher order stable implicit methods for solving parabolic partial differential equations*, J. Math. Phys., 40 (1961), pp. 220–231.

[10] G. WANNER, E. HAIRER AND S. P. NØRSETT, *Order stars and stability theorems*, BIT, 18 (1978), pp. 475–489.

[11] A. ISERLES AND M. J. D. POWELL, *On the A-Acceptability of Rational Approximations to the Exponential*, DAMTP NA/3, Univ. of Cambridge, 1980.

[12] S. P. NØRSETT AND G. WANNER, *The real-pole sandwich for rational approximations and oscillation equations*, BIT, 19 (1979), pp. 79–94.

[13] W. J. CODY, G. MEINARDUS AND R. S. VARGA, *Chebyshev rational approximations to* e^{-x} *in* $[0, +\infty)$ *and applications to heat-conduction problems*, J. Approx. Th., 2 (1969), pp. 50–65.