Piecewise Padé-Approximation of e^{-x} on $[0, \infty)$

H. U. OPITZ AND K. SCHERER

Institut für Angewandte Mathematik, Universität Bonn, Bonn, West Germany

Communicated by Charles A. Micchelli

Received December 18, 1982

INTRODUCTION

Recently the problem of uniform approximation of e^{-x} on $[0, \infty)$ by rational functions has received much attention. One main motivation was given by the fundamental paper [3] of Cody *et al.*, where the application to the construction of numerical procedures for initial value problems was considered. Besides the special questions arising from this application, the asymptotic behaviour of the approximation error has been considered [1, 7, 8, 11, 13, 14]. To our knowledge the best rate of approximation so far has been obtained in [11] using an appropriate translation of a Padéapproximation (a still better rate was announced in [6] using Laguerre– Padé-approximation, however, the proof is incorrect and incomplete).

Here we introduce as a new aspect the question whether the rate of approximation can be improved by using piecewise Padé-approximation (the total number of parameters remaining unchanged). We show that this rate indeed improves significantly. The thorough analysis of the local error of Padé-approximation constitutes the main part of our work, thereby extending and sharpening the results of [8] and [11] for the error on $[0, \infty)$. This was made possible by the special form of the Padé-approximation of e^{-x} but from our result a similar improvement of the rate of approximation may also be expected in the case of best rational approximation.

1. POINTWISE ERROR ESTIMATES

It is well known (e.g., [10]) that the (m, n) Padé-approximation of e^x has the form

$$R_{m,n}(x) = \frac{\int_0^\infty t^n (t+x)^m e^{-t} dt}{\int_0^\infty t^m (t-x)^n e^{-t} dt}.$$
(1.1)

0021-9045/84 \$3.00

From this it follows directly that

$$e^{-x} - R_{m,n}(-x) = -\frac{\int_0^x (u-x)^m u^n e^{-u} du}{\int_0^\infty u^m (u+x)^n e^{-u} du}.$$
 (1.2)

We introduce now the following functions $(m \leq n)$

$$f(t) := e^{-x't} t(1-t)^r, \qquad g(t) := e^{-x't} (1+t) t^r$$
(1.3)

where

$$x' := x/n, \qquad r := m/n.$$
 (1.4)

Then a simple calculation yields by (1.2) for x > 0

$$|e^{-x} - R_{m,n}(-x)| = \int_0^1 f(t)^n dt \bigg/ \int_0^\infty g(t)^n dt.$$
 (1.5)

The investigation of this error is done by Laplace's method or "méthode du col" according to which the values of the integrals in (1.4) are determined essentially by the maxima of their integrands. This is carried out in detail to show that the method yields very precise estimates (exact up to a factor of order $n^3 \log n$) in a simple manner. To our knowledge such estimates do not exist in literature (however, see [4, 6.3.3]).

THEOREM 1. For all x > 0, $n \ge 2$ and $r = m/n \in (0, 1]$ there holds

$$\frac{Cnx}{x+n+m} E\left(\frac{x}{n}, r\right)^n \leq |e^{-x} - R_{rn,n}(-x)| \leq \frac{8enx}{x+n+m} E\left(\frac{x}{n}, r\right)^n$$

with the constant $C = [3e(n+1)(1+2n)^2 \log(1+2n)]^{-1}$. The quantity E(y,r) is defined by

$$E(y,r) := e^{h-y} \frac{p-h}{p+h} \left(\frac{h+q}{h-q}\right)^r$$
(1.6)

via the abbreviating notations

$$p := p(y, r) := y + 1 + r, \qquad q := q(y, r) := y - 1 - r$$

$$h := h(y, r) := \sqrt{p^2 - 4y} = \sqrt{q^2 + 4ry}. \qquad (1.7)$$

Proof. By direct calculation we find that (with y = x' in (1.7))

$$x't^* = p/2 - \sqrt{p^2/4 - x'}$$
(1.8)

$$x't^{2} = -q/2 + \sqrt{q^{2}/4 + rx'}$$
(1.9)

determine the points t^* and t° of the maxima of the functions f(t) and g(t), respectively. We show then for the numerator in (1.5) that

$$\frac{t^*f(t^*)^n}{e(n+1)} \leqslant \int_0^1 f(t)^n \, dt \leqslant 4t^*f(t^*)^n. \tag{1.10}$$

To this end we observe that for any $a \in (0, t^*)$

$$\int_0^1 f(t)^n \, dt \ge \int_{t^*-a}^{t^*} f(t)^n \, dt \ge a[(t^*-a)f(t^*)/t^*]^n.$$

The maximum with respect to *a* is obtained for $a = t^*/(n+1)$ which gives the left-hand side in (1.10). For the upper estimate we split $\int_0^1 = \int_0^{t^*} + \int_{t^*}^1$ and use

$$\int_{t^*}^{1} f(t)^n dt = f(t^*)^n \int_{0}^{1-t^*} \left[e^{-x't} \left(1 + \frac{t}{t^*} \right) \left(1 - \frac{t}{1-t^*} \right)^r \right]^n dt.$$

By (1.8) we have $t^* = (p - x't^*)^{-1}$. Hence

$$e^{-x't}\left(1-\frac{t}{1-t^*}\right)^r \leq \exp\left\{-t\left(x'+\frac{r}{1-t^*}\right)\right\} = e^{-t/t^*}$$

and after the substitution $u = n + nt/t^*$ we get

$$\int_{t^*}^{1} f(t)^n dt \leqslant \frac{t^*}{n} f(t^*)^n \int_{n}^{n/t^*} e^{n-u} (u/n)^n du$$
$$\leqslant \frac{t^*}{n} f(t^*)^n \left(\frac{e}{n}\right)^n n!$$
$$\leqslant \sqrt{\frac{2\pi}{n}} e^{1/12n} t^* f(t^*)^n.$$

Together with the trivial estimate

$$\int_0^{t^*} f(t)^n \, dt \leqslant t^* f(t^*)^n$$

the right-hand side of (1.10) now follows.

As to the denominator in (1.5) we show

$$\frac{1}{ex}g(t^{\hat{}})^n \leqslant \int_0^\infty g(t)^n \, dt \leqslant \frac{3(1+2n)^2 \log(1+2n)}{x} \, g(t^{\hat{}})^n. \tag{1.11}$$

Here the left-hand side follows similarly as above. For any a > 0 we have

$$\int_0^\infty g(t)^n dt \ge \int_{t^*}^{t^*+a} g(t)^n dt \ge a e^{-ax} g(t^*)^n$$

and maximization with respect to a yields a = 1/x which gives the desired estimate.

For the proof of the right-hand side of (1.11) we write with b > 0

$$\int_{0}^{\infty} g(t)^{n} dt \leq \int_{0}^{(1+b)(t^{2}+1/x)} + \int_{(1+b)(t^{2}+1/x)}^{\infty} \equiv I_{1} + I_{2}.$$
 (1.12)

We have

$$I_2 = (1+b) \int_{t^*+1/x}^{\infty} \left[e^{-(1+b)x't} (1+b)^r t^r (1+(1+b)t) \right]^n dt$$
$$\leq (1+b)^{1+rn+n} e^{-bx(t^*+1/x)} \int_{t^*+1/x}^{\infty} g(t)^n dt.$$

The choice $b = 2(1 + 2n) \log(1 + 2n)$ yields

$$(1+b)^{1+rn+n} e^{-b} \leq (1+b)^{1+2n} e^{-b} \leq \frac{1}{5}$$

for $n \ge 2$ so that

$$I_2 \leq (I_1 + I_2)/5$$
 or $I_2 \leq I_1/4$. (1.13)

Now by (1.9) there holds $x't^{\uparrow} \leq 1 + r \leq 2$ and consequently

$$I_1 \leq (1+b)(t^{\hat{}}+1/x) g(t^{\hat{}})^n \leq (1+b)(2n+1) g(t^{\hat{}})^n/x,$$

establishing the right-hand side of (1.11) by (1.12) and (1.13).

The theorem now follows from (1.10) and (1.11) taking into account $E(x', r) = f(t^*)/g(t^*)$ and the inequality

$$\frac{x}{x+n+m} \leqslant x't^* \leqslant \frac{2x}{x+n+m},$$

which is a consequence of (1.8).

We remark that the result of Theorem 1 is a more precise statement of the classical property $|e^{-x} - R_{rn,n}(-x)| = O(|x|^{m+n+1}), |x| \to 0$, of Padé-approximation since one can easily show that $E(y, r)^n = O(|y|^{m+n}), |y| \to 0$, for fixed r > 0.

For the following we need the counterpart of Theorem 1 in case x < 0. To this end we assume *n* to be even since otherwise there is a pole in (1.1). After replacing x by -x in (1.2), (1.5) a short calculation yields

$$|e^{x} - R_{m,n}(x)| = \frac{\int_{0}^{1} \left[e^{x't}t(1-t)^{r}\right]^{n} dt}{\int_{0}^{\infty} \left[e^{-x't}(1-t)t^{r}\right]^{n} dt} = \frac{e^{x}}{1+A^{-1}}$$
(1.14)

where

$$A := e^x \int_0^1 \varphi(s)^n \, ds \bigg/ \int_0^\infty \psi(s)^n \, ds$$

and

$$\varphi(s) := e^{-x's}(1-s)s^r, \qquad \psi(s) := e^{-x's}s(1+s)^r.$$

But $\varphi(s)^n$, $\psi(s)^n$ are equal to $f(s)^n$ and $g(s)^n$, respectively, except for the interchange of *m* and *n* (see (1.3)). Thus *A* can be estimated in Theorem 1 (cf. (1.10), (1.11)) leading to

$$Cs^*x[\varphi(s^*)/\psi(s^{\hat{}})]^n \leqslant e^{-x}A \leqslant 4es^*x[\varphi(s^*)/\psi(s^{\hat{}})]^n$$
(1.15)

with same constant C. The numbers s^* and $s^{\hat{}}$ are given by (note the interchange of m and n)

$$x's^* = p/2 - \sqrt{p^2/4} - rx'$$

$$x's^{-} = -q/2 + \sqrt{q^2/4 + x'}.$$

Substitution of these formulas into (1.15) yields

$$Cs^*xF(x/n,r)^n \leq A \leq 4es^*xF(x/n,r)^n$$

where

$$F(y,r) = e^{\tilde{h}} \frac{\tilde{h} + q}{\tilde{h} - q} \left(\frac{p - \tilde{h}}{p + \tilde{h}}\right)^r, \qquad (1.16)$$

and

$$\tilde{h} := \tilde{h}(y, r) := \sqrt{(y - 1 - r)^2 + 4y}.$$
 (1.17)

Now, using the inequality

$$\frac{rx}{x+n+m} \leqslant x's^* \leqslant \frac{2rx}{x+n+m}$$

and in (1.14)

$$\frac{1}{2}\min(1,A) \leqslant \frac{1}{1+A^{-1}} \leqslant \min(1,A),$$

we obtain finally as the counterpart of Theorem 1.

THEOREM 2. For all x > 0, n even and $r = m/n \in (0, 1]$ there holds

$$\frac{Cmxe^x}{2(x+n+m)}\min(1, F(x/n, r)^n)$$
$$\leqslant |e^x - R_{rn,n}(x)| \leqslant \frac{8emxe^x}{x+n+m}\min(1, F(x/n, r)^n)$$

where C is defined in Theorem 1 and F(x/n, r) is given by (1.16), (1.17).

2. LOCAL ERROR ESTIMATES

The pointwise error estimates of the preceding section are now used to derive asymptotically exact error estimates for intervals of the form [0, a], a > 0. The crucial point here is the optimal choice of the parameter r = m/n and of a number $\alpha \in [0, a]$ by which the Padé-approximation (1.1), (1.2) may be translated. These questions can be attacked successfully because the functions E(y, r) and F(y, r) introduced above depend on y, r, respectively, in a rather simple manner, as the following two lemmas of technical nature will show.

LEMMA 1. The following assertions are true:

(i) E(y, r) is strictly monotone increasing in y for fixed $r \in (0, 1]$ if $y < \varphi(r)$,

$$\varphi(r) := \frac{(1+r)^2}{2(1-r)} \qquad (\varphi(1) := \infty), \tag{2.1}$$

and strictly monotone decreasing for $y > \varphi(r)$.

(ii) E(y, r) is strictly monotone increasing in r for fixed $y \in (0, \infty)$ if r + 1 < y and strictly monotone decreasing in r if r + 1 > y.

Proof. We only show part (i) in detail. Its proof relies on the formulas (1.6), (1.7) and the relations between the quantities p, q and h appearing in them. We have

$$E(y,r)^{-1} \frac{\partial}{\partial y} E(y,r)$$

= $h' - 1 + \frac{1 - h'}{p - h} - \frac{(1 + h')}{p + h} + \frac{r(1 + h')}{h + q} - \frac{r(h' - 1)}{h - q}$
= $h' - 1 + \frac{2(h - ph')}{p^2 - h^2} + 2r \frac{(h - h'q)}{h^2 - q^2}$
= $h' - 1 + \frac{2h - h'(p + q)}{2y} = \frac{h}{y} - 1$

where we have used the relations $p^2 - h^2 = 4y$, $h^2 - q^2 = 4ry$, p + q = 2y. The solution of the equation h(y, r) = y is then $y = \varphi(r)$ given by (2.1).

Part (ii) follows from the relation

$$\frac{\partial}{\partial r}E(y,r) = E(y,r)\ln\left(\frac{h+q}{h-q}\right).$$
(2.2)

Concerning F(y, r) we have the simpler statement of

LEMMA 2. F(y, r) is strictly monotone increasing in $y \in [0, \infty)$ for fixed $r \in (0, 1]$ and strictly monotone decreasing in $r \in (0, 1]$ for fixed $y \in (0, \infty)$.

Lemma 2 is a direct consequence of the following formulas obtained from (1.16), (1.17):

$$\frac{\partial}{\partial y}F(y,r) = \frac{\tilde{h}}{y}F(y,r), \qquad \frac{\partial}{\partial r}F(y,r) = F(y,r)\ln\left(\frac{p-\tilde{h}}{p+\tilde{h}}\right).$$
(2.3)

As a first step we derive from Lemma 1 local error estimates for the Padéapproximation (1.2) on [0, a] with $r \in (0, 1]$ being fixed.

LEMMA 3. For each a > 0 and $r \in (0, 1]$ there holds

$$\{\|e^{-x} - R_{[rn],n}(-x)\|_{\infty,[0,a]}\}^{1/n} \approx \min(1,a)^{1/n} E(\min(a/n,\varphi(r)),r).$$

Here and in the following the symbol \approx stands for equality up to a factor which is bounded by positive absolute constants from above and below and which tends to 1 as $n \rightarrow \infty$.

The proof follows immediately from Theorem 1 and Lemma 1, together with the observation that $\varphi(r) \ge \frac{1}{2}$ and the inequalities

$$\frac{1}{4}\min\left(1,\frac{x}{n}\right) \leqslant \frac{x}{x+n+m} \leqslant \min\left(1,\frac{x}{n}\right)$$

which are a consequence of (1.18).

OPITZ AND SCHERER

Lemma 3 shows that $a = \varphi(r) \cdot n$ gives the critical length of the interval [0, a] beyond that the local error is always equal to the global error on $[0, \infty)$. In the latter case an easy calculation yields

$$L(r) := E(\varphi(r), r) = \frac{1}{2}(2r)^r (1-r)^{1-r}.$$
(2.4)

This function has already been found by Ni *et al.* [8] in their study of the global error on $[0, \infty)$. It is strictly convex in $r \in [0, 1]$ and attains its minimum at r = 1/3 where L(1/3) = 1/3. Hence we obtain as a corollary of Lemma 3 one of the main results in [8]:

COROLLARY 1. For $r \in (0, 1]$ define

$$A(r) := \lim_{n \to \infty} \|e^{-x} - R_{[rn],n}(-x)\|_{\infty, [0,\infty)}^{1/n}.$$

Then A(r) has a minimum for r = 1/3 with A(1/3) = 1/3.

With the help of Lemma 3 this result can now be extended to any interval [0, a]. To this end we define for a > 0, $n \in \mathbb{N}$ and $r \in (0, 1]$

$$A_n(a,r) := \{ \| e^{-x} - R_{[rn],n}(-x) \|_{\infty,[0,a]} \}^{1/n}.$$
 (2.5)

THEOREM 3. There holds

$$\inf_{1/n \le r \le 1} A_n(a, r) \approx \min(1, a^{1/n}) \min(\frac{1}{3}, E(a/n, 1)).$$

Remark. This shows that the optimal choice for r is r = 1 if $a \le a_0$, where a_0 is determined by the equation $E(a_0/n, 1) = 1/3$ ($a_0 = n \cdot 1.660605$). In case $a > a_0$ the local error and the global error are asymptotically equal according to Corollary 1. The optimal choice here is r = 1/3.

Proof of Theorem 3. If $a/n \leq 1/2$ we have $\min(a/n, \varphi(r)) = a/n$ so in view of the monotonicity property in Lemma 1(ii) the assertion follows directly from Lemma 3. In the other case we define r_a (uniquely) by $a/n = \varphi(r_a)$. Then Lemma 1(ii) shows that

$$\inf_{1/n \le r \le 1} E(\min(a/n, \varphi(r)), r)
= \min(\inf_{r_a < r \le 1} E(a/n, r); \lim_{1/n \le r \le r_a} L(r))
= \min(\min(E((\varphi(r_a), r_a); E(a/n, 1)); \inf_{1/n \le r \le r_a} L(r)))
= \min(E(a/n, 1); \inf_{1/n \le r \le r_a} L(r)).$$

But since $L(r) \ge L(1/3) = 1/3$ and since for $a/n \ge a_0/n > 4/3$ there holds $\varphi(r_a) > 4/3$ or $r_a > 1/3$ we conclude

$$\inf_{1/n \le r \le 1} E(\min(a/n, \varphi(r)), r) = \min(E(a/n, 1), 1/3)$$

establishing Theorem 3.

The error estimate of Theorem 3 can still be improved when using an appropriate translate of the Padé-approximation (1.2), i.e., we consider instead of (2.5) the (smaller) quantity

$$B_n(a,r) := \inf_{0 \leqslant \alpha \leqslant a} \|e^{-x} - e^{-\alpha} R_{[rn],n}(-x+\alpha)\|_{\infty,[0,a]}^{1/n}.$$
 (2.6)

In order to study its behaviour we need as a counterpart to Lemma 3

LEMMA 4. For each $\alpha > 0$ and $r \in (0, 1]$ there holds

$$\{\|e^{x} - R_{[rn],n}(x)\|_{\infty,[0,\alpha]}\}^{1/n} \approx [r\min(1,\alpha)]^{1/n} e^{\alpha/n}\min(1,F(\alpha/n,r)).$$

The Lemma follows from Theorem 2 and Lemma 2.

THEOREM 4. Let

$$R := \inf_{0 < r \leq 1} \inf_{y > 0} \max\{F(y, r), e^{-y}L(r)\}$$
(2.7)

and define the function G(x) for x > 0 implicitly by

$$G(x) := \{F(z, 1) : F(z, 1) = e^{-x}F(x - z, 1), 0 \le z \le x\}.$$
 (2.8)

Then

$$\inf_{1/n \le r \le 1} B_n(a, r) \approx \min(1, a)^{1/n} \min(R, G(a/n)).$$
(2.9)

Remark. The number *R* describing the asymptotic behaviour of the global error $(a = \infty)$ has the value $(4.0982107...)^{-1}$ and was already computed by Rahman and Schmeisser [11] by a somewhat different method. It is attained in (2.7) for the values (cf. [11])

$$r = r^* = 0.4832939..., \quad y = y^* = 0.3598078....$$
 (2.10)

In case $a < a_1$, where $a_1/n = 3.428985...$ is given by

$$R = G(a_1/n) \tag{2.11}$$

the local error is less than the global one and described by the function

OPITZ AND SCHERER

G(a/n) showing that the choice r = 1 is then optimal for Padéapproximation.

Proof of Theorem 4. For fixed r and $\alpha \in [0, a]$ we have by Lemmas 3 and 4

$$\begin{split} \|e^{-x} - e^{-\alpha} R_{[rn],n}(-x+\alpha)\|_{\infty,[0,a]}^{1/n} \\ &= e^{-\alpha/n} \|e^{-y} - R_{[rn],n}(-y)\|_{\infty,[-\alpha,a-\alpha]}^{1/n} \\ &= e^{-\alpha/n} \max\{\|e^{y} - R_{[rn],n}(y)\|_{\infty,[0,\alpha]}^{1/n}, \|e^{-y} - R_{[rn],n}(-y)\|_{\infty,[0,a-\alpha]}^{1/n}\} \\ &\approx \max\{[r\min(1,\alpha)]^{1/n} \cdot \min(1,F(\alpha/n,r)), \\ &\min(1,a-\alpha)^{1/n} e^{-\alpha/n} E(\min((a-\alpha)/n,\varphi(r)), r)\}. \end{split}$$

From this we get

$$\inf_{1/n \le r \le 1} B_n(a, r) \approx \min(1, a)^{1/n} \cdot I$$
 (2.12)

where

$$I := \inf_{1/n \leq r \leq 1} \inf_{0 \leq \alpha \leq a} \max\{F(\alpha/n, r), e^{-\alpha/n}E(\min((a-\alpha)/n, \varphi(r)), r)\}.$$
(2.13)

(The upper estimate here is immediate; the lower one requires some additional arguments showing, e.g., that the cases $F(\alpha/n, r) > 1$, $\alpha > a/2$ and $\alpha < \min(1, a/10)$ do not give the infimum.)

For fixed a > 0 we now introduce

$$I_{1} := \inf_{1/n \le r \le 1} \inf_{\min(\varphi(r), a') \le a' - \alpha' \le a'} \max\{F(\alpha', r), e^{-\alpha'}L(r)\}$$
(2.14)

$$I_2 := \inf_{1/n \leqslant r \leqslant 1} \inf_{0 \leqslant a' - a' \leqslant \min(\varphi(r), a')} \max\{F(a', r), e^{-a'}E(a' - a', r)\}$$
(2.15)

where we put a' := a/n, $\alpha' = \alpha/n$. Then we have $I = \min(I_1, I_2)$. Now in (2.15) apart from the case r = 1 we need only consider those α' with $a' - \alpha' \ge 1 + r$ since otherwise both terms forming the "max" are decreasing in r so that the infimum is attained for r = 1 again. Also we have then necessarily $r \ge 1/3$ since in case $1 + r > \varphi(r)$ the feasible set in (2.15) is empty. Thus in view of $\varphi(1) = \infty$ and

$$\inf_{0 \le a' - \alpha' \le a'} \max\{F(\alpha', 1), e^{-\alpha'}E(a' - \alpha', 1)\}$$

=
$$\inf_{0 \le a' - \alpha' \le a'} \max\{F(\alpha', 1), e^{-a'}F(a' - \alpha', 1)\} = G(a')$$

we find

$$I_2 = \min(G(a'), \inf_{\substack{1/3 \le r \le 1}} H(r))$$
(2.16)

where we have introduced (provided that the feasible domain is non-void)

$$H(r) = H(r, a')$$

:= $\inf_{r+1 \le a' - a' \le \min(\varphi(r), a')} \max\{F(a', r), e^{-a'}E(a' - a', r)\}.$ (2.17)

Below we shall prove the following crucial

LEMMA 5. If $a' \leq 3.6$, then on each subinterval [c, d] of $[\max(r_1, 1/3), 1]$ the infimum of H(r) has a value $\geq \min(H(c), G(a'))$ where $r_1 = r_1(a) \in [1/3, 1]$ is defined by

$$F(a' - \varphi(r_1), r_1) = e^{-a' + \varphi(r_1)} L(r_1).$$
(2.18)

If no such r_1 exists then we set $r_1 := 0$.

Now, if $1/3 \le r \le r_1$ in (2.16) we have $a' - \varphi(r) > a' - \varphi(r_1) > 0$ and therefore by (2.18) and Lemmas 1 and $2 F(\alpha', r) > e^{-\alpha'}E(a' - \alpha', r)$ for all feasible α' in (2.17) which does certainly not give the infimum. Hence we may assume $r \ge \max(1/3, r_1)$ in (2.16). But then with the help of Lemma 5 we see that for $\alpha' \le 3.6$

$$I_2 \ge \min(G(a'), H(\max(1/3, r_1))) \ge \min(G(a'), I_1)$$

Here we have used that by (2.18) $H(r_1) = \max\{F(a' - \varphi(r_1), r_1), e^{-a' + \varphi(r_1)}L(r_1)\} \ge I_1$ as well as $H(1/3) = \max\{F(a' - 4/3, 1/3), e^{-a' + 4/3}L(1/3)\} \ge I_1$. Hence we have from the foregoing estimate for I_2 and by (2.7)

$$I = \min(I_1, I_2) \ge \min(I_1, G(a')) \ge \min(R, G(a')).$$

On the other hand clearly $I \leq \min(I_1, G(a'))$.

Now G(x) is monotone increasing in x (this follows directly from the definition (2.8) by use of $e^{-x}F(x-z, 1) = e^{-z}E(x-z, 1)$). Thus $G(a') \leq R \leq I_1$ for all $a \leq a_1 \leq 3.6n$. In the complementary case $a > a_1$ we have $G(a') > R = I_1$ since then the numbers $r = r^* = 0.4832939$... and a' = y = 0.3598078... from (2.10) satisfy the constraints in (2.14): a' > 3.428985... > $a' + \varphi(r^*)$. Thus the assertion of the theorem is proved.

Proof of Lemma 5. We treat the case $r_1 = 0$ first. Then we must have $F(a' - \varphi(1/3), 1/3) < e^{-a' + \varphi(1/3)}L(1/3)$ and so for all feasible a' in (2.17) by Lemmas 1 and 2

$$F(\alpha', r) \leqslant F(a' - 1 - r, r)$$

$$\leqslant F(a' - 4/3, 1/3)$$

$$< e^{-a' + 4/3} E(4/3, 1/3)$$

$$\leqslant e^{-a' + 1 + r} E(a' - (a' - 1 - r), r) \leqslant e^{-\alpha'} E(a' - \alpha', r).$$

From this we find

$$H(r) = \inf_{\substack{r+1 \leqslant a' - \alpha' \leqslant \min(\varphi(r), a')}} e^{-\alpha'} E(a' - \alpha', r)$$
$$= e^{-a' + 1 + r} E(1 + r, r)$$

which is strictly increasing for $r \in [1/3, 1]$. Thus the assertion of Lemma 5 holds.

Now let $r_1 \ge 1/3$. Let us assume further on that $r \le r_2$ where $r_2 \in [r_1, 1]$ is defined by

$$F(a'-1-r_2,r_2) = e^{-a'+1+r_2}E(1+r_2,r_2)$$
(2.19)

and $r_2 := 1$ if no such r_2 exists. Then for every $r \in [r_1, r_2]$ we have $F(\alpha', r) \leq (\geq) e^{-\alpha'} E(\alpha' - \alpha', r)$ where $\alpha' = \alpha' - \min(\varphi(r), \alpha')$ $(\alpha' = \alpha' - 1 - r)$ respectively. This shows that for any $r \in [r_1, r_2]$ there is a unique $\beta(r) \in [\alpha' - \min(\varphi(r), \alpha'), \alpha' - 1 - r]$ such that

$$H(r) = F(\beta(r), r) = e^{-\beta(r)} E(a' - \beta(r), r).$$
(2.20)

From this defining equation and with the help of the formulas in Lemmas 1 and 2 we derive

$$\beta'(r) = \frac{e^{-\beta(r)} \partial E/\partial r - \partial F/\partial r}{\partial F/\partial y + e^{-\beta(r)} E(a' - \beta(r), r) + e^{-\beta(r)} \partial E/\partial y}$$
$$= \frac{\ln((p + \tilde{h})/(p - \tilde{h})) + \ln((h + q)/(h - q))}{\tilde{h}/\beta(r) + h/(a' - \beta(r))}$$
(2.21)

where (with values for y according to (2.20))

$$p := \beta(r) + 1 + r, \qquad \tilde{h} := \sqrt{p^2 - 4r\beta(r)} = \sqrt{(\beta - 1 - r)^2 + 4\beta(r)}$$
(2.22)

$$q := a' - \beta(r) - 1 - r, \qquad h := \sqrt{q^2 + 4r(a' - \beta(r))}.$$
 (2.23)

An easy calculation using (2.21) shows

$$\operatorname{sgn} H'(r) = \operatorname{sgn} \left[\frac{a' - \beta(r)}{h} \ln \left(\frac{h + q}{h - q} \right) - \frac{\beta(r)}{h} \ln \left(\frac{p + \tilde{h}}{p - \tilde{h}} \right) \right]. \quad (2.24)$$

We want to show now that the expression in brackets which we shall denote by K(r) has at most one zero. For this it will be sufficient to prove that $(\beta(r) \equiv \beta)$

$$0 > \frac{d}{dr}K(r) = -\frac{(\beta'h + (a' - \beta)h')}{h^2} \ln\left(\frac{h + q}{h - q}\right) + \frac{q'h - qh'}{2rh} - \frac{(\beta'\tilde{h} - \beta\tilde{h}')}{\tilde{h}^2} \ln\left(\frac{p + \tilde{h}}{p - \tilde{h}}\right) - \frac{p\tilde{h}' - p'\tilde{h}}{2r\tilde{h}}.$$
 (2.25)

We need some auxiliary inequalities. In view of $r \in [r_1, r_2]$ we have

$$1 + r \leqslant a' - \beta \leqslant \varphi(r) = (1 + r)^2 / 2(1 - r).$$
(2.26)

Since $e^{1+r}F(1+r,r) \ge e^{4/3}F(4/3,1) > 1 \ge L(r) \ge E(a'-\beta,r) = e^{\beta}F(\beta,r)$ we know further

$$1 + r > \beta \equiv \beta(r). \tag{2.27}$$

Now there holds $\beta' > 0$ by (2.21) and

$$\beta' h + (a' - \beta) h'$$

= (1/h){\beta' [(a' - \beta)(r - 1) + (1 + r)^2] + (a' - \beta)(a' - \beta + 1 + r)}

so that by (2.26) the first term in (2.25) is negative. From (2.26) and (2.27) it follows that $\beta < a' - \beta$ and hence that

$$\frac{\tilde{h}^2}{\beta^2} = 1 + \frac{2(1-r)}{\beta} + \left(\frac{1+r}{\beta}\right)^2$$
$$\geqslant 1 - \frac{2(1-r)}{a'-\beta} + \left(\frac{1+r}{a'-\beta}\right)^2 = \frac{h^2}{(a'-\beta)^2}.$$

In view of the inequality $\ln((1+x)/(1-x)) \ge 2x$ for $x \in [0, 1)$ this implies by (2.21) that $(x \equiv \tilde{h}/p)$

$$\beta' \ge \frac{\ln((p+\tilde{h})/(p-\tilde{h}))}{2\tilde{h}/\beta} \ge \frac{\beta}{p}.$$
(2.28)

From this estimate and the formula

$$\beta'\tilde{h} - \beta\tilde{h'} = (1/\tilde{h})[\beta'[p^2 - 2\beta p + \beta(\beta + 1 - r)] - \beta(p - 2\beta)\}$$

we see that also the third term in (2.25) is negative. The remaining terms are

$$\frac{q'h-qh'}{2rh} - \frac{p\bar{h}'-p'\bar{h}}{2r\bar{h}}$$

$$= -\beta' \left\{ \frac{1+r+a'-\beta}{h^2} - \frac{1+r-\beta}{\bar{h}^2} \right\}$$

$$- \left\{ \frac{(a'-\beta)(a'-\beta-1+r)}{rh^2} - \frac{\beta(\beta+1-r)}{r\bar{h}^2} \right\}$$
(2.29)

so that in order to establish (2.25) we only need to show that both terms in brackets are positive. The second one we can transform to

$$\frac{(a'-\beta)(a'-\beta-1+r)}{rh^2} - \frac{\beta(\beta+1-r)}{rh^2}$$
$$= \frac{1}{rh^2h^2} \{\tilde{h}^2a'(a'-2\beta-1+r) - \beta(\beta+1-r)(h^2-h^2)\}.$$

But because of $\tilde{h}^2 \ge \beta(\beta + 1 - r)$ and $h^2 - \tilde{h}^2 = a'^2 - 2a'(\beta + 1 - r) \le a'(a' - 2\beta - 1 + r)$ this expression is positive.

For the first term in brackets in (2.29) we have

$$\frac{1+r+a'-\beta}{h^2} - \frac{(1+r-\beta)}{h^2}$$
$$= \frac{a'}{h^2h^2} \left[-\beta^2 + \beta[2(1+r)+a'] + (1+r)(3-r-a')]\right].$$

By (2.27) the function $Q(\beta) := -\beta^2 + \beta [2(1+r) + a'] + (1+r)(3-r-a')$ is strictly monotone increasing for all $\beta = \beta(r)$ of (2.20). Assuming then that

$$\beta = \beta(r) \ge 0.64r, \quad r \in [r_1, r_2]$$
 (2.30)

we see that the term in question is positive if

$$Q(0.64r) = -0.1296r^2 + r(3.28 - 0.36a') + 3 - a' > 0$$

for all $r \ge 1/3$. But since Q(0.64r) is monotone increasing for $a' \le 8$ this is true if $Q(0.64/3) \ge -0.0144 + 1.09 + 3 - 1.12a' > 0$ which is the case for $a' \le 3.6$. Hence under assumption (2.30) we have proved that (2.25) holds. By (2.24) this means that H(r) is either monotone (decreasing or increasing) in r or else has only one interior extremum which must be a maximum then. Hence for $a' \le 3.6$ and for any subinterval $[c, d] \subset [r_1, r_2]$

$$\inf_{c\leqslant r\leqslant d}H(r)=\min(H(c),H(d)). \tag{2.31}$$

In order to prove (2.30) let us assume the contrary. Then (2.20), (2.26) imply

$$e^{0.64r}F(0.64r,r) \ge E(a'-\beta(r),r) \ge E(r+1,1)$$
(2.32)

for some $r \in [r_1, r_2]$. Then we consider the function (for any fixed b > 0)

$$\Psi(r) := e^{br} F(br, r) / E(r+1, r).$$

Similarly to the computation of (2.21), (2.24) we get with the help of the formulas of Lemmas 1 and 2

$$\Psi'(r) = \frac{e^{br}F(br,r)}{E(r+1,r)} \left\{ b + \frac{\tilde{h}}{r} + \ln \frac{p-\tilde{h}}{p+\tilde{h}} - \frac{h}{1+r} + 1 \right\}$$

where, according to the choice of y, we have to set (cf. (2.22), (2.23))

$$p := 1 + (1+b)r, \quad \tilde{h} := \sqrt{p^2 - 4br^2}, \quad h = 2\sqrt{r(r+1)}.$$
 (2.33)

We shall now show $\Psi'(r) > 0$ for b = 0.064. We have

$$\frac{d}{dr}\left\{b+\frac{\tilde{h}}{r}+\ln\frac{p-\tilde{h}}{p+\tilde{h}}-\frac{h}{1+r}+1\right\}$$
$$=\frac{h-h'(1+r)}{(1+r)^2}+\frac{\tilde{h}'r-\tilde{h}}{r^2}+\frac{p'\tilde{h}-p\tilde{h}'}{2r^2b}.$$

With the help of (2.33) we can verify that

$$[h - h'(1 + r)](1 + r)^{-2} = -2[h(1 + r)]^{-1} < 0$$

and that for $b \in (0, 1)$

$$\frac{(\tilde{h'}r-\tilde{h})\,2b+p'\tilde{h}-p\tilde{h'}}{2r^2b}=-\frac{(p-2r)}{r^2\tilde{h}}=\frac{-1+(1-b)\,r}{r^2\tilde{h}}<0.$$

Thus $\Psi'(r)$ has at most one zero for any $b \in (0, 1)$ and will turn from positive to negative values there. But for b = 0.64

sgn
$$\Psi'(1) =$$
sgn $\left\{ b + \sqrt{b^2 + 4} + \ln \frac{b + 2 - \sqrt{b^2 + 4}}{b + 2 + \sqrt{b^2 + 4}} - \sqrt{2} + 1 \right\}$
= sgn {0.15...} > 0

which shows that $\Psi'(r) > 0$ for b = 0.64. Now this implies a contradiction to (2.32) because

$$\max_{1/3 \leqslant r \leqslant 1} \frac{e^{0.64r} F(0.64r, r)}{E(r+1, r)} = \frac{e^{0.64} F(0.64, 1)}{E(2, 1)}$$
$$= 0.9605... < 1.$$

Hence our assumption $\beta(r) \ge 0.64r$ in (2.30) is proved. In order to complete

the proof of Lemma 5 it remains to study H(r) on the interval $[r_2, 1]$ (if it is non-void). We observe that

$$H(r) \ge \tilde{H}(r) := \inf_{0 \le a' - \alpha' \le \min(\varphi(r), a')} \max\{F(\alpha', r), e^{-\alpha'}E(a' - \alpha', r)\}.$$
(2.34)

Then for $r \ge r_1$ there is always a feasible $\tilde{\beta}(r)$ satisfying

$$\tilde{H}(r) = F(\tilde{\beta}(r), r) = e^{-\tilde{\beta}(r)}E(a' - \tilde{\beta}(r), r).$$

For $\tilde{\beta}'(r)$ and $\tilde{H}'(r)$ there hold the same formulae as in (2.21), (2.24). However, if $r \in [r_2, 1]$ with $r_2 < 1$ by (2.19) we must have then $\tilde{\beta}(r) > a' - 1 - r$ (in contrast to (2.26)) whence q < 0 in (2.23). An immediate consequence of (2.24) is then sgn $\tilde{H}'(r) < 0$ so that by (2.34) for all $r \in [r_2, 1]$

$$H(r) \ge \tilde{H}(r) \ge \tilde{H}(1) = G(a').$$

Together with (2.31) this finally establishes the complete assertion of the lemma.

Remark. The proof of Lemma 5 is complicated by the fact (cf. (2.24)) that for H(r) we can only show that it is either monotone or concave. Indeed, this behaviour is confirmed by numerical results for the critical values of a.

3. Optimal Piecewise Padé-Approximation

The local error estimates of the preceding section enable us to achieve the final goal of this paper, namely, to investigate whether the rate of approximation (on $[0, \infty)$) can be improved by using piecewise Padé-approximation with equal total number of parameters. We look for the optimal distribution of the pieces and the degrees of the approximating rational functions. On every single interval $I = [b, c] \subseteq [0, \infty)$ we use Padé-approximation with optimal choice of the center and the ratio r of the degrees of the numerator- and denominator-polynomial. Hence as a generalization of (2.6) for n = 0, 1, 2,... we consider the local error

$$E(n+1,I) := \inf_{1/n \le r \le 1} \inf_{b \le \alpha \le c} \|e^{-x} - e^{-\alpha}R_{[rn],n}(-x+\alpha)\|_{\infty,[b,c]}$$
(3.1)

with $R_{[rn],n}(x)$ defined by (1.1). Our aim is then to determine the asymptotic behaviour $(N \to \infty)$ of the error

$$E_N := \inf \left\{ \sup_i E(n_i, I_i) : \sum n_i \leq N, \bigcup_i I_i = [0, \infty) \right\}.$$
(3.2)

The crucial point in the subsequent analysis is the following scaling property (easily verified from the definition (3.1))

$$E(n+1, I+a) = e^{-a}E(n+1, I), \qquad a > 0$$
(3.3)

where I + a denotes the interval I translated by a. In [12] it has been shown that for any family with property (3.3) the sequence $\{E_N\}_{N \in \mathbb{N}}$ can be determined as a fixed point of the following operator T defined on the class of sequences $\vec{e} := \{e_N\}_{N \in \mathbb{N}}$ of numbers

$$(T\vec{e})_{N} := \begin{cases} \min(e_{N}, e_{N}), & N \ge 2\\ e_{1}, & N = 1 \end{cases}$$
(3.4)

where

$$\underline{e}_{N} := \inf_{\substack{0 < a < \infty \\ 1 \leq v \leq N-1}} \{ e^{-a} e_{v} : e^{-a} e_{v} = E(N - v, [0, a]) \}.$$
(3.5)

The application of the operator T to a sequence $\{e_{\nu}\}$ of (global) errors may be interpreted as the attempt to improve the error e_{N} by adding a new subinterval with $N - \nu$ parameters in an optimal way to the partition with ν parameters and error e_{ν} .

A further important property established in [12] is the monotonicity of T, i.e., for any two sequences $\vec{e}^{(1)} \leq \vec{e}^{(2)}$ (to be understood component-wise), there holds

$$T\vec{e}^{(1)} \leqslant T\vec{e}^{(2)}.$$
 (3.6)

Finally, we need

LEMMA 6. Define the operator \tilde{T} as in (3.4), (3.5) but with E(n, [0, a]) replaced by $\tilde{E}(n, [0, a])$ where $K\tilde{E}(n, [0, a]) \leqslant E(n, [0, a])$ for some constant K > 0. Then, if a sequence \tilde{e} is a fixed point of \tilde{T} , the sequence \tilde{e} with $e_v := K\tilde{e}_v$ is a fixed point of the operator T.

This fact has not been explicitly formulated in [12] but follows from the easily verified inequality $e_N \ge e_N$ for e_v defined as above. Now we prove

LEMMA 7. Define \tilde{T} as in Lemma 7 with

$$\tilde{E}(n, [0, a]) := \min(R, G(a/n))^n$$
(3.7)

where R and the function G(x) are introduced in Theorem 4. For $\gamma \ge 0$ denote by $\vec{e}^{(\gamma)}$ the sequence of numbers $e_{\gamma,n} := e^{-\gamma n}$. Then $\vec{e}^{(\gamma)}$ is a fixed point of \tilde{T} for any $\gamma \ge \gamma_0$, where $\gamma_0 = 1.88716923...$ is defined by

$$e^{-\gamma_0} = G(\gamma_0). \tag{3.8}$$

Proof. In view of the definition

$$\underline{e}_{\gamma,N} := \inf_{\substack{0 < a < \infty \\ 1 \leq \nu \leq N-1}} \{ e^{-a} e_{\gamma,\nu} : e^{-a} e_{\gamma,\nu} = \tilde{E}(N-\nu, [0, a]) \}$$

we have (N - v = n, a' := a/n)

$$\frac{\underline{e}_{\gamma,N}}{e_{\gamma,N}} = \lim_{\substack{0 < a' < \infty \\ 1 \le n \le N-1}} \{ e^{(\gamma - a')n} : (\gamma' - a) \ n = \gamma N + n \log \min(R, G(a')) \}.$$
(3.9)

The constraint here is equivalent to

$$-(N-n) \gamma/n = a' + \log \min(R, G(a')).$$
(3.10)

Since the left-hand side of (3.10) is always ≤ 0 and since G is a monotone increasing function this means that $a' \leq \gamma_0$ where

$$0 = \gamma_0 + \log \min(R, G(\gamma_0)).$$

However,

$$G(\gamma_0) = e^{-\gamma_0} = (6.600657267...)^{-1} < R^{-1} = (4.0982...)^{-1}.$$
 (3.11)

(We remark that by (2.8) γ_0 can be easily computed successively from the equations $1 = F(\gamma_0 - z, 1)$, $e^z F(z, 1) = e^{-(\gamma_0 - z)} \cdot F(\gamma_0 - z, 1)$ in the unknowns $\gamma_0 - z$ and z.)

The restriction $a' \leq \gamma_0$ shows now that $\gamma \geq \gamma_0$ in (3.9) implies $\underline{e}_{\gamma,N} \geq \underline{e}_{\gamma,N}$. But by definition (3.4), (3.5) for \tilde{T} this is just the assertion of the lemma.

The main result of this section is now easily established:

THEOREM 5. The error E_N of optimal piecewise Padé-approximation of e^{-x} on $[0, \infty)$ defined by (3.1), (3.2) can be estimated by

$$C_3 e^{-\gamma_0 N} \leqslant E_N \leqslant C_4 e^{-\gamma_0 N} \tag{3.12}$$

with positive constants C_3 , C_4 of size $O(N^3 \log N)$. It can be obtained (up to these constants) by approximating e^{-x} on $[0, \gamma_0 N]$ by the Padé-approximant of Theorem 4 and on $[\gamma_0 N, \infty)$ by the zero function.

Proof. From Lemmas 6 and 7 we conclude that

$$e_n = K e^{-\gamma_0 n}, \qquad K \leqslant \min(1, a) \tag{3.13}$$

is a fixed point of the operator T defined by (3.1), (3.2) and (3.4), (3.5). Here, we have used estimate (2.9) of Theorem 4 and assumed that K is a constant not depending on a. But this assumption is justified in view of the

98

fact that for any fixed point of \tilde{T} the feasible *a* may be bounded from below by $a \ge \gamma_0 n$ (see (3.9), (3.10)).

On the other hand, from the monotonicity of T we know (cf. [12, Corollary 3]) that any fixed point \vec{e} of T with $e_n \leq C_5 R^n \leq E(n, [0, \infty))$ (with C_5 according to Theorem 4) gives a lower bound for E_N . Since this condition for e_n is satisfied by (3.11) and (3.13) the lower estimate of the theorem is established.

Concerning the upper bound we consider the special approximation on the intervals $[0, \gamma_0 N]$ and $[\gamma_0 N, \infty)$ as described above. By Theorem 4 we immediately verify that this yields the estimate

$$E_N \leq C_4 \max(G(\gamma_0), e^{-\gamma_0})^N$$

where C_4 has size $O(N^3 \log N)$.

Theorem 5 shows that passing from global Padé-approximants to Padésplines we can improve the rate of approximation of e^{-x} on $[0, \infty)$ from R > 1/(4.1) (see (2.9)) to $G(\gamma_0) < 1/(6.6)$ (see (3.11)). If one uses best rational approximation instead, a further improvement should be possible with a better rate, than that of best rational approximation on $[0, \infty)$ which is still unknown. We cannot hope to achieve this by using a piecewise rational approximation which employs Padé-approximation on the initial interval [0, a]. This can be seen by the same arguments as above since applying repeatedly the operator T to such an initial approximation would again lead to the lower bound of Theorem 5.

From the above analysis it is also clear that we have to study the local error on intervals of the form $[0, \alpha n]$ $(n \to \infty \text{ and } \alpha \text{ fixed})$ so that the sharp estimates of [2, 5] cannot be used for this purpose. Better estimates than for classical Padé-approximation might be obtained by using Padé-Laguerreapproximation for which explicit formulae have been obtained by Németh [6]. They are of the same nature as (1.1), (1.2) but involve complex integrals. An asymptotic estimate for the pointwise error is then derived as in Theorems 1 and 2 of Section 1. However, the "méthode du col" has then to be replaced by its modification to the complex case known as the "saddlepoint method." From the formulas in [6] one can easily derive that the error for the (m, n) Laguerre-Padé-approximation has at least m + n + 1 zeros on $[0, \infty)$. However, it is not clear which one of the m + n extrema of the error is estimated in [6] where the method is applied only formally. In addition the estimate in [6] becomes definitely wrong for general intervals $[0, \alpha n], \alpha > 0$. The correct applications of the saddle-point method (cf. [9]) certainly still requires much work.

OPITZ AND SCHERER

References

- 1. H. P. BLATT AND D. BRAESS, Zur Rationalen Approximation von e^{-x} auf $[0, \infty)$, J. Approx. Theory **30** (1980), 169–172.
- 2. D. BRAESS, On the conjecture of Meinardus on rational approximation of e^x , preprint, Ruhr-Universität Bochum, 1982.
- 3. W. J. CODY, G. MEINARDUS, AND R. S. VARGA, Chebyshev rational approximation to e^{-x} in $[0, \infty)$ and applications to heat conduction problems, J. Approx. Theory 2 (1969), 50-65.
- J. GILEWICZ, Approximants de Padé, in "Lecture Notes in Mathematics No. 667," Springer-Verlag, Berlin/New York, 1978.
- 5. G. NÉMETH, Relative rational approximation of the function e^x , Math. Notes 21 (1977), 325-328.
- G. NÉMETH, Notes on generalized Padé approximation, in "Approximation and Function Spaces," Proc. Int. Conf. Gdansk, 1979 (Z. Ciesielsky, Ed.), pp. 484–508.
- 7. D. J. NEWMAN, Rational approximation to e^{-x} , J. Approx. Theory 10 (1974), 301-303.
- W.-C. NI, E. B. SAFF AND R. S. VARGA, Geometric convergence of rational approximation to e^{-z} in infinite sectors, Numer. Math. 26 (1976), 211-225.
- 9. F. W. J. OLVER, "Asymptotics and Special Functions," Academic Press, New York, 1974.
- 10. O. PERRON, "Die Lehre von den Kettenbrüchen, II, "Teubner, Leipzig, 1957.
- 11. Q. I. RAHMAN AND G. SCHMEISSER, Rational approximation to e^{-x} , II, Trans. Amer. Math. Soc. 235 (1978), 395–402.
- 12. K. SCHERER, On optimal global error bounds obtained by scaled local error estimates, *Numer. Math.*, in press.
- 13. A. SCHÖNHAGE, Zur rationalen Approximierbarkeit von e^{-x} über $[0, \infty)$, J. Approx. Theory 7 (1973), 395-398.
- 14. A. SCHÖNHAGE, Rational approximation to e^{-x} and related \mathcal{L}^2 -problems, preprint, Universität Tübingen, 1981.