A Note on the Relative Merits of Padé and Maehly's Diagonal Convergents in Computing e^x

By R. Sankar and V. Malini

Abstract. Methods for calculating functions to a high degree of accuracy have assumed increased importance following the advent of the computers. It has been found that rational approximations require fewer operations on a computer than the older polynomial approximations. Among the known methods those due to Padé [1] and Maehly [2] are perhaps the most important. In this paper we have analyzed these methods as applied to the exponential function. It is observed that Maehly's method is superior to the Padé method in the sense of yielding better accuracy over a given range on the real axis for a given order of approximation. Maehly's formulas for computing e^x correct to eight decimal places have been worked out.

1. Introduction. A direct calculation of an nth degree polynomial

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n, \tag{1}$$

would require $(2n - 1)$ multiplications. The time for this calculation in a computer will be substantially equal to the time for $(2n - 1)$ multiplications, since the time for addition and subtraction will generally be small in comparison. $f(x)$ can however be evaluated with only n multiplications by the method of nested multiplication defined by the formula [3]

$$f(x) = [\{(a_nx + a_{n-1})x + a_{n-2}\}x + \cdots + a_0], \tag{2}$$

which has the added advantage of not requiring any intermediate recording.

But a rational approximation $P_m(x)/Q_n(x)$, where the suffixes denote the degree of the polynomials is equivalent to a polynomial approximation of degree $(m + n)$. By expressing the rational function $P_m(x)/Q_n(x)$, as a continued fraction it can be evaluated in m or $n + 1$ operations according as $m \geqq n$ or $m < n$. Thus we find that a polynomial approximation of degree $2n$, requiring $2n$ operations can, if transformed into a rational approximation, be calculated in only n operations. This economy of effort achieved by rational approximations makes it important to investigate the relative merits of available methods for common functions. For e^x the Padé method has been studied by Kogbetliantz [4]. We have analyzed the exponential function and established that Maehly's method is superior to the Padé method in the sense of yielding better accuracy over a given range on the real axis for a given order of approximation.

2. The Padé Method. From Kogbetliantz[4] we have the Padé formula,

$$e^x = \frac{P_m(x)}{P_m(-x)} + R_m(x), \tag{3}$$

Received December 8, 1962, revised April 10, 1963.

where

$$R_m(x) = (-1)^m \frac{m!\Gamma\frac{1}{2}}{(2m)!} \cdot \frac{e^{x/2}}{P_m(-x)} \cdot x^{m+1/2} I_{m+1/2}(x/2)$$

and

$$P_m(x) = \frac{m!}{(2m)!} \sum_{r=0}^{m} \frac{(2m-r)!}{r!(m-r)!} x^r.$$

It can be shown that $e^{x/2}/|P_m(-x)|$ is monotonic increasing for $x < 2$. Also it is easy to see that

$$\left| x^{m+1/2} I_{m+1/2}\left(\frac{x}{2}\right) \right|$$

is symmetric about $x = 0$ and is monotonic increasing for $x > 0$. So it follows that for $0 < x < 2$

$$|R_m(-x)| < |R_m(x)| < |R_m(2)|.$$

Hence for $-a \leqq x \leqq a < 2$, $|R_m(x)| \leqq |R_m(a)|$.
Again since

$$0 < I_{m+1/2}\left(\frac{a}{2}\right) \Gamma(m + \tfrac{3}{2}) \cdot \left(\frac{4}{a}\right)^{m+1/2} < e^{a^2/[8(2m+3)]},$$

it follows that

$$|R_m(a)| < \left[\frac{m!}{(2m)!}\right]^2 \frac{1}{2m+1} \cdot \frac{a^{2m+1}}{|P_m(-a)|} \cdot e^{a/2+a^2/[8(2m+3)]}$$

Thus for $-a \leqq x \leqq a < 2$, and for all m,

$$|R_m(x)| < \left[\frac{m!}{(2m)!}\right]^2 \frac{1}{2m+1} \cdot \frac{a^{2m+1}}{|P_m(-a)|} \cdot e^{a/2+a^2/[8(2m+3)]}. \tag{4}$$

3. Maehly's Method. We know that

$$e^{ax} = I_0(a) + 2\sum_{n=1}^{\infty} I_n(a) T_n(x), \qquad -1 \leqq x \leqq 1.$$

where $T_n(x)$ is the Chebyshev Polynomial defined by

$$T_n(x) = \cos(n \cos^{-1} x).$$

Hereafter we shall use I_m to denote $I_m(a)$. We assume a rational approximation of the form

$$\frac{\sum_{r=0}^{m} a_r T_r(x)}{\sum_{r=0}^{m} b_r T_r(x)}.$$

The coefficients a_r's and b_r's are evaluated by identifying the Chebyshev expansion of

$$e^{ax}\sum_{r=0}^{m} b_r T_r(x)$$

with

$$\sum_{r=0}^{m} a_r T_r(x)$$

up to the term $T_{2m}(x)$.

These provide us with the following $(2m + 1)$ equations to determine the a_r's and b_r's.

$$(5)\qquad \begin{cases} a_0 = \sum_{s=0}^{m} b_s I_s\,, \\ a_r = \sum_{s=0}^{m} b_s[I_{r+s} - I_{|r-s|}] \qquad r = 1, 2, 3, \cdots, m. \end{cases}$$

$$(6)\qquad \sum_{s=0}^{m} b_s[I_{r+s} - I_{r-s}] = 0, \qquad r = m + 1, m + 2, \cdots, 2m,$$

and

$$(7)\qquad d_0 = \sum_{s=0}^{m} b_s[I_{2m+1+s} + I_{2m+1-s}],$$

where d_0 is defined among other d's by the equation,

$$e^{ax} = \frac{\sum_{r=0}^{m} a_r T_r(x)}{\sum_{r=0}^{m} b_r T_r(x)} + \frac{\sum_{r=0}^{\infty} d_r T_{2m+1+r}(x)}{\sum_{r=0}^{m} b_r T_r(x)}.$$

Since the d_r's and b_r's decrease rapidly and $b_0 = 1$, the error

$$\frac{\sum_{r=0}^{\infty} d_r\, T_{2m+1+r}(x)}{\sum_{r=0}^{m} b_r\, T_r(x)}$$

can be approximated by $d_0 T_{2m+1}(x)$. Again since $|\, T_{2m+1}(x)\,| \leqq 1$, the error is bounded by d_0. From equations (6) and (7) we have

$$(8)\qquad |\, d_0\,| = |\, 2D/\Delta\,|,$$

where

$$D = \begin{vmatrix} (I_m + I_{m+2}), & (I_{m-1} + I_{m+3}), & \cdots & (I_1 + I_{2m+1}), & I_{m+1} \\ (I_{m+1} + I_{m+3}), & (I_m + I_{m+4}), & \cdots & (I_2 + I_{2m+2}), & I_{m+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (I_{2m-1} + I_{2m+1}), & (I_{2m-2} + I_{2m+2}), & \cdots & (I_m + I_{3m}), & I_{2m} \\ (I_{2m} + I_{2m+2}), & (I_{2m-1} + I_{2m+3}), & \cdots & (I_{m+1} + I_{3m+1}), & I_{2m+1} \end{vmatrix}$$

and, Δ is the minor of D obtained by deleting the last row and last column of D. If we transform D into a triangular matrix, the last element in the principal diagonal

TABLE 1
Error bounds for Padé and Maehly's methods

a	m			
	2	3	4	5
0.1	1.5×10^{-8} 8.7×10^{-10}	1.1×10^{-12} 1.5×10^{-14}	4.3×10^{-17} 1.4×10^{-19}	1.1×10^{-21} 8.8×10^{-24}
0.2	5.4×10^{-7} 2.8×10^{-8}	1.8×10^{-10} 2.2×10^{-12}	2.4×10^{-14} 9.6×10^{-17}	2.5×10^{-18} 7.2×10^{-21}
0.3	4.6×10^{-6} 2.1×10^{-7}	2.9×10^{-9} 3.2×10^{-11}	1.0×10^{-12} 3.2×10^{-15}	2.4×10^{-17} 3.2×10^{-20}
0.4	2.1×10^{-5} 8.4×10^{-7}	3.2×10^{-8} 1.2×10^{-10}	1.6×10^{-11} 4.7×10^{-14}	6.2×10^{-15} 2.9×10^{-18}
0.5	7.2×10^{-5} 2.7×10^{-6}	6.4×10^{-7} 1.2×10^{-9}	1.3×10^{-10} 2.0×10^{-13}	8.0×10^{-14} 1.0×10^{-16}

gives the value of D/Δ. The values of I_n's were taken from the British Association Tables [7].

4. Comparison of the Methods. The error bounds for Padé and Maehly's methods given by equations (4) and (8) were computed for some values of a and m, and are shown in Table 1. In each equare, the first entry corresponds to the Padé method and the second entry to Maehly's method. a stands for the range of applicability $-a \leqq x \leqq a$, and m for the order of the rational approximation.

It is seen from the table that Maehly's method is superior to the Padé method for the range $-a \leqq x \leqq a$, if $a \geqq \cdot 1$. Also the superiority of Maehly's method increases with m. On the basis of rough calculations it is felt that there should exist a small range $-a \leqq x \leqq a$, where the Padé method would be superior to Maehly's method. We have not considered large ranges, since large ranges can be reduced to small ones by any one of the conventional methods.

5. Maehly's Formulas for e^x. Maehly's formulas for $\sin x$, $\cos x$, $\tan x$ (correct to 10 decimal places) and $\cot x$, and $\log x$ (correct to 8 decimal places), are available [5]. We present here Maehly's formulas for e^x correct to 8 decimal places. The error table shows that the cases $m = 2$, $a = .1$ and $m = 3$, $a = .5$ give eight decimal place accuracy. In the computation the values of I_n correct to 10 decimal places were taken from the British Association Tables [7].

For $m = 2$, $a = .1$ we have by solving equations (5) and (6)

$$e^{.1x} = \frac{8.32916782 \times 10^{-4}x^2 + 4.998125716 \times 10^{-2}x + 9.995836456 \times 10^{-1}}{8.327087326 \times 10^{-4}x^2 - 4.997709787 \times 10^{-2}x + 9.995836456 \times 10^{-1}}$$

$$= 1.000249846 + \frac{1.200549936 \times 10^{2}}{(x - 6.001500366 \times 10)} + \frac{1.200250173 \times 10^{3}}{(x - 2.498148174 \times 10^{-3})} \tag{9}$$

$$-1 \leqq x \leqq 1.$$

For $m = 3$, $a = .5$ we have by solving equations (5) and (6)

$$e^{.5x} = \frac{1.027910364 \times 10^{-3}x^3 + 2.470655174 \times 10^{-2}x^2 + 2.470292178 \times 10^{-1}x + 9.876743802 \times 10^{-1}}{-1.023329716 \times 10^{-3}x^3 + 2.465124084 \times 10^{-2}x^2 - 2.468079636 \times 10^{-1}x + 9.876743797 \times 10^{-1}}$$

$$(10) \qquad = -1.004476219 - \frac{4.834036983 \times 10}{(x - 2.407138535 \times 10) + \dfrac{2.007302412 \times 10^2}{(x - 8.923162505 \times 10^{-3}) + \dfrac{4.002103141 \times 10}{(x - 8.937067425 \times 10^{-3})}}}$$

$$-1 \leqq x \leqq 1.$$

The values of e^x calculated with formulas (9) and (10) agree to 8 decimal places with the 18 figure tables of the exponential function published by the National Bureau of Standards [6].

Acknowledgments. The authors are thankful to Dr. P. Nilakantan, Director, National Aeronautical Laboratory, for his kind interest and encouragement. They also thank the referee for pointing out a flaw in the first draft of equation (4).

National Aeronautical Laboratory,
Bangalore, India

1. H. Padé, "Sur la répresentation approchee dúne function par des fractions rationnelles," *Ann. Sci. École. Norm. Sup.*, Paris, v. 9, 1892, p. 1–93; v. 16, 1899, p. 315–426.
2. H. Maehly, *First Interim Progress Report on Rational Approximations*, Project NR 44-196, Princeton University, June 23, 1958.
3. National Physical Laboratory, *Modern Computing Methods*, Notes on Applied Science, No. 16, Her Majesty's Stationery Office, London, 1961, p. 53.
4. E. G. Kogbetliantz, "Computation of e^N for $-\infty < N < \infty$ using an electronic computer," IBM *J. Res. Develop.*, v. 1, no. 2, 1957, p. 110–115.
5. Kurt Spielberg, "Efficient continued fraction approximations to elementary functions," *Math. Comp.*, v. 15, 1961, p. 409–417.
6. Nat. Bur. Standards, "Tables of the exponential function e^x," *Appl. Math. Ser.* 14, Department of Commerce, Washington, D. C., 1951.
7. British Association for the Advancement of Science, Comittee on Mathematical Tables, Vol. X, *Bessel Functions, Part II*, Cambridge University Press, 1952, p. 220–237.