E. G. Kogbetliantz

Computation of e^N for $-\infty < N < +\infty$ Using an Electronic Computer

Machine	Approx.	Computation	Single Precision			Double Precision		
			M	PC	Dg	M	PC	Dg
Binary	R	floating point	3	19	9	5	21	19
			4	8	10	6	8	18
		fixed point	3	35	10	6	10	21
			4	8	10			
Decimal	P	floating point	5	17	8	8	30	18
						9	24	17
		fixed point	5	25	10	9	32	20
			6	19	10			

Abstract: Rational *R* and polynomial *P* approximations to the exponential function e^N are studied. They allow e^N to be computed for any value of the exponent *N* in the infinite range from minus infinity to plus infinity in a minimum number *M* of multiplications (and divisions, for the rational approximations). This minimum is attained without unduly increasing the number *PC* of precomputed and stored constants and also without limiting the number *Dg* of the first correct significant digits. The main results are presented in the table at the left.

Introduction

The aim of this paper is to formulate shorter procedures for the evaluation of e^N which involve fewer operations and therefore consume less time. Approximations must always be adapted to particular features of the electronic computer used. Thus, for instance, if division is too slow then rational approximations cannot be used since they involve divisions. On the other hand, for the IBM 704, rational approximations are the most economical. In reducing the infinite range $(-\infty, +\infty)$ to a shorter range in which the approximation to be used is sufficiently accurate, different procedures must be used for binary and for decimal machines.

The number of precomputed constants involved in a subroutine and stored in the memory of the machine is also important. It is always possible to decrease the number of multiplications and divisions (additions are so rapid that we omit them from the consideration), increasing the number of precomputed constants; but, on the other hand, it is not advisable to load the memory of the machine by too many constants. These conflicting considerations make the choice of the best procedure a very important question. In this paper we disregard the case of routines based on the use of many precomputed constants. We will try to reduce the number of multiplications and/or divisions, without increasing the number of precomputed constants above an upper bound equal to 35, the choice of which is dictated by our experience.

Two mathematical tools are considered here: approximating polynomials are derived from the classical expansion of the exponential function into Fourier Series of Tchebychev polynomials $T_n(x)$, [1], while our rational approximations are the diagonal elements of the well known Padé Table for e^x, [2] . In the Tchebychev expansion

$$e^{ax} = I_0(a) + 2\sum_{m=1}^{\infty} I_m(a) \cdot T_m(x) \qquad (|x| \leqq 1) \qquad \text{(I)}$$

$I_m(z)$ is the modified Bessel function $J_m(t)$, namely $I_m(z) = i^{-m} \cdot J_m(iz)$, while $T_m(x) = \cos(m \cdot \arccos x)$.

Let $Q_n(a,x)$ denote the sum of first $n+1$ terms of the series (I), while $E_n(a,x)$ is the absolute error made in replacing e^{ax} by $Q_{n-1}(a,x)$, so that

$$e^{ax} = Q_{n-1}(a,x) + E_n(a,x). \qquad (1)$$

With $|x| \leqq 1$, E_n admits the following upper bound

$$|E_n(a,x)| \leqq 2 \cdot \sum_{m=n}^{\infty} I_m(a) \leqq 2I_n(a)/[1 - \tfrac{1}{2}(n+1)^{-1} \cdot a]. \qquad (2)$$

The (m,n)-th element in the m-th row and n-th column of the Padé Table for e^x is $P_{mn}(x)/P_{mn}(-x)$, where the polynomial of n-th degree $P_{mn}(x)$ is defined by

$$(m+n)! \cdot P_{mn}(x) = n! \sum_{s=0}^{n} (m+n-s)! x^s / [s!(n-s)!].$$

It is known [2] that

$$e^x \cdot P_{mn}(-x) = P_{mn}(x) + (-1)^m \cdot x^{m+n+1} \times \int_0^1 e^{ux} \cdot u^m \cdot (1-u)^n \cdot du/(m+n)!$$

Therefore, the absolute value of relative error, made in approximating e^x for given values of x and of the sum $m+n=s$, but for variable $m=0, 1, 2, \ldots s$, by $P_{mn}(x)/P_{mn}(-x)$, reaches its minimum for $n=m$, that is for a diagonal element $P_{mm}(x)=P_m(x)$ of the Padé Table. Now in

$$e^z=P_m(z)/P_m(-z)+e^z\cdot R_m(z), \tag{3}$$

where

$$(2m)!P_m(z)=m!\sum_{s=0}^{m}(2m-s)!z^s/[s!(m-s)!], \tag{4}$$

the relative error $R_m(z)$ is given by:

$$R_m(z)=(-1)^m\cdot e^{-z}\cdot z^{2m+1}\times\int_0^1 e^{uz}(u(1-u))^m\cdot du/[(2m)!P_m(-z)].$$

The substitution $2u=1+t$ shows that the definite integral is equal to

$$m!\Gamma(\tfrac{1}{2})\cdot e^{z/2}\cdot I_{m+\frac{1}{2}}(z/2)/z^{m+\frac{1}{2}}.$$

Therefore, the inequality

$$0<I_{m+\frac{1}{2}}(z/2)\cdot 2^{2m+1}\cdot z^{-m-\frac{1}{2}}\cdot\Gamma(m+3/2)<e^{z^2}/[8(2m+3)],$$

yields the following upper bound for $|R_m(z)|$:

$$|R_n(z)|<(2n+1)^{-1}\cdot(n!/(2n)!)^2\cdot|z|^{2n+1}\cdot e^{-z/2+z^2/[8(2n+3)]}. \tag{5}$$

The expression (3) will be used in a sufficiently small range $|z|\leqq 2^{-k}\cdot\ln 2$ where the choice of the positive integer k depends on the accuracy required. For a given accuracy we can minimize the number of multiplications and divisions in (1) or (3) by choosing a small value of n and sufficiently large value of k.

The number of precomputed constants is equal to 2^{k-1} plus a constant and it increases rapidly with k. This precludes the use of larger values of k. We shall consider only four cases: $k=2$, 3, 4 and 5. Four values $n=2$, 3, 4, 5 in (3) and two values $n=5$, 6 in (1) will be considered. The reader can easily extend our results to other values of k and n, if needed.

The rational approximations to e^z studied in this paper are not new. They are a very particular case of the generalized Taylor series formed in 1876 by the French mathematician Gaston Darboux [3]:

$$[f(z)-f(a)]\cdot P^{(n)}(0)=\sum_{m=1}^{n}(-1)^m[P^{(n-m)}(0)\cdot f^{(m)}(a)-P^{(n-m)}(1)\cdot f^{(m)}(z)]\cdot(z-a)^m+R_n$$

$$R_n=(-1)^n(z-a)^{n+1}\cdot\int_0^1 P(t)\cdot f^{(n+1)}[a+t(z-a)]\cdot dt.$$

Particularizing this expansion for $f(z)=e^z$, $P(t)=t^n\cdot(t-1)^n$, $a=0$, one obtains our approximations. Another French mathematician H. Padé [4] formulated in 1892 a general method for constructing a complete table of all rational functions approximating a function $f(x)$. The diagonal of his table for e^x contains our approximations.

In 1949, seventy three years after the Darboux paper was published, Messrs. P. M. Hummel and C. L. Seebeck rediscovered [5] the particular case $P(t)=t^n\cdot(t-1)^n$ of the Darboux expansion and, applying it to e^x, formed again the diagonal of the classical Padé table for e^x.

Reproducing again the work of Darboux in his book on numerical analysis [6], Dr. C. Lanczos ascribes it to Messrs. P. M. Hummel and C. L. Seebeck.

We add that the same sequence of approximations to e^x can be obtained from Lambert's well known continued fraction [7].

$$\tanh(x/2)=\frac{x/2|}{|1}+\sum_{n=1}^{\infty}\frac{(x/2)^2|}{|(2n+1)}.$$

Let us denote the n-th convergent of this continued fraction by $A_n(x)/B_n(x)$. Then $e^x=[1+\tanh(x/2)]\cdot[1-\tanh(x/2)]^{-1}$ is approximated as follows:

$$e^x\approx[B_n(x)+A_n(x)]/[B_n(x)-A_n(x)].$$

The second member is identical with our approximation.

Part A—Binary machine; rational approximations

1. Reduction to small range

Multiplying the exponent N in e^N, $-\infty<N<+\infty$, by $\log_2 e$ and denoting the integral part of the product by M

$$N\cdot\log_2 e=M+F=M+a(k)+f,\quad(0<F=a(k)+f<1) \tag{1}$$

we reduce first the infinite range of N in $e^N=2^M\cdot e^{F/\log_2 e}$ to the range (0, ln 2) of the exponent $F/\log_2 e=F\cdot\ln 2$. Choosing a fixed positive integer k, we subdivide the interval (0,1) into 2^k subintervals $[2^{-k}\cdot j;\ 2^{-k}\cdot(j+1)]$ with $0\leqq j\leqq 2^k-1$. Beginning with $f_0=F$, k numbers $f_1, f_2, \ldots f_{k-1}, f_k=f$ are computed successively in k additions by letting $f_{i+1}=f_i-s_i/2^{i+1}$, $(0\leqq i\leqq k-1)$, where s_i denotes the sign of f_i, namely: $s_i=\text{signum}\,(f_i)$.

Thus, we have in $F=a(k)+f$

$$f=f_k=F-a(k)=F-\sum_{i=0}^{k-1}s_i/2^{i+1}.$$

There are 2^{k-1} different possible values of $a(k)$ since $k-1$ signs s_i, $1\leqq i\leqq k-1$, are involved ($s_0=+1$) and each of them can take either one of two values ± 1. The 2^{k-1} constants to be stored are the different possible values of $2^{a(k)}$ since $e^{F\ln 2}=e^{a(k)\cdot\ln 2}\cdot e^{f\ln 2}=2^{a(k)}\cdot e^{f\ln 2}$. It is easy to prove by induction that $|f|\leqq 2^{-k}$ so that the range of the exponent z in $e^z=e^{f\ln 2}$ is:

$$-2^{-k}\cdot\ln 2<f\cdot\ln 2=z<2^{-k}\cdot\ln 2.$$

This reduction to as small a range as we wish (k can be chosen at our convenience) is the most important step since it allows us to obtain any desired accuracy.

2. Study of the relative error $R_m(z)$ in (3)

The maximum of the upper bound (5) for the absolute value of relative error is attained for $z=-2^{-k}\cdot\ln 2$. Denoting this maximum by $M(k,m)$ we computed the following values of this function of two parameters k and m:

Table 1 **Values of M (k, m)**

$m\backslash k$	2	3	4	5
2	2.4×10^{-7}	7.2×10^{-9}	2.2×10^{-10}	6.7×10^{-12}
3	5.1×10^{-11}	3.8×10^{-13}	3.0×10^{-15}	2.3×10^{-17}
4	5.1×10^{-15}	1.1×10^{-17}	2.2×10^{-20}	4.2×10^{-23}
5	4.6×10^{-19}	2.2×10^{-22}	1.0×10^{-25}	5.0×10^{-29}

The number d of correct significant digits in an approximate value of e^N computed by (3) and expressed in decimal numeration depends on the value of $M(k,m)$. If the binary representation of the exponent $z=f\cdot\ln 2$ is considered as exact, the first h significant digits in e^N will be correct, if $M(k,m)$ is less than $\frac{1}{2}\cdot10^{-h}$. But if the value of z is affected by an absolute error dz the condition $M(k,m)<\frac{1}{2}\cdot10^{-h}$ is necessary, but not sufficient. Since $e^{-z}\cdot d(e^z)=dz$, an *absolute* error dz in the exponent z generates an equal relative error in e^z.

Even if the decimal representation N_{10} of N is known to be exact so that there is no error DN_{10} in the given value of N_{10}, the conversion of N_{10} into the binary representation N_2 of N introduces an error $DN_2\neq0$. In a single precision fixed point computation with a 35-bit binary machine, we can have $DN_2=2^{-35}\cdot N$ and, if the double precision is used, $DN_2=2^{-70}\cdot N$. For floating point computations the corresponding conversion errors can reach $2^{-27}\cdot N$ and $2^{-54}\cdot N$.

Let us consider the case when N is large and has q digits in the integral part of its decimal representation N_{10}, so that $10^q>N>10^{q-1}$. In this case, the absolute error dz in $z=f\cdot\ln 2$ can reach 3.10^{q-11} and there will be at most only $10-q$ correct significant digits in the final value of e^N.

Another cause of possible loss of accuracy unrelated to the value of $M(k,m)$ is the generation of an error dz in the multiplication of $\log_2 e$ by N, if N is large. Suppose that the binary value of $\log_2 e$ stored in the memory of the machine has 35 bits, so that the absolute error in $\log_2 e$ is less than 2^{-36}. If the integral part of N_{10} has q digits, then the absolute error Df in $f=N\cdot\log_2 e-M-a(k)$ can reach the value $2^{-36}\cdot10^q\cdot\log_2 e$ so that $dz=Df\cdot\ln 2=2^{-36}\cdot10^q=3.10^{q-11}$ and again, instead of ten, only $10-q$ first significative digits will be correct, if $M(k,m)<5.10^{-11}$. To avoid the loss of accuracy which may be caused by the conversion of N_{10} into N_2 and/or by the multiplication of $\log_2 e$ by N_2, it is advisable, if N is large, to use the *double precision* binary representations of $\log_2 e$ and N_2 in computing f and then continue a single precision computation. This will insure $21-q$ correct digits, if $10^{q-1}<N<10^q$, provided $M(k,m)>\frac{1}{2}10^{q-21}$.

In what follows, we suppose that the double precision is used in computing f, so that the accuracy of the result will depend on $M(k,m)$ only.

3. Number of operations

To be able to choose among combinations (k,m) insuring the same accuracy we have to compare the number of operations and of precomputed constants involved in each of these procedures. Using (4) for $m=2, 3, 4, 5$, forming the corresponding expressions of quotients $P_m(z)/P_m(-z)$ and replacing in them z by its value $z=f\cdot\ln 2=f/g$, where $g=\log_2 e=\ln^{-1}2$, we finally obtained the following practical rules for computing the products $\Pi_m\simeq2^{a(k)}\cdot P_m(z)/P_m(-z)$:

$$\Pi_2=2^{a(k)}+a_2\cdot[f-c_2+b_2\cdot f^{-1}]^{-1} \quad (\dagger)$$

$$\Pi_3=-2^{a(k)}+a_3\cdot[b_3-f-c_3\cdot(f+d_3\cdot f^{-1})^{-1}]^{-1}$$

$$\Pi_4=2^{a(k)}+a_4\cdot[b_4\cdot f^{-1}-c_4+d_4\cdot f+h_4\cdot(f+b_4\cdot f^{-1})^{-1}]^{-1}$$

$$\Pi_5=2\cdot2^{a(k)}\cdot\{\tfrac{1}{2}+f\cdot[b_5-f-c_5(f^2+d_5-h_5[f^2+r_5]^{-1})^{-1}]^{-1}\},$$

where

$a_2=12g\cdot2^{a(k)}$; $b_2=12g^2$; $c_2=6g$; $a_3=24g\cdot2^{a(k)}$;

$b_3=12g$; $c_3=50g^2$; $d_3=10g^2$; $a_4=42g\cdot2^{a(k)}$;

$b_4=42g^2$; $c_4=21g$; $d_4=1.05$; $h_4=102.9g^2$; $b_5=30g$;

$c_5=9240g^3$; $d_5=4116g^2/11$; $h_5=244,\ 944g^4/121$;

$r_5=504g^2/11$.

Since

$$e^N\simeq2^M\cdot2^{a(k)}\cdot P_m(z)/P_m(-z)=2^M\cdot\Pi_m$$

it is seen that the number of multiplications (divisions being counted as multiplications) for $m=2, 3, 4, 5$ is equal to $m+1$ because one more multiplication is needed to find f. The factor 2^M is accounted for by a shift. The number of additions in computing Π_m is also equal to $m+1$. Adding k additions necessary for the determination of f, we have in all $m+k+1$ additions. The precomputed constants to be stored are: 2^{k-1} numbers of the type $2^{a(k)}$, as well as 2^{k-1} numbers a_m for $m=2$, 3 and 4 and 2^{k-1} numbers $2\cdot2^{a(k)}$ for $m=5$. This gives 2^k+m constants necessary to compute Π_m for $2\leqq m\leqq4$, while the computation of Π_5 necessitates $2^{k-1}+m$ constants. Adding to it the constant g, we summarize these results in Table 2.

Table 2 **Values of M, PC, Dg**

	$k=2$			$k=3$			$k=4$			$k=5$		
	M	PC	Dg	M	PC	Dg	M	PC	Dg	M	PC	Dg
$m=2$	3	7	6	3	11	7	3*	19	9	3*	35	10
$m=3$	4*	8	10	4	12	12	4	20	14	4	36	16
$m=4$	5	9	14	5	13	16	5*	21	19	5	37	22
$m=5$	6*	8	18	6*	10	21	6	14	24	6	22	28

* Important Combinations

The number of divisions for $m=2, 3, 4$ and 5 is equal to 2, 3, 3 and 3 respectively. Using the rational approximations it is not possible to eliminate divisions completely, but in some cases it is preferable to reduce their number. A simple algebraic transformation reduces the number of divisions for $m=3, 4, 5$ to one, replacing a division by two

† Suggested by Dr. George E. Collins, IBM.

multiplications. Thus, Π_3, Π_4 and Π_5 can be computed in 4, 5 and 6 multiplications plus one division, using the following equivalent expressions:

$$\Pi_3 = A_3 \cdot \{a^*_3 + f \cdot (b^*_3 + f^2) \cdot [c^*_3 + d^*_3 \cdot f^2 - f(b^*_3 + f^2)]^{-1}\}$$

$$\Pi_4 = A_4 \cdot \{a^*_4 + f(b^*_4 + f^2) \cdot [c^*_4 + (d^*_4 + h^*_4 \cdot f^2) f^2 - f(b^*_4 + f^2)]^{-1}\}$$

$$\Pi_5 = A_5 \cdot \{a^*_5 + f \cdot [k_5 + f^2(b^*_5 + f^2)] \cdot [c^*_5 + (d^*_5 + h^*_5 \cdot f^2) f^2 - f(k_5 + f^2(b^*_5 + f^2))]^{-1}\},$$

where

$A_3 = A_4 = A_5 = 2 \cdot 2^{a(k)}$; $a^*_3 = a^*_4 = a^*_5 = \frac{1}{2}$;

$b^*_3 = 60g^2$, $b^*_4 = 42g^2$, $b^*_5 = 420g^2$; $c^*_3 = 120g^3$, $c^*_4 = 84g^3$,

$c^*_5 = 30{,}240g^5$; $d^*_3 = 12g$, $d^*_4 = 9g$, $d^*_5 = 3360g^3$;

$h^*_4 = (20g)^{-1}$, $h^*_5 = 30g$ and $k_5 = 15{,}120g^4$.

Two single precision subroutines for the computation of e^N, based on (3), were coded a year ago for the IBM 704. They yield ten correct digits, but only eight are needed in the floating point computation. Their characteristics are as follows:

Table 3 **IBM 704 Subroutines for e^N**

	fixed point	*floating point*
Combinations used	$k = 1, m = 4$	$k = 2, m = 3$
Multiplications	4	3
Divisions	2	2
Precomputed constants	5	7
Time (in milliseconds)	2.80	2.63

Six combinations of Table 2 are important for single and double precision, fixed and floating point subroutines for the computation of e^N. They are given in Table 4.

Table 4 **Important Combinations**

Combination	*M*	*PC*	*Dg*
(1) $m = 2, k = 4$	3	19	9
(2) $m = 2, k = 5$	3	35	10
(3) $m = 3, k = 2$	4	8	10
(4) $m = 4, k = 4$	5	21	19
(5) $m = 5, k = 2$	6	8	18
(6) $m = 5, k = 3$	6	10	21

Among them (1), (5) should be used in single precision floating point, double precision floating point and (4) or (6)—in double precision fixed point computations respectively. In single precision fixed point computations the combination (3) needs one more multiplication than (2), but it involves only 8 stored constants while (3) has 35 constants.

Part B—Decimal machine; polynomial approximations

4. Reduction to small range

Multiplying N in e^N by $g = \log_{10} e$, we have $Ng = N \cdot \log_{10} e = M^* + F = M^* + a(k) + f$, where M^* is the integral and F the fractional part of the product. Dividing the range (0;1) of F into 2^k subintervals exactly as in Part A and using the same notations $a(k)$ and f, we have

$$e^N = 10^{M^*} \cdot 10^{a(k)} \cdot e^{f \ln 10}$$

with $|f| \leqq 2^{-k}$. Here the 2^{k-1} precomputed constants are equal to $10^{a(k)}$, where $a(k)$ takes the same 2^{k-1} values $(2j-1)/2^k$, $1 \leqq j \leqq 2^{k-1}$, as in Part A.

5. Rational approximations

Without repeating the analysis of Part A, we give (in Table 5) its results, which can be useful only in the case when division is not too slow an operation for a given decimal machine:

Table 5 **Values of M(k, m)**

$m \backslash k$	2	3	4	5
2	1.9×10^{-4}	4.0×10^{-6}	1.0×10^{-7}	3.0×10^{-9}
3	4.5×10^{-7}	2.4×10^{-9}	1.5×10^{-11}	1.1×10^{-13}
4	5.9×10^{-10}	7.8×10^{-13}	1.3×10^{-15}	2.2×10^{-18}
5	4.9×10^{-13}	1.6×10^{-16}	6.6×10^{-20}	3.0×10^{-23}
6	2.9×10^{-16}	2.4×10^{-20}	2.4×10^{-24}	2.7×10^{-28}

In the case $m = 6$, e^z is approximated by $P_6(z)/P_6(-z)$, where the coefficients c_k of

$$P_6(z) = 1 + z/2 + \sum_{k=2}^{6} c_k \cdot z^k$$

are: $c_2 = 5/44$, $c_3 = 1/66$, $c_4 = c_3/12$, $c_5 = c_4/20$ and $c_6 = c_5/42$. The quotient $P_6(z)/P_6(-z)$ is computable in one multiplication (necessary to form z^2) and four divisions:

$$P_6(z)/P_6(-z) = 1 + z \cdot [1 - z/2 + z^2/Q(z^2)]^{-1},$$

where

$$Q(t) = 84 - a_6 \cdot \{t + b_6 + c_6 \cdot (t + d_6)^{-1}\}^{-1}$$

with $a_6 = 43{,}344$, $b_6 = 12294/43$, $c_6 = 53{,}824{,}320/43^2$, $d_6 = 3960/43$.

Thus, for a decimal machine allowing the use of division the rational approximations yield the same results as for a binary machine, but the number of operations and of precomputed constants, for the same accuracy, is somewhat greater. We cite some of them in Table 6.

Table 6 **Values of M, PC, Dg**

m	*k*	*Mult.*	*Add.*	*Prec. Const.*	*Correct Digits*
2	5	3	8	35	8
3	3	4	7	12	8
3	4	4	8	20	10
5	4	6	10	14	18
6	3	7	10	12	19
5	5	6	11	22	22

The rational approximations should not be used for decimal machines in which the division is too slow. For such machines polynomial approximations deduced from the series (I) are much more economical. They are more

Table 7 **Values of $M^*(n,k)$**

$k\backslash n$	5	6	7	8	9	10
2	1×10^{-4}	7×10^{-6}	3×10^{-7}	1×10^{-8}	3×10^{-10}	$*1\times10^{-11}$
3	2×10^{-6}	5×10^{-8}	1×10^{-9}	$*2\times10^{-11}$	3×10^{-13}	5×10^{-15}
4	5×10^{-8}	6×10^{-10}	$*6\times10^{-12}$	5×10^{-14}	4×10^{-16}	3×10^{-18}
5	2×10^{-9}	$*7\times10^{-12}$	4×10^{-14}	2×10^{-16}	7×10^{-19}	3×10^{-21}

economical than the partial sums of exponential power series, even when the latter are shortened by relaxation of last terms.

6. *Polynomial approximations*

We study now the relative error $R_n(z)=e^{-z}\cdot E_n(a,x)$, where $E_n(a,x)$ verifies the inequality (2), $z=ax$, $a=2^{-k}\cdot\ln10$ and $|x|\leqq1$. Therefore

$|R_n(z)|\leqq2.10^a\cdot I_n(a)[1-a(n+1)^{-1}/2]^{-1}=M^*(n,k)$

Denoting the upper bound of $|R_n(z)|$ by $M^*(n,k)$, we have the result given in Table 7.

It is interesting to compare $M^*(n,k)$ to the relative error made in approximating e^z by the partial sum $\sum_{j=0}^{n-1} z^j/j!$

This relative error is essentially equal to $e^{-z}\cdot z^n/n!\cdot(1-z/(n+1))^{-1}$, the maximum value of which (obtained for $z=-\ln10/2^k$) we denote by $B(n,k)$:

$$B(n,k)=10^{2^{-k}}\cdot(\ln10/2^k)^n(n!)^{-1}/[1-z(n+1)]$$

Thus, we have approximately $B(n,k)=M^*(n,k)\cdot2^{n-1}$ which shows that the relative error made using the Maclaurin expansion of e^z is 2^{n-1} times greater than the relative error of our polynomial approximation. For $n=6$, 8, 10 the factor 2^{n-1} takes the values 32, 128, 512.

Different combinations (n,k) insure the same number of correct digits in the final result. Thus, for instance ten correct digits can be obtained using either one of four combinations (6,5), (7,4), (8,3) and (10,2).

The coefficients c_k of the polynomial approximation $Q_{n-1}(x)$

$$e^{ax}\approx Q_{n-1}(x)=I_0(a)+2\cdot\sum_{m=1}^{n-1}I_m(a)\cdot T_m(x)=\sum_{j=0}^{n-1}c_j\cdot x^j$$

are obtained, replacing the Tchebychev polynomial $T_m(x)$ by its expression

$$2T_n(x)=n\cdot\sum_{m=0}^{2m\leq n}(-1)^m\cdot\binom{n-m}{m}\cdot(2x)^{n-2m}/(n-m).$$

The final result is

$$e^z\approx Q_{n-1}(z)=\sum_{j=0}^{n-1}(1-b_{nj})z^j/j!$$

where b_{nj} are small, so that our approximating polynomial $Q_{n-1}(x)$ appears as the sum of the first n terms of the exponential series for e^{ax} with slightly modified coefficients.

We now illustrate the computation of coefficients and the relaxation of $Q_{n-1}(x)$ on a particular example of approximation giving ten correct significant digits. The combination used in this example is $n=6$ and $k=5$, so that $a=(\ln10)/32$. Replacing in the expression for $Q_5(x)$ the polynomials $T_m(x)$ by their expansions and omitting the argument a of the Bessel functions $I_m(a)$, we have

$$Q_5(x)=I_0+2I_1\cdot x+2I_2\cdot(2x^2-1)+2I_3\cdot(4x^3-3x)+2I_4\cdot(8x^4-8x^2+1)+2I_5\cdot(16x^5-20x^3+5x).$$

Thus,

$$e^z\approx Q_5(x)=\sum_{j=0}^{5}c_j(2x)^j,$$

where

$$c_0=I_0-2I_2+2I_4;\ c_1=I_1-3I_3+5I_5;\ c_2=I_2-4I_4;$$
$$c_3=I_3-5I_5;\ c_4=I_4;\ c_5=I_5.$$

7. *Relaxation*

The last term $2^5\cdot c_5\cdot x^5$ can be replaced by an approximating polynomial of the fourth degree, the error made being less than 10^{-11}. In fact we will have two approximating polynomials: one in $0\leqq x\leqq1$ and the other in $-1\leqq x\leqq0$. The modified Tchebychev polynomial

$$T_5(x)=512x^5-1280x^4+1120x^3-400x^2+50x-1\quad(0\leqq x\leqq1)$$

is less in absolute value than one, if $0\leqq x\leqq1$. Changing the sign of its argument, we have another polynomial

$$T^*_5(x)=-(512x^5+1280x^4+1120x^3+400x^2+50x+1)\quad(-1\leqq x\leqq0)$$

which is less than one in absolute value, if $-1\leqq x\leqq0$. Therefore, in $0\leqq x\leqq1$ we take

$$x^5=5x^4/2-35x^3/16+25x^2/32-25x/256+1/512+g(x),\quad(0\leqq x\leqq1)$$

where $|g(x)|\leqq2^{-9}$ for $0\leqq x\leqq1$, while in $-1\leqq x\leqq0$ we have

$$x^5=-5x^4/2-35x^3/16-25x^2/32-25x/256-1/512+h(x),\quad(-1\leqq x\leqq0)$$

where again $|h(x)|\leqq2^{-9}$ for $-1\leqq x\leqq0$. Now, since $a=(\ln10)/32$, we have $2^5\cdot c_5=2^5\cdot I_5(a)\leqq10^{-7.79125}$ and therefore $2^5\cdot c_5\cdot g|(x)|$ and $2^5\cdot c_5\cdot|h(x)|$ are less than $10^{-10.50052}$, which proves that the absolute error made in dropping $g(x)$ or $h(x)$ is less in absolute value than 3.2×10^{-11}. The corresponding relative error is less in absolute value than 3.44×10^{-11}. Thus, replacing x^5 by a polynomial of the fourth degree we obtain

$$2^5\cdot c_5x^5=I_5\cdot[\pm5\cdot(2x)^4-35\cdot(2x)^3/4\pm25\cdot(2x)^2/4-25\cdot(2x)/16\pm1/16],$$

so that the relaxed coefficients c^*_k become

$$c^*_0=I_0-2I_2+2I_4\pm I_5/16;\ c^*_1=I_1-3I_3+55I_5/16;$$
$$c^*_2=I_2-4I_4\pm25I_5/4;\ c^*_3=I_3-55\cdot I_5/4;\ c^*_4=I_4\pm5I_5.$$

In c^*_0, c^*_2 and c^*_4 the plus sign should be taken, if f is positive, and the minus sign if f is negative. The relaxation of the term in x^4 does not work, the resulting error being of the order of 10^{-6}.

Thus, e^z can be computed with the aid of two approximating polynomials of the fourth degree in five multiplications:

$$e^z \simeq d_0 + f\cdot[d_1 + f\cdot(d_2 + f\cdot[d_3 + d_4\cdot f])]$$

where

$$d_j = 2^{(k+1)j}\cdot c^*{}_j\cdot = 64^j\cdot c^*{}_j$$

Adding to the eight coefficients d_j, the $2^4+1=17$ precomputed constants necessary for the reduction of the infinite range $-\infty < N < \infty$ to $|f| \leqq 2^{-5}$, we obtain, in all, 26 precomputed constants. The number of operations is: nine additions and five multiplications.

This example, it is hoped, shows clearly the procedure which is to be followed in the computation of the coefficients d_j of a relaxed approximating polynomial $Q_{n-1}(z)$, when n and k have fixed known values.

The number of precomputed constants is equal to $2^{k-1} + n + [n/2] + 1$, $[n/2]$ denoting the integral part of n/2. The number of multiplications and additions is equal to $n-1$ and $n+3$.

8. *Briggs' Method*

To conclude we describe a rather curious adaptation of an old method, used by Briggs in 1624 for compiling his table of logarithms, for the computation of e^N with the aid of an electronic computer. To fix our ideas let us consider only the case where first ten significant correct digits are required, the computer being a binary machine.

To compute the factor 2^F in $e^N = 2^M\cdot 2^F$, $0 < F < 1$, with ten correct digits we will use 17 precomputed and stored constants $c_k = \log_2(1+2^{-k})$, $1 \leqq k \leqq 17$. Let $f_0 = F$ and define recursively f_i, a_i as follows: $f_{i+1} = f_i$ and $a_{i+1} = 0$, if $f_i < c_{i+1}$, but $f_{i+1} = f_i - c_{i+1}$ and $a_{i+1} = 1$, if $f_i > c_{i+1}$. Then

$$F = \sum_{k=1}^{17} a_k\cdot c_k + f^*,$$

where $0 < f^* = f_{17} < c_{17} = 1.1\times 10^{-5}$, so that $\frac{1}{2}(f^*\cdot \ln 2)^2 < 3.10^{-11}$ and $2^{f^*} \simeq 1 + f^*\cdot \ln 2$. Therefore

$$2^F = (1+f^*\cdot \ln 2)\cdot \prod_{k=1}^{17} (1+2^{-k})^{a_k} \qquad (a_k = 0 \text{ or } 1)$$

The multiplications by factors $1+2^{-k}$ are in fact performed as additions:

$$A\cdot(1+2^{-k}) = A + 2^{-k}\cdot A$$

In all only two multiplications are used: one to form F, another to compute the product $f^*\cdot \ln 2$. The number of additions depends on N. It varies between 18 and 35. There are 19 precomputed constants: ln2, $\log_2 e$ and seventeen c_k.

If first eight correct digits are required (floating point computation), then $f^* = f_{14}$. Six correct digits are obtained, if $f^* = f_{12}$.

References

1. G. N. Watson, *Theory of Bessel Functions*, 1952; p. 15; form. (3) for $z = ia$, $it = e^{i\theta}$, $\cos\theta = x$.
2. O. Perron, *Die Lehre von den Kettenbruchen*, pp. 435, 436; Chelsea Publ. Co., New York, 1950.
3. *Jour de Math.* (3), 1876, vol. II, p. 271. See also: *Modern Analysis* by Whittaker and Watson, Ch. VII, p. 125.
4. *Thesis. Annales de l'Ecole Norm.* (3), 1892, vol. 9, pp. 1–93. See also: Ch. XVIII, *Analytic Theory of Continued Fractions* by H. S. Wall, 1948, Van Nostrand.
5. *Amer. Mathem. Monthly*, April 1949, V. 56, p. 243.
6. *Applied Analysis*, Prentice Hall, 1956, Ch. VI, § 19, pp. 419–427.
7. *Continued Fractions*, H. S. Wall, p. 349, formula (91.6).

Received December 12, 1956.