E. G. Kogbetliantz

Computation of Sin N, Cos N and $\sqrt[m]{N}$ Using an Electronic Computer

Abstract: Rational Padé approximations to Sin N in the interval $0 \leqslant N \leqslant 41\pi/256$ and to Cos N in $0 \leqslant N \leqslant 87\pi/256$ allow the computation of both functions in $0 \leqslant N \leqslant \pi/2$ with the first ten correct significant digits in four multiplications and divisions only. If the infinite range $0 \leqslant N < \infty$ is considered, one more multiplication reduces it to the range $0 \leqslant N \leqslant \pi/2$ so that the total number of operations is five. The method is flexible and gives any desired accuracy. Thus if eighteen first correct significant digits are required, they are obtained in seven operations for any N in $(0, \infty)$.

The same method applied to $\sqrt{N}$ and $\sqrt[3]{N}$ yields a very accurate first guess which then is improved by Newton's method. For the radicals $\sqrt[m]{N}$ with $m > 4$, Newton's method is too slow, and rational Padé approximations studied in this paper yield better subroutines.

Part I: Subroutines for Sine and Cosine

• *Padé approximations*

Given the formal expansion

$$f(x) \sim \sum_{n=0}^{\infty} c_n x^n$$

of $f(x)$ into a convergent or divergent power series, a Padé approximation to $f(x)$ is rational function $P_M(x)/Q_N(x)$, where

$$P_M(x) = c_0 + \sum_{k=1}^{M} p_k x^k \, ; \, Q_N(x) = 1 + \sum_{k=1}^{N} q_k x^k$$

are defined by the identity

$$Q_N(x) \sum_{n=0}^{\infty} c_n x^n - P_M(x) \equiv x^{M+N+1} \sum_{n=0}^{\infty} A_n x^n \, . \qquad (1)$$

Therefore q_k, $1 \leqslant k \leqslant N$, satisfy the system

$$\sum_{m=1}^{N} q_m c_{N-m+j} = -c_{N+j} \qquad (1 \leqslant j \leqslant N) \, , \qquad (2)$$

while p_k, $1 \leqslant k \leqslant M$, are computed with the aid of

$$p_k = c_k + \sum_{m=1}^{k} q_m c_{k-m} \qquad (1 \leqslant k \leqslant M) \, . \qquad (3)$$

We have also

$$A_n = \sum_{m=0}^{N} q_m c_{2N-m+n+1} \qquad (n \geqslant 0) \, . \qquad (4)$$

The coefficient A_n decreases very rapidly and the first term $A_0 x^{M+N+1}$ in the righthand member of (1) divided by $Q_N(x)$ is a good estimate of the absolute error made in approximating $f(x)$ by $P_M(x)/Q_N(x)$. Such approximations are useful in general for small range of $|x|$ since the error increases as a power of $|x|$, if $|x| > 1$. For $|x| < 1$ the value of $Q_N(x)$ differs very little from one since q_k decreases rapidly when k increases, and in general q_1 is already small enough. This shows that the order of magnitude of the upper bound B of the absolute error in a given range $(0, x_0)$, where $x_0 < 1$, is approximately represented by $|A_0||x_0|^{M+N+1}$. Thus, it is sufficient to compute A_0 to estimate B.

Eliminating the coefficients q_m from the N equations (2) and the expression (4) of A_0, we have $A_0 = D_N/d_N$, where the elements of the determinant $D(d_{ij})$ are defined by

$$d_{ij} = c_{i+j-1} \qquad (1 \leqslant i, j \leqslant N+1) \, ,$$

while the determinant d_N is the principal minor of D_N obtained by omitting in D_N the last row and the last column.

The accuracy of approximation P_M/Q_N for a fixed N depends on M, and it is known that for a fixed value of N the choice $M = N$ (diagonal of the Padé Table) gives the best approximation. The accuracy of the approximation P_N/Q_N increases rapidly with N. Once N is fixed, the rational function P_N/Q_N should be expanded into an

equivalent continued fraction, since in this form the computation of the approximation is performed much more rapidly than in the form P_N/Q_N, the number of operations (multiplications and divisions) being halved.

• *Reduction of the infinite range*

The values of Cos N and Sin N, where $0<N<\infty$, are computed reducing the argument N to a small range as follows. Forming the product $N\pi^{-1}$ and denoting the fractional part of it by f_0, so that $N\pi^{-1}=n+f_0$ where n is an integer, we have Cos $N=(-1)^n$Cos πf_0, Sin $N=(-1)^n$Sin πf_0, with $0<f_0<1$. Subtracting f_0 from $\frac{1}{2}$ and defining $S_1=\text{sign}(\frac{1}{2}-f_0)$, so that $f_0<\frac{1}{2}$ if $S_1=+1$, and $f_0>\frac{1}{2}$ if $S_1=-1$, we form $f_1=\frac{1}{2}-(\frac{1}{2}-f_0)S_1$, that is, $f_1=f_0$ if $f_0<\frac{1}{2}$, but $f_1=1-f_0$ if $f_0>\frac{1}{2}$, we have Sin $\pi f_0=$ Sin πf_1 and Cos $\pi f_0=S_1$Cos πf_1, where $0<f_1<\frac{1}{2}$. Therefore the subroutine computes Sint and Cost, where $0<t<\pi/2$. Now Sin πf_1 is approximated by P_2/Q_2 if f_1 belongs to the interval $(0, 41/256)$. When f_1 exceeds $41/256$, then Sin πf_1 is computed as a cosine, namely as Cos $[\pi(\frac{1}{2}-f_1)]$, where $\frac{1}{2}-f_1$ belongs to the interval $(0, 87/256)$.

Cos πf_1 is computed, using the approximation P_3/Q_3, if $0<f_1\leqslant 87/256$, but if f_1 exceeds $87/256$ it is computed as Sin $[\pi(\frac{1}{2}-f_1)]$, where $\frac{1}{2}-f_1$ belongs to the interval $(0, 41/256)$. The fractions $41/256$ and $87/256$ are represented in the binary numeration by $(0.0010\ 1001)_2$ and $(0.\ 0101\ 0111)_2$. The angle $87\pi/256$ expressed in degrees is equal to $61°52'30''$.

• *Subroutine for* Cos N, $0<N\leqslant 87\pi/256$

Applying this method to Cos z and letting $z^2=x$, we have $c_n=(-1)^n/(2n)!$ and choose $M=N=3$. To estimate the accuracy of the approximation P_3/Q_3 we compute $A_0=D_3/d_3$. In this case we find easily that $10!8!d_3=59$ and $8!11!15!D_3=11{,}367.25$, so that

$$A_0=D_3/d_3\approx 1.33\times 10^{-11}\ .$$

If the first ten correct significant digits are required, we should choose the range $(0, x_0)$ so that the upper bound B for the *absolute* error is at most equal to 5.10^{-11}. The relative error is not important here because Cos $z>0.1$ for $z<1.46$. Moreover, for cosine all the coefficients q_m are positive and $Q_3(x)>1$, which allows us to take $10^{11}B=10^{11}A_0x_0^7=4z_0^{14}/3$. The condition $B\leqslant 5.10^{-11}$ now yields $z_0\leqslant 0.3408\pi$, so that $87\pi/256=0.3398\ 4375\pi$ does insure first ten correct significant digits, using $P_3(x)/Q_3(x)$ as approximation to Cos $z=$Cos$(x^{\frac{1}{2}})$. But if we use $P_3(x)/Q_3(x)$ as it is without transforming it into a continued fraction, we will need seven multiplications and one division to compute the cosine while the usual polynomial approximation

$$S_{12}(x)=\sum_{k=0}^{6}(-1)^k x^{2k}/(2k)!$$

yields in $0<x<\pi/4$ the same accuracy in seven multiplications. The use of a Padé approximation can be justified only if we can economize the machine time without losing the accuracy, and this is done by transforming P_3/Q_3 into a continued fraction, namely

$$P_3(x)/Q_3(x)\equiv C_0+\sum_{m=1}^{3}\frac{C_m^*|}{|x+D_m^*}\ . \tag{5}$$

The righthand member of (5) can be computed in three divisions and one multiplication only, but here we are confronted with the following complication, which is of course a common feature of almost all applications of this method: the constant C_0 is equal to $-14{,}615/127$, so that our approximation is a small difference of two large numbers. In a single precision computation the continued fraction (5), when used in a 35-bit binary computer, will give only ten decimal digits, the first three of which will be lost in subtracting $14{,}615/217$, so that the final value of a cosine will have only first seven correct decimals instead of ten.

Fortunately it is possible to eliminate this cause of loss of accuracy, reducing the value of C_0 below one in absolute value and making it positive. This can be done without increasing the number of divisions and multiplications which are costly in machine time. Adding one more term to $P_3(x)$, we consider $P_4(x)/Q_3(x)$, where

$$P_4(x)=1+p_1x+p_2x^2+p_3x^3-tq_3x^4\ .$$

The numerical value of the constant t will be chosen later and in such a way that the value taken by C_0 will satisfy the condition $0<C_0<1$. The replacement of $P_3(x)$ by $P_4(x)$ changes the first equation of the system (2) which now becomes

$$c_4+c_3q_1+c_2q_2+(c_1+t)q_3=0\ .$$

Solving the system (2), we have

$$\begin{aligned}12!(59-34t)q_1&=10!(229-76t)\\ 12!(59-34t)q_2&=\ 8!(297-42t)\\ 12!(59-34t)q_3&=\ 6!127.\end{aligned} \tag{6}$$

Both A_0 and $C_0=(p_3+tq_2)/q_3$ also become functions of t:

$$\begin{aligned}127C_0&=-2{,}352t^2+33{,}264t-14{,}615\\ 44.15!A_0&=(45{,}469+9{,}336t)/(59-34t)\ ,\end{aligned}$$

while

$$\begin{aligned}12!(59-34t)p_1&=10!(-3{,}665+2{,}168t)\\ 12!(59-34t)p_2&=\ 8!(-2{,}133+4{,}456t)\\ 12!(59-34t)p_3&=\ 6!(-14{,}615+16{,}632t)\ .\end{aligned} \tag{7}$$

The reduction to the small range leads to the computation of Cos πf, where the number f is known. Therefore the rational approximation to Cos πf is to be expressed in terms of f. Dividing $P_4(\pi^2f^2)$ by $Q_3(\pi^2f^2)$, we obtain

$$\text{Cos}\ \pi f\approx C_0-\pi^2tf^2+\sum_{m=1}^{3}\frac{C_m|}{|f^2+D_m}\ .$$

Here the continued fraction is the expansion of the quotient

$$(u_0+u_1f^2+u_2f^4)\left[1+\sum_{m=1}^{3}q_m(\pi f)^{2m}\right]^{-1}\ , \tag{8}$$

so that

$$q_3u_0=q_3-(p_3+tq_2)$$
$$\pi^{-2}q_3u_1=q_3(p_1+t)-q_1(p_3+tq_2)$$
$$\pi^{-4}q_3u_2=q_3(p_2+tq_1)-q_2(p_3+tq_2)\,. \qquad (9)$$

We now take $\pi^2t=4.5=(100.1)_2$, choosing the factor π^2t in such a way that $C_0=0.49304\ 82724$ while $A_0=1.987\times10^{-11}<2.10^{-11}$. Computing $B=A_0(87\pi/256)^{14}$ we find $B=4.10^{-11}$ which insures ten correct significant digits, if $0<f\leqslant87/256$. At the same time the factor $4.5=(100.1)_2$ allows so rapid a computation of the term $4.5f^2$ that we do not count it as an operation.

The coefficients C_m, D_m in the final result

$$\text{Cos}\ \pi f\approx0.49304\ 82724-4.5f^2+\sum_{m=1}^{3}\frac{C_m|}{|f^2+D_m} \qquad (10)$$

are computed as follows. First (6) and (7) give the values of p_k, q_k using the value $4.5/\pi^2$ of t. Then u_0, u_1, u_2 are obtained by (9) and the products $v_k=\pi^{2k}q_k$ are computed. The last operation consists in the transformation

$$\left(\sum_{k=0}^{2}u_kf^{2k}\right)\Big/\left(\sum_{k=0}^{3}v_kf^{2k}\right)\equiv\sum_{m=1}^{3}\frac{C_m|}{|f^2+D_m}\,. \qquad (11)$$

Denoting the combinations of C_m and D_m as follows:

$$D_1D_2D_3+C_2D_3+C_3D_1=W_1\,,$$
$$D_1D_2+D_2D_3+D_3D_1+C_2+C_3=W_2\,,$$
$$D_1+D_2+D_3=W_3\,,$$
$$C_1(D_2D_3+C_3)=W_4\,,$$
$$C_1(D_1+D_2)=W_5\,,$$

and letting $C_1=W_6$, we see that the identity (11) yields a system of six linear equations for our six unknowns W_j, $1\leqslant j\leqslant6$. Solving this system, we now have

$$D_1=W_2-W_5/W_6\,;\ C_2=W_2-W_4/W_6-D_1(W_3-D_1)\,;$$
$$D_3=(W_1-W_4/W_6)/C_2\,;\ D_2=W_3-D_1-D_3\,;$$
$$C_3=W_4/W_6-D_2D_3\,.$$

• *Subroutine for* Sin N, $0<N\leqslant41\pi/256$.

Trying to approximate $x^{-1}\,\text{Sin}\,x$ by the rational function $P_2(x^2)/Q_2(x^2)$, we find again that the constant term C_0 in

$$\text{Sin}\ \pi f\approx f\left(C_0+\sum_{k=1}^{2}\frac{C_k|}{|f^2+D_k}\right) \qquad (10)$$

is a large number, $C_0\approx23$. Therefore, we approximate $x^{-1}\text{Sin}x$ as follows:

$$x^{-1}\text{Sin}x\approx(1+a_1x^2+a_2x^4-tb_2x^6)/(1+b_1x^2+b_2x^4)$$

and determine the unknown coefficients in the usual way. Thus, denoting the coefficient of x^{2k} in the Maclaurin series of $x^{-1}\text{Sin}x$ by $c_k=(-1)^k/(2k+1)!$ and solving the system

$$(t+c_1)b_2+c_2b_1+c_3=0$$
$$c_2b_2+c_3b_1+c_4=0\,,$$

we obtain

$$36(11+60t)b_1=13+30t$$
$$1008(11+60t)b_2=\ 5\,. \qquad (12)$$

Since $a_1=b_1+c_1$ and $a_2=b_2+c_1b_1+c_2$, we have also

$$36(11+60t)a_1=-(53+330t)$$
$$1008(11+60t)a_2=(551+5460t)/15 \qquad (13)$$

as well as

$$11!(11+60t)A_0=(11{,}528-31{,}920t)/1008\,.$$

Finally, replacing x by πf,

$$\text{Sin}\ \pi f\approx f\left[C_0(t)-t\pi^3f^2+\sum_{k=1}^{2}\frac{C_k^*|}{|f^2+D_k^*}\right],$$

where

$$75C_0(t)=\pi(12{,}600t^2+10{,}920t+551)\,,$$

since

$$C_0(t)=(a_2+tb_1)\pi/b_2\,.$$

Here we take $t\pi^3=-51/32=-(1.11011)_2$, so that in a binary machine the term $t\pi^3f^2$ is computed very rapidly. The corresponding value $t_0=-51/32\pi^3$ of t is equal approximately to $t_0=-0.05140\ 08830$, and it gives to $C_0(t)$ the value $0.96309\ 49114=C_0(t_0)$.

Computing the logarithm of A_0, we find $\log A_0=-7.38358$. The upper bound of the relative error is equal to $A_0(\pi f^*)^{11}/\text{Sin}\ \pi f^*$, where $f^*=41/256$. Therefore its logarithm is equal to $-10{,}34813$ which proves that first ten significant digits of our approximation are correct for this choice of t:

$$\text{Sin}\ \pi f=f\left[0.96309\ 49114+51f^2/32+\sum_{k=1}^{2}\frac{C_k^*|}{|f^2+D_k^*}\right]$$
$$+E10^{-11}, \qquad (14)$$

where $|E|<4.5\ \text{Sin}\ \pi f$, provided $f\leqslant f^*=41/256$.

The four coefficients C_k^*, D_k^* are computed as follows. First b_1, b_2, a_1, a_2, are found, using $t=t_0$. Then the numbers $u_1=(a_1+t)\pi^3$, $u_2=(a_2+tb_1)\pi^5=b_2\pi^4C_0(t_0)$, $v_1=b_1\pi^2$ and $v_2=b_2\pi^4$ are computed. The continued fraction in (14) is an expansion of a rational function, namely

$$(\pi+u_1f^2+u_2f^4)/(1+v_1f^2+v_2f^4)-u_2/v_2\equiv\sum_{k=1}^{2}\frac{C_k^*|}{|f^2+D_k^*}\,.$$

Therefore C_k^*, D_k^* are obtained from the identity

$$C_1^*(f^2+D_2^*)(1+v_1f^2+v_2f^4)\equiv(A+Bf^2)\,[C_2^*+D_1^*D_2^*+(D_1^*+D_2^*)f^2+f^4]\,,$$

where $A=\pi-u_2/v_2$ and $B=u_1-v_1u_2/v_2$. First $C_1^*=B/v_2$ is computed and then the three unknowns D_2^*, $D_1^*+D_2^*=n$ and $D_1^*D_2+C_2^*=m$ are obtained solving the system

$$An+Bm-C_1^*v_1D_2^*=C_1^*$$
$$Am-C_1^*D_2^*=0$$
$$Bn-C_1^*v_2D_2^*=C_1^*v_1-A\,.$$

Finally,

$$D_1^*=n-D_2^*\ \text{and}\ C_2^*=m-D_1^*D_2^*\,.$$

Having described how to program a subroutine for Sin N and Cos N yielding the first ten correct significant digits in *five* operations, we add that the same method

gives a subroutine which yields the first eighteen correct significant digits in *seven* operations only. To obtain this accuracy, Padé approximations of order $M=N=5$ for cosine and of order $M=N=4$ for sine are to be used in the same intervals $0<f\leqslant 87/256$ and $0<f\leqslant 41/256$.

The final formulae are:

$$\operatorname{Cos} \pi f \approx C_0-\alpha f^2+\sum_{K=1}^{5}\frac{C_k|}{|f^2+D_k}$$

and

$$\operatorname{Sin} \pi f \approx f\left[E_0+\beta f^2+\sum_{K=1}^{4}\frac{E_k|}{|f^2+G_k}\right],$$

the factors α and β being chosen in such a way that C_0 and E_0 are small constants and the multiplications by α and β are very rapid operations.

If subroutines for the computation of Sin z and Cos z are required in the interval $|z|<\pi/2$ only, then the approximations (10) and (14) are not the most economical because they are expressed in terms of $f=z/\pi$ and they necessitate the division of the given argument z by π. To economize this operation, the same approximations can be expressed in terms of z directly. Approximating Cos z by $P_4(z^2)/Q_3(z^2)$ and choosing in $P_4(z^2)$ $t=t^*=117/256$ $=(0.0111\ 0101)_2$, we have

$$\operatorname{Cos} z \approx C_0^*-t^*z^2+\sum_{m=1}^{3}\frac{C_m|}{|z^2+D_m}, \qquad (|z|\leqslant 87\pi/256) \tag{10*}$$

where $C_0^*=0.75911\ 39425\ 44$. The correct value of C_0^* and the coefficient C_m, D_m are computed, transforming $P_4(z^2)/Q_3(z^2)$ to the form (10*) with the aid of (6) and (7), where $t=117/256$.

The same holds for (14) which becomes

$$\operatorname{Sin} z \approx z\left[C_0^{**}-t^{**}z^2+\sum_{m=1}^{2}\frac{C_m|}{|z^2+D_m}\right] \qquad (|z|\leqslant 41\pi/256) \tag{14*}$$

with $C_0^{**}=0.32342\ 18088\ 78$ and $t^{**}=-105/2048=-(0.00001\ 10100\ 1)_2$.

Part II: Subroutines for radicals

$\sqrt{N}$ is generated usually by successive approximations. A first guess x_0 is improved with the aid of Heron's method

$$x_{n+1}=\tfrac{1}{2}(x_n+Nx_n^{-1}) \tag{15}$$

and the sequence $x_0, x_1, x_2 \ldots$ converges to $\sqrt{N}$, the relative error $e_n=x_nN^{-\frac{1}{2}}-1$ decreasing quadratically: $|e_{n+1}|\approx\tfrac{1}{2}e_n^2$. The number of iterations depends on the accuracy required and on the choice of x_0. This is why it is important to begin with as small e_0 as possible. A procedure for the computation of a good initial guess x_0 developed in the sequel is based on Padé approximations.

Heron's method (15), two milennia old, is a particular case of Newton's method for solving any equation $f(x)=0$ by successive approximations

$$x_{n+1}=x_n-f(x_n)/f'(x_n). \tag{16}$$

If $f(x)\equiv x^m-N=0$, then

$$\sqrt[m]{N}\approx x_{n+1}=[(m-1)x_n+Nx_n^{1-m}]/m, \tag{17}$$

which proves that this method becomes less and less economical, when the order m of the radical $N^{1/m}$ increases. For a cubic root ($m=3$) three operations are necessary to compute x_{n+1} knowing x_n:

$$N^{1/3}\approx x_{n+1}=\tfrac{1}{2}x_n+(N+\tfrac{1}{2}N)/(2x_n^2+Nx_n^{-1}), \tag{18}$$

while for $N^{1/5}$ five operations are needed.

Therefore, rational Padé approximations to roots of orders $m\geqslant 3$ are even more important than for square roots.

• *Square root—binary machine*

Letting $N=2^{2m}f$, where $0.25<f\leqslant 1$ and m is an integer, we have $\sqrt{N}=2^m\sqrt{f}$. Let us consider the general case $f^{1/n}$, $n\geqslant 2$, $2^{-n}<f\leqslant 1$. The accuracy of a rational approximation $P_s(f)/Q_s(f)$ to $f^{1/n}$ depends on the degree s of polynomials P_s, Q_s and on the range $a\leqslant f\leqslant b$ of f. It increases with s, but only small values of s are of interest for us. Keeping s small we can increase the accuracy by decreasing the ratio $r=b/a$ which implies a subdivision of the whole range $(2^{-n}, 1)$ into subintervals.

Given a subinterval (a, b), we choose an interior point c, $\frac{1}{2}b<c<b$, and introduce a new variable t, letting $f=c(1+t)$. If the range of t is denoted by $(-t_1, t_2)$, we have

$$rt_1+t_2=r-1. \qquad (0<t_1, t_2<1) \tag{19}$$

A second equation for t_1, t_2 is obtained by equating the absolute values of the minimum and maximum of relative error which happen to correspond to $t=-t_1$ and $t=t_2$, the relative error being an increasing function of t. Given s, the coefficient p_{si}, q_{si} of P_s and Q_s are determined by the identity

$$\left[\sum_{j=0}^{\infty}\binom{1/n}{j}t^j\right]\left(\sum_{i=0}^{s}q_{si}t^i\right)\equiv\sum_{i=0}^{s}p_{si}t^i+t^{2s+1}\sum_{0}^{\infty}A_{si}t^i. \tag{20}$$

Denoting the binomial coefficient $\binom{1/n}{j}$ by c_j, we have $p_{s0}=q_{s0}=1$ and

$$\sum_{i=0}^{s}c_{s+k-i}q_{si}=0 \qquad (1\leqslant k\leqslant s)$$

$$p_{sk}=\sum_{i=0}^{k}c_{s-i}q_{si} \qquad (1\leqslant k\leqslant s)$$

$$A_{s0}=\sum_{i=0}^{s}c_{2s+1-i}q_{si}.$$

Once P_s and Q_s are determined, we replace t by $f/c-1$ and approximate $f^{1/n}$ by

$$f^{1/n}\approx c^{1/n}\left(1+\sum_{i=1}^{s}B_{si}f^i\right)\left(1+\sum_{i=1}^{s}C_{si}f^i\right)^{-1}.$$

Finally, transforming this rational function of f into a continued fraction, it is seen that the number of operations needed to compute our approximation is equal to s.

The sum of infinite series in the second member of (20) is fairly well represented by $A_{s0}(1+t)^{-1}$ since the ratio $A_{s,i+1}/A_{si}$ rapidly approaches minus one when i increases. Therefore the function

$$R(t)=-A_{s0}t^{2s+1}(1+t)^{-(n+1)/n}[Q_s(t)]^{-1}$$

is a good estimate of the relative error. Thus, the upper bound B_s of $|R(t)|$ in the range $(-t_1, t_2)$ is minimized, if

$$R(-t_1)=|R(t_2)|=B_s\,. \qquad (21)$$

The equations (19) and (21) determine the numerical values of t_1, t_2 and with them the number c. Applying this method to first four values $s=1, 2, 3, 4$ of s, we find for the range (a, b) with $b=2a$, that is, for $r=2$:

$p_{11}=(n+1)/2n$, $q_{11}=(n-1)/2n$

$p_{21}=(2n+1)/2n$, $p_{22}=(n+1)p_{21}/6n$,
$q_{21}=(2n-1)/2n$, $q_{22}=(n-1)q_{21}/6n$

$p_{31}=(3n+1)/2n$, $p_{32}=(2n+1)p_{31}/5n$,
$p_{33}=(n+1)p_{32}/12n$
$q_{31}=(3n-1)/2n$, $q_{32}=(2n-1)q_{31}/5n$,
$q_{33}=(n-1)q_{32}/12n$

$p_{41}=(4n+1)/2n$, $p_{42}=3(3n+1)p_{41}/14n$,
$p_{43}=(2n+1)p_{42}/9n$, $p_{44}=(n+1)p_{43}/20n$
$q_{41}=(4n-1)/2n$, $q_{42}=3(3n-1)q_{41}/14n$,
$q_{43}=(2n-1)q_{42}/9n$, $q_{44}=(n-1)q_{43}/20n$.

In general $A_{s0}=p_{ss}\,q_{ss}/(2s+1)n>0$.

Solving the equations (19) and (21) with $r=2$, we find that t_1 and t_2 vary so slowly with n that for small n we can take $t_1=0.3$ and $t_2=0.4$ so that $c=10a/7$.

Returning now to the square root, we have two subintervals $0.25<f\leqslant 0.5$ and $0.5\leqslant f<1$. First taking $s=1$ so that $p_{11}=3/4$ and $q_{14}=1/4$, we have $A_{10}=2^{-5}$. In the first interval, $0.25<f\leqslant 0.5$, $c=5/14$ and $\sqrt{c}\approx 0.597\,614\,3$. Computing the upper bound B_2 of the relative error e_0, we find that in $\sqrt{f}\approx P_1(f)/Q_1(f)$, that is, in

$$\sqrt{f}=c_{10}-\frac{c_{11}}{f+c_{12}}+e_0\sqrt{f} \qquad (0.25<f\leqslant 0.5) \qquad (22)$$

one has $|e_0|\leqslant B_1=14\times 10^{-4}$. Applying Heron's method twice to the initial approximation $x_0=c_{10}-c_{11}(f+c_{12})^{-1}$, we reduce

$|e_0|$ to $|e_2|<\frac{1}{2}10^{-12}$, since $|e_2|\approx\frac{1}{2}|e_1^2|\approx|e_0|^4/8<B_1^4/8=\frac{1}{2}10^{-12}$.

This proves that (22) allows the computation of *twelve* correct significant digits in three operations. The constants are: $c_{10}=1.792\,843$, $c_{11}=1.707\,469$ and $c_{12}=1.071\,429$. If $0.5<f\leqslant 1$ another set of constants is to be used in (22), namely $c_{10}^*=c_{10}\sqrt{2}=2.535\,463$, $c_{11}^*=2c_{11}\sqrt{2}=4.829\,452$ and $c_{12}^*=2c_{12}=2.142\,858$.

Thus the number of stored constants is six.

To mention some examples: (22) gives for $f_1=0.36$ and $f_2=0.81$ the approximations $0.6-10^{-6}$ and $0.9-65\times 10^{-6}$ so that the relative errors are -16.10^{-7} and -7.10^{-5}. We add that the approximate values $t_1=0.3$ and $t_2=0.4$ do not balance $R(-t_1)$ and $|R_2(t_2)|$ exactly: their ratio is equal to 1.4.

Now taking $s=2$, we have $p_{21}=5/4$, $p_{22}=5/16$, $q_{21}=3/4$, $q_{22}=1/16$, $A_{20}=2^{-9}$ and $B_2=10^{-5}$. Applying (1) this time to the first guess

$$\sqrt{f}\approx x_0=c_{20}-\frac{c_{21}|}{|f+c_{22}}-\frac{c_{23}|}{|f+c_{24}} \qquad (0.25<f\leqslant 0.5) \qquad (23)$$

only once, first ten correct significant digits are obtained in three operations: $|e_1|\leqslant\frac{1}{2}B_2^2=\frac{1}{2}10^{-10}$. Applying (1) to (23) twice, we obtain twenty correct significant digits in four operations. The constants are: $c_{20}=5(5/14)^{\frac{1}{2}}$; $c_{21}=20c_{20}/7$; $c_{22}=47/14$; $c_{23}=4/49$ and $c_{24}=3/14$. In the interval $0.5<f\leqslant 1$ the constants to be used in (23) change:

$c_{20}^*=c_{20}\sqrt{2}$; $c_{21}^*=2c_{21}\sqrt{2}$; $c_{22}^*=2c_{22}$; $c_{23}^*=4c_{23}$ and $c_{24}^*=2c_{24}$.

In the computation of square roots the values of s exceeding 2 are not economical since (1) necessitates only one operation.

• *Square root—decimal machine*

Here $N=10^{2m}f$ and $10^{-2}<f\leqslant 1$. The range $(10^{-2}, 1)$ is subdivided into four subintervals because if $s=2$ our method does not work for $r\geqslant 3.8$. Estimates of relative error are based on the convergence of the series of general term $(-t)^k$ and the necessary condition $t_2<1$ is a limitation imposed on $r=b/a$. From it follows $t_1>1-2/r$ since $r(1-t_1)=1+t_2$. To satisfy (21) we must have

$R(2/r-1)<|R(1)|$, that is in our case $(n=2)$:

$$r^{2s-\frac{1}{2}}Q_s(2/r-1)>Q_s(1)\,(r-2)^{2s+1}\,.$$

This inequality proves that $r\leqslant r_s$, where $r_1\cong 3.72$, $r_2\cong 3.81$ etc., r_s increasing with s. Since $\sqrt[3]{100}\approx 4.64$, while $\sqrt[4]{100}\approx 3.16$ it is necessary to subdivide the interval $(10^{-2};\ 1)$ as follows: $(10^{-2}r^{k-1};\ 10^{-2}r^k)$ for $k=1, 2, 3, 4$ and $r=10^{\frac{1}{2}}$. In the first subinterval $(k=1)$ we have, taking $s=2$, $t_1=0.4475$; $221c=4$ and thus

$$\sqrt{f}\approx x_0=d_0-\frac{d_1|}{|f+d_2}-\frac{d_3|}{|f+d_4} \qquad (1<100f\leqslant r) \qquad (24)$$

where $d_0=0.674\,055$; $d_1=0.098\,002$; $d_2=0.170\,836$; $d_3=0.000\,211\,4$; $d_4=0.010\,904$ and $|e_0|<17.10^{-6}$. For instance $\sqrt{0.0289}\approx x_0=0.17-15.10^{-6}$, but $\sqrt[4]{10}\approx x_0=\sqrt[4]{10}-3.10^{-6}$.

Applying (1) to x_0 once, first nine correct significant digits are obtained in three operations. The values of d_j in the other three intervals $(k=2, 3, 4)$ are obtained by observing that if f is in the range (ra, rb), then f/r belongs to the range (a, b). Therefore denoting the coefficient in (24) by d_j for the range (a, b) those d_j^* for the range (ra, rb) are: $d_0^*=d_0r^{\frac{1}{2}}$, $d_1^*=d_1r^{3/2}$, $d_2^*=d_2r$,

$d_3^* = d_3 r^2$, $d_4^* = d_4 r$. The number of stored constants is equal to thirteen.

Applying (1) twice, eighteen correct digits are obtained in four operations.

• *Cubic root—binary machine*

It is possible to accelerate the convergence of Newton's method when it is applied to computation of radicals. Instead of the equation $x^{2m+1} - N = 0$ (radicals of even order are omitted because they are reduced to those of odd order and square roots), we consider the equation

$$f(x) = x^{m+1} - Nx^{-m} = 0 \qquad (m \geqslant 1) \tag{25}$$

which for $m = 1$ yields our recurrence formula (18) for the sequence of approximations to $N^{1/3}$. When (25) is used the relative error e_j of the $j-$th approximation x_{mj} to $N^{1/m}$ decreases more rapidly than the relative errors of approximations deduced from the equation $x^{2m+1} - N = 0$:

$$|e_{j+1}| \approx (m+1)m|e_j|^3/3 . \tag{26}$$

To prove (26) replace $x_{m,j+1}$ in

$$x_{m,j+1} - x_{mj} = (Nx_{mj}^{-m} - x_{mj}^{m+1})[(m+1)x_{mj}^{m} + mNx_{mj}^{-m-1}]^{-1}$$

by $(1+e_{j+1})N^{1/(2m+1)}$

and x_{mj} by $(1+e_j)N^{1/(2m+1)}$.

The result is

$$e_{j+1} = e_j + (1+e_j)[1-(1+e_j)^{2m+1}][(m+1)(1+e_j)^{2m+1} + m]^{-1}$$

and this gives

$$e_{j+1} = (m+1)me_j^3[1+0(e_j)][3+0(e_j)]^{-1}.$$

In particular, applying (18) ($m=1$) we have $|e_{j+1}| \leqslant 2|e_j|^3/3$. For the fifth root $N^{1/5}$ (25) takes the form $x^3 - Nx^{-2} = 0$ and $|e_{j+1}| \leqslant 2|e_j|^3$, the recurrence relation being

$$x_{n+1} = x_n[2(x_n^4 + N/x_n) + N/x_n][2(x_n^4 + N/x_n) + x_n^4]^{-1}. \tag{27}$$

Returning to the cubic root, we have $N = 2^{3m}f$ and $N^{1/3} = 2^m f^{1/3}$, where $2^{-3} < f \leqslant 1$. Subdividing this range into three intervals $(2^{-3}; 2^{-2})$, $(2^{-2}; \frac{1}{2})$ and $(\frac{1}{2}; 1)$, we consider first the case $2^{-3} < f \leqslant 2^{-2}$, so that $r = 2$. We found general expressions of p_{si}, q_{si} for any radical. For a cubic root ($n=3$) $s = 1, 2, 3, 4$ they give the coefficients of P_s and Q_s as well as A_{s0}. For instance $A_{10} = 2.3^{-4}$, $A_{20} = 7.3^{-7}/2$, $A_{30} = 2.3^{-9}$ and $A_{40} = 3^{-13}143/14$. Taking $t_1 = 0.3$, $t_2 = 0.4$ and computing the upper bounds B_s for the absolute value of the relative error, we found $B_1 = 12.10^{-4}$, $B_2 = 83.10^{-7}$, $B_3 = 56.10^{-9}$, $B_4 = 38.10^{-11}$. Extrapolating we conclude that $B_5 = 3.10^{-12}$ which proves that in five operations the Padé approximation to cube root $P_5(f)/Q_5(f)$ yields first eleven correct significant digits without using Newton's method at all, while $P_4(f)/Q_4(f)$ yields nine digits in four operations and twenty-eight in seven, if (18) is applied once. Since $2B_1^3/3 \approx 10^{-9}$, applying (18) once to $P_1(f)/Q_1(f)$, that is, to

$$f^{1/3} \approx x_0 = 1.12625 - \frac{0.30167}{f + 0.35714}, \qquad (2^{-3} < f \leqslant 2^{-2}) \tag{28}$$

first eight correct significant digits are obtained in four operations, provided $2^{-3} < f \leqslant 2^{-2}$. Take for instance $f = 0.125$, so that $x_0 = 0.50056$. Applying (18), we find $x_1 = 0.5 + 4.10^{-10}$. If $f = 0.25$, then $x_0 = 0.62938$ [the true value of $(0.25)^{1/3}$ is equal to 0.62996 05247] and $x_1 = 0.62996$ 05275.

Naturally in two other intervals $(2^{-2}; 2^{-1})$ and $(2^{-1}; 1)$ the constants in (28) change. Thus, in $(2^{-2}; 2^{-1})$ $x_0 = 1.418\,986 - 0.760\,1607(f + 0.71428)^{-1}$ while in $(2^{-1}; 1)$ $x_0 = 1.78781 - 1.91548(f + 1.42856)^{-1}$. In all, nine stored constants are needed in the case $s = 1$. If $s = 2$, then

$$\sqrt[3]{f} \approx x_0 = a_0 - \frac{a_1|}{|f + b_1} - \frac{a_2|}{|f + b_2} \tag{29}$$

the values of fifteen constants being:

	Interval		
	$2^{-3} \leqslant f \leqslant 2^{-2}$	$2^{-2} \leqslant f \leqslant 2^{-1}$	$2^{-1} \leqslant f \leqslant 1$
a_0	1.576 745	1.986 574	2.502 926
a_1	1.267 028	3.192 710	8.045 125
b_1	1.153 061	2.306 122	4.612 244
a_2	0.022 490 6	0.089 962 4	0.359 849 6
b_2	0.096 938 8	0.193 877 6	0.387 755 2

Here $c = 5/28$ and $\sqrt[3]{c} = 0.563\,123\,4$, $t_1 = 0.3$, $t_2 = 0.4$. The values in the first interval are: $a_0 = 2.8\sqrt[3]{c}$; $a_1 = 2.25\sqrt[3]{c}$; $b_1 = 113/98$, $a_2 = 54/2401$ and $b_2 = 19/196$. Using (29), first four correct significant digits are obtained in two divisions only: $|e_0| < 75.10^{-7}$. Applying to x_0 the recurrence relation (18) once, fifteen correct digits are obtained in five operations since $|e_1| \leqslant 2|e_0|^3/3 < 3.10^{-16}$.

• *Cubic root—decimal machine*

The range $10^{-3} \leqslant f \leqslant 1$ should be subdivided into five intervals $I_k[10^{-0.6(6-k)}; 10^{-0.6(5-k)}]$ for $k = 1, 2, 3, 4, 5$ since the ratio $r = 1000^{1/5} = 3.981 < 4$ and then $s = 3$ can be taken, but not $s \leqslant 2$. We have then $t_1 = 0.51836$, $t_2 = 0.91744$ and the upper bound B_3 of the relative error $|e_0|$ is equal to 3.10^{-5}. Therefore in three operations four correct significant digits are obtained without using (18). If (18) is applied once to the first guess $x_0 = P_3(f)/Q_3(f)$, then thirteen correct digits are obtained in six operations since $|e_1| \leqslant 2|e_0|^3/3 < 2.10^{-14}$. The number of stored constants is thirty-five, since in each of five intervals I_k seven constants are needed.

Received January 12, 1959