E. G. Kogbetliantz

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

Abstract: Rational (R) and polynomial (P) approximations to Arctan N are studied with the aim of computing this function, to any prescribed accuracy and without unduly increasing the number PC of stored constants, in a minimum number M of multiplications (and divisions for R approximations). The number Dg of first correct *significant* digits in principle is not bounded. The results corresponding to the values 8, 10, 18 and 20 of this number are as follows:

Approximation	*Point (Computation)*	*Single Precision* Dg	M	PC	*Double Precision* Dg	M	PC
Rational (R)	Floating	8	4*	21	18	6*	19
			5	9		7	17
	Fixed	10	5	14	20	6*	30
			6*	9		7	18
Polynomial (P)	Floating	8	5	10	17	8	21
			6	8	18	9	14
	Fixed	10	6	11	20	9	22
			7	9		10	15

If M is increased, subroutines with smaller PC are easily deduced from our general results. Thus, for instance, rational approximations with $Dg = 6$ can be obtained in three multiplications only, if $PC = 19$ (combination $m^* = 3$, $q = 10$); but *the same accuracy* $Dg = 6$ characterizes also the cases $M = 4$ with $PC = 11$ and $M = 5$ with $PC = 7$ (combinations $m^* = 4$, $q = 6$ and $m = 5$, $q = 4$).

If polynomial approximations are used, $Dg = 6$ is obtained for $M = 5$, $PC = 7$, but also for $M = 4$ and $PC = 11$. No subroutines with a stored table of values of Arctan x are considered.

Introduction

The aim of this paper is to formulate the most economical procedures for the approximate evaluation of Arctan N adapted to binary and/or decimal computing machines and sufficiently flexible to yield as many correct significant digits as desired, and this for any value of N in $(-\infty, +\infty)$.

Two mathematical tools are used here to form our rational and polynomial approximations to the function Arctan N. Successive convergents (approximants) $K_m(t)$ of the classical continued fraction found in 1812 by Gauss*

$$\text{Arctan } t = \frac{t|}{|1} + \sum_{s=1}^{\infty} \frac{s^2 \cdot t^2|}{|2s+1} \qquad \text{(I)}$$

yield a sequence of rational approximations the accuracy of which increases very rapidly with m. Our approximating polynomials are partial sums $S_m(x)$ of first $(m+1)$ terms of the series.

*C. F. Gauss: *Werke*, 1876. v. III. See also: H. S. Wall, *Continued Fractions*, Van Nostrand, 1948, p. 343; form. (90.3).

$$\text{Arctan}\,(x\cdot\tan 2\theta) = 2\cdot\sum_{n=0}^{\infty}\frac{(-1)^n\tan^{2n+1}\theta}{2n+1}\cdot T_{2n+1}(x). \qquad (|x|\leqq 1) \quad \text{(II)}$$

This expansion of the Arctangent into a Fourier series of Tchebychev polynomials $T_n(x)$ converges absolutely and uniformly in $|x|\leqq 1$ for $0<\theta<\pi/4$, so that Arctan N, $N=x\tan 2\theta$, is represented by (II) in as large an interval $0<N<\tan 2\theta$ as we please, but the convergence becomes very slow when $\theta<\pi/4$ is near $\pi/4$. Expression (II) will be used for small values of θ. For $x=1$, the series (II) reduces to the classical Gregory series

$$\theta = \sum_{n=0}^{\infty}\frac{(-1)^n}{2n+1}\cdot\tan^{2n+1}\theta,$$

while $\theta=\pi/4$ yields a curious generalization of Leibnitz' series for $\pi/4$:

$$\pi/4 = \sum_{n=0}^{\infty}\frac{(-1)^n}{2n+1}\cdot T_{2n+1}(x). \quad (0<x\leqq 1)$$

Both (I) and (II) will be used in a reduced range $0\leqq N\leqq \tan(\pi/2q)$, the parameter q having positive and integral values. The number Dg of first correct significant digits of an approximation K_m or S_m is an increasing function of m and q. So also is the number PC of precomputed and stored constants. The number PC increases quite rapidly with q and this precludes the use of too large values of q.

The problem studied in this paper can be formulated as follows: For a given value of Dg (that is for a prescribed accuracy), find the optimum combination (m,q) characterized by the least possible values of M and PC, M denoting the number of multiplications and/or divisions.

In the sequel we consider in detail the five most interesting cases: Dg = 6, 8, 10, 18 and 20. These cases correspond to fixed and floating-point computations with single or double precision.

Rational approximations

• *1. Proof of the expansion (II)*

Let us expand the function of t

$$F(t) = \text{Arctan}\,[2a\cdot\text{Cos}\,t/(1-a^2)]$$

in the interval $0\leqq t\leqq\pi$ into its Fourier Cosine series

$$F(t) = \tfrac{1}{2}A_0(a) + \sum_{n=1}^{\infty}A_n(a)\cdot\text{Cos}\,nt, \quad (0<a<1)$$

and compute the coefficients $A_n(a)$ by

$$\pi A_n(a) = 2\int_0^{\pi}F(t)\cdot\text{Cos}\,nt\cdot dt.$$

Since

$$(1+2a^2\,\text{Cos}\,2t+a^4)F'(t) = -2a(1-a^2)\cdot\text{Sin}\,t,$$

an integration by parts gives, for $n\geqq 1$:

$$\pi\cdot n\cdot A_n(a) = 2a(1-a^2)[j_{n-1}(a)-j_{n+1}(a)],$$

where

$$j_n(a) = \int_0^{\pi}(1+2a^2\,\text{Cos}\,2t+a^4)^{-1}\cdot\text{Cos}\,nt\,dt. \qquad (1)$$

Substituting into (1) the expansion

$$(1-a^4)(1+2a^2\,\text{Cos}\,2t+a^4)^{-1} = 1+2\sum_{m=1}^{\infty}(-1)^m\cdot a^{2m}\,\text{Cos}\,2mt,$$

we find first of all that $j_{2n+1}(a)=0$, $n\geqq 0$, so that $A_{2n}=0$.

On the other hand

$$(1-a^2)\cdot j_{2n}(a) = (-1)^n a^{2n}(1+a^2)^{-1}\int_0^{\pi}2\,\text{Cos}^2\,2nt\cdot dt = (-1)^n\cdot\pi a^{2n}/(1+a^2)$$

and therefore $A_{2n+1}=2(-1)^n\cdot a^{2n+1}/(2n+1)$, so that

$$F(t) = \tfrac{1}{2}A_0(a) + 2\sum_{n=0}^{\infty}(-1)^n a^{2n+1}\cdot\text{Cos}\,[(2n+1)t]/(2n+1). \qquad (2)$$

The substitution $t=\pi/2$ proves that $A_0(a)=0$, since $F(\pi/2)=0$. For $a=\tan\theta$ and Cos $t=x$, (2) becomes (II).

We add that (II) can also be transformed into a Tchebychev expansion of the function $f(x)=\text{Log}\,[(1+2ax+a^2)/(1-2ax+a^2)]$, namely into

$$f(x) = 4\sum_{n=0}^{\infty}a^{2n+1}\cdot T_{2n+1}(x)/(2n+1) \quad (|a|<1,\ |x|\leqq 1) \qquad \text{(III)}$$

To deduce (III) replace a by ia in (2) and use the relation

$$2i\,\text{Arctan}\,(iz) = \text{Log}\,[(1-z)/(1+z)].$$

The series (III) converges in $-1\leqq x\leqq 1$ absolutely and uniformly, provided $|a|<1$. It is a source of very accurate polynomial approximations to the natural logarithm Log N, the argument x being defined by $x=\alpha-\alpha(N+1)^{-1}$ with $2\alpha=a+a^{-1}$, since the constant a can be chosen very small. Series (III) was obtained by Mr. Germizoglou (IBM-France) by a direct integration of the generating function of Tchebychev polynomials.

• *2. Reduction to a smaller range*

Let us denote the integral part of a number z by $[z]$. Given a known integer q, we subdivide the infinite range $(0,\infty)$ of N into $\gamma=[q/2]+1$ intervals $I_k[a_{k-1}, a_k]$; $1\leqq k\leqq\gamma$; where $a_0=0$, $a_\gamma=\infty$ and $a_k=\tan[(k-\frac{1}{2})\pi/q]$ for $1\leqq k\leqq\gamma-1$. The intervals I_k are half-open intervals, so that N belongs to I_k (denoted by $N\subset I_k$), if $a_{k-1}\leqq N<a_k$.

The range $(0, \pi/2)$ of θ = Arctan N is also subdivided by points $\theta_k=(k-\frac{1}{2})\cdot\pi/q$ into γ intervals i_k, $1\leqq k\leqq\gamma$, and θ = Arctan N belongs to i_k, if $\theta_{k-1}\leqq\theta<\theta_k$. Here $\theta_0=0$ and $\theta_\gamma=\pi/2$. If $N\subset I_k$, then $\theta\subset i_k$ and vice versa. The length of i_1 is $\theta_1=\pi/2q$. If q is *even* then the last interval $i_\gamma=(\theta_{\gamma-1};\pi/2)$ is equal to the first, but the $\gamma-2$ interior intervals i_k, $2\leqq k\leqq\gamma-1$, are of the length π/q. If q is *odd*, then the length of i_γ is also equal to π/q and there is only one, namely the first, interval of length $\pi/2q$. In general, there are $[(q-1)/2]$ intervals of length π/q and in them θ = Arctan N is computed with the aid of the addition theorem:

$$\text{Arctan } N = k\pi/q + \text{Arctan } z_k \quad (N \subset I_k) \tag{3}$$

where

$$z_k = z_k(N,q) = \alpha_k - \beta_k(N+\alpha_k)^{-1} \tag{4}$$

with $\alpha_k = \text{Cotan}(k\pi/q)$ and $\beta_k = 1+\alpha_k^2$.

For an *even* q and $N \subset I_\gamma$ we will use in this last interval (of length $\pi/2q$ in θ) the relation

$$\text{Arctan } N = \pi/2 - \text{Arctan } (N^{-1}). \quad (N \subset I_\gamma, q = \text{even}) \tag{5}$$

In the first interval, $N \subset I_1$, Arctan N is computed directly as such.

Given N, the first thing to do is to locate N in some I_k, $1 \leqq k \leqq \gamma$. Therefore, $[q/2]$ constants $a_k = \tan\theta_k = \tan[(k-\frac{1}{2})\pi/q]$ are to be stored. In many cases it is possible to reduce their number, expressing some of them in terms of the others.

The computation of z_k using (4) involves, for each one of $[(q-1)/2]$ intervals where (4) is to be used, two constants, namely Cotan $(k\pi/q)$ and Cosec^2 $(k\pi/q)$. As will be seen below, it is possible to save one multiplication in approximating Arctan z_k with the aid of $K_{2m}(z_k)$ and computing this convergent of *even* order $2m$ as a function $K^*_{2m}(t_k)$, not of z_k, but of $t_k = \lambda_m^{-1} \cdot z_k$, λ_m, this being a suitable constant. Then (4) will take the form

$$t_k = \lambda_m^{-1} \cdot z_k = \alpha^*_k - \beta^*_k \cdot (N+\gamma_k)^{-1} \quad (N \subset I_k) \tag{4*}$$

which involves three constants: γ_k = Cotan $(k\pi/q)$, $\alpha^*_k = \lambda_m^{-1} \cdot$ Cotan $(k\pi/q)$ and $\beta^*_k = \lambda_m^{-1} \cdot$ Cosec^2 $(k\pi/q)$, instead of two. Therefore, in saving one multiplication the number of stored constants has been increased by $[(q-1)/2]$.

In the last step (3) again $[(q-1)/2]$ constants $k\pi/q$ are needed. Finally, there are also m constants involved in the computation of Arctan z_k with the aid of $K_m(z_k)$, $K^*_m(t_k)$ or $S_m(z_k)$. If q is even, (5) adds, using (15) or (16), m constants, but as will be shown below, the use of (5) can be avoided.

Thus, if K_m or S_m are used, the total number of stored constants is at most equal to

$PC = 3[(q-1)/2] + [q/2] + m[3+(-1)^q]/2$

while, if K^*_m, m = even, is applied, $PC^* = PC + [(q-1)/2]$.

These two numbers should be considered in fact as upper bounds for PC. Some of the constants used in the subroutine are linear combinations of other constants computable by the machine in one or two additions only. Such constants need not be stored.

• 3. *Relative error of rational approximations*

The numerator and denominator of the m-th convergent $K_m(t)$ of (I) are denoted in the sequel by $t \cdot P_m$ and Q_m. They are odd and even polynomials of degrees $2[(m-1)/2]+1$ and $2[m/2]$ respectively:

$$P_m = \sum_{s=0}^{2s \leqq m-1} p_{sm} \cdot t^{2s}; \quad Q_m = \sum_{0=s}^{2s \leqq m} q_{sm} \cdot t^{2s} \tag{6}$$

Here $p_{0m} = q_{0m} = (2m-1)!!$; $p_{13} = 4$, $q_{13} = 9$; $p_{14} = 55$, $q_{14} = 90$, $q_{24} = 9$ and, for $5 \leqq m \leqq 10$, see Tables 1a and 1b below.

We will need in the sequel the following expressions:

$q_{m;2m} = [(2m-1)!!]^2$; $\quad q_{m-1,2m} = m(2m+1) \cdot q_{m,2m}$;
$q_{m,2m+1} = [(2m+1)!!]^2$; $\quad p_{m,2m+1} = (2m!!)^2$;
$p_{m,2m+2} = p_{m+1,2m+3} - q_{m,2m+1} = [(2m+2)!!]^2 - [(2m+1)!!]^2$.

The absolute value $R_m(t)$ of the relative error in the *first* interval I_1, namely $0 < R_m(t) = (-1)^m \cdot [1 - K_m(t)/\text{Arctan } t]$, is an even and increasing function of t. It is sufficient to consider $R_m(t)$ for $t > 0$. If the range of $|t|$ is $(0, T)$,

$$R_m(t) \leqq R_m(T). \quad (0 < |t| \leqq T) \tag{7}$$

Equation (7) is justified by proving that $R'_m(t) > 0$ in $0 < t \leqq T$. Denoting the absolute value of the absolute error by $E_m(t)$ so that

$E_m(t) = (-1)^m[\text{Arctan } t - K_m(t)]$,

Table 1a **Values of $p_{sm}/9$**

m	$s=1$	$s=2$	$s=3$	$s=4$
5	245/3	64/9		
6	1,190	231		
7	19,250	5,943	256	
8	345 × 1,001	147,455	15,159	
9	6,825 × 1,001	3,735 × 1,001	638,055	16,384
10	29,580 × 1,001	19,782 × 5,005	962,676 × 25	61,567 × 25

Table 1b **Values of $q_{sm}/9$**

m	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
5	350/3	25			
6	1,575	525	25		
7	24,255	11,025	1,225		
8	420 × 1,001	242,550	44,100	1,225	
9	8,100 × 1,001	5,670 × 1,001	161,700 × 9	99,225	
10	172,175 × 1,001	5,670 × 1,001 × 25	47,250 × 1,001	99,225 × 55	99,225

one easily finds that

$E'_m(t)=(m!)^2t^{2m}\cdot(1+t^2)^{-1}\cdot Q_m^{-2}(t)\geqq 0$

and

$E''_m(t)=2(m!)^2\cdot G_m(t)\cdot(1+t^2)^{-2}\cdot Q_m^{-3}(t)\cdot t^{2m-1}$

where

$G_m(t)=[m(1+t^2)-t^2]Q_m(t)-t(1+t^2)\cdot Q'_m(t).$

Now

$(1+t^2)$ (Arctan $t)^2\cdot R'_m(t)=F(t)=(1+t^2)$ Arctan $t\cdot E'_m(t)-E_m(t)$ and, thanks to $E_m(o)=0$ and to $(1+t^2)$ Arctan $t\geqq t$,

$$F(t)\geqq t\cdot E'_m(t)-E_m(t)=\int_o^t uE''_m(u)\cdot du.$$

Therefore, $F(t)$ and $R'_m(t)$ are positive in $(0, T)$, if $G_m(t)$ is, since then $E''_m(u)$ is positive. But the expression of $G_m(t)$ is as follows:

$$G_m(t)=mq_{om}+\sum_{s=1}^{2s\leqq m}[(m-2s)(q_{sm}+q_{s-1,m})+q_{s-1,m}]t^{2s}-r_m(t),$$

where $r_m(t)\equiv 0$ if m is odd, but

$r_{2m}(t)=q_{m,2m}\cdot t^{2m+2}.$

It is seen, therefore, that $G_{2m+1}(t)>0$ for all values of t, but for large t $G_{2m}(t)$ can become negative. Omitting in $G_{2m}(t)$ all positive terms except the term for which $s=m$, one has the inequality

$G_{2m}(t)>(q_{m-1,2m}-q_{m,2m}\cdot t^2)\cdot t^2{}_m.$

This result proves that $G_{2m}(t)$ remains positive at least for $t^2>q_{m-1,2m}/q_{m,2m}$, that is for $t^2<m(2m+1)$ and a fortiori for $t\leqq\sqrt{21}$ if $m\geqq 3$. Since only the values $q\geqq 2$ and $T=\tan(\pi/2q)$ are considered it can be concluded that for $q\geqq 2$ and $m\geqq 3$ the lemma is proved and $R_m(T)$ is the upper bound R_{mq} of $R_m(t)$ in $0\leqq t\leqq\tan(\pi/2q)$:

$R_{mq}=R_m(T).\quad (T=\tan(\pi/2q))$

For $q\leqq 6$ (and $3\leqq m\leqq 10$) $R_m(T)$ was computed directly. For $q\geqq 7$ an upper bound B_{mq} was used. It is obtained as follows: Since Arctan $T=\pi/2q$, we have

$$R_m(T)=2q\pi^{-1}\cdot E_m(T)=2q\cdot\pi^{-1}\cdot\int_o^T E'_m(u)\cdot du,$$

where

$E'_m(u)\leqq(m!)^2u^{2m}\cdot Q_m^{-2}(0)=(m!)^2u^{2m}/[(2m-1)!!]^2.$

Thus

$R_{mq}=R_m(T)\leqq 2q\cdot C_m(T/2)^{2m+1}$

the constant C_m being very near to one:

$0.9312<C_m=2^{2m+1}(m!)^2/\{(2m-1)!!(2m+1)!!\pi\}<0.9775.$
$(3\leqq m\leqq 10)$

Replacing C_m by one, B_{mq} is defined as follows:

$R_{mq}<B_{mq}=2q\cdot[\frac{1}{2}\tan(\pi/2q)]^{2m+1}.$

If $q\geqq 7$ this upper bound is sufficiently accurate for our purpose. How good it is for large q can be illustrated on the example of $B_{7,18}$. For $m=7$, $q=18$ it is found that $B_{7,18}=1.48\times 10^{-19}$, that is, Log $B_{7,18}=-18.83$.

A direct computation of $R_{7,18}$ based on the formulae

$K_7[\tan(\alpha/2)]=N_7(\alpha)/D_7(\alpha)\quad(\alpha=\pi/18)$

$N_7(\alpha)=45{,}619\ \text{Sin}\ \alpha+29{,}155\ \text{Sin}\ 2\alpha+5{,}155\ \text{Sin}\ 3\alpha+181.5\ \text{Sin}\ 4\alpha$

$D_7(\alpha)=85{,}750+116{,}620\ \text{Cos}\ \alpha+34{,}300\ \text{Cos}\ 2\alpha+3{,}500\ \text{Cos}\ 3\alpha+70\ \text{Cos}\ 4\alpha$

and the values of Sine and Cosine of angles equal to 10°, 20°, 30°, 40° taken with first twenty correct digits after the dot, yields $R_{7,18}=1.23\times 10^{-19}$, or Log $R_{7,18}=-18.91$.

To insure in the final value of Arctan N first Dg correct significant digits we compare $L_{mq}=|\text{Log}_{10}B_{mq}|$ to $Dg+0.3$. The combination (m,q) yields Dg correct significant digits if $L_{mq}>Dg+0.3$. It is sufficient to know L_{mq} with an accuracy of 0.05. But then the error made in replacing in the definition of B_{mq} tan $(\pi/2q)$ by its argument $\pi/2q$ is negligible for $q\geqq 7$, $m\leqq 10$ and the expression of L_{mq} can be simplified:

$L_{mq}=|\text{Log}_{10}B_{mq}|\approx(2m+1)\text{Log}_{10}(4q/\pi)-\text{Log}_{10}(2q)$

that is

$L_{mq}\approx(2\ \text{Log}\ q+0.21)\cdot m-0.20.$

Thus, for a fixed value of $q\geqq 7$, L_{mq} is a linear function of m and the same fact is confirmed for $q\leqq 6$ by a direct computation of R_{mq}, which gave slightly smaller coefficients of m than 2 Log $q+0.21$:

q =	2	3	4	5	6
Coefficient of m =	0.76	1.14	1.40	1.60	1.89

This result is represented in Fig. 1, where five horizontal lines mark the critical values 6.3; 8.3; 10.3; 18.3; and 20.3 of L_{mq}. Combinations represented by points (m, q) immediately above or on a horizontal line—(marked by circles)—insure the corresponding accuracy of first 6, 8, 10, 18 or 20 significant digits respectively.

Thus, many combinations (m, q) have the same accuracy (see Table 2).

Table 2 **Combinations (m, q) with same Dg**

Dg = 6	for	(3; 10), (4; 6), (5; 4), (6; 3), (9; 2).
Dg = 8	for	(3; 20), (4; 9), (5; 6), (6; 5), (7; 4), (8; 3).
Dg = 10	for	(4; 16), (5; 9), (6; 6), (7; 5), (8; 4), (10; 3).
Dg = 18	for	(7; 18), (8; 12), (9; 9), (10; 7).
Dg = 20	for	(8; 15), (9; 12), (10; 9).

A direct computation of the relative error rejected the combinations (4, 5); (6, 4) and (7, 16) since for $t=T$ the errors are equal to 6×10^{-7}, 6.4×10^{-9} and 6.2×10^{-19}.

To choose between many combinations listed in Table 2 with the same value of Dg, it now is necessary to study the number M of multiplications and the number PC of constants involved in each of these combinations.

• 4. *Study of M and PC*

The convergents $K_{2m}(t)$ and $K_{2m+1}(t)$ of even and odd order can be computed in the same optimum number $m+2$ of

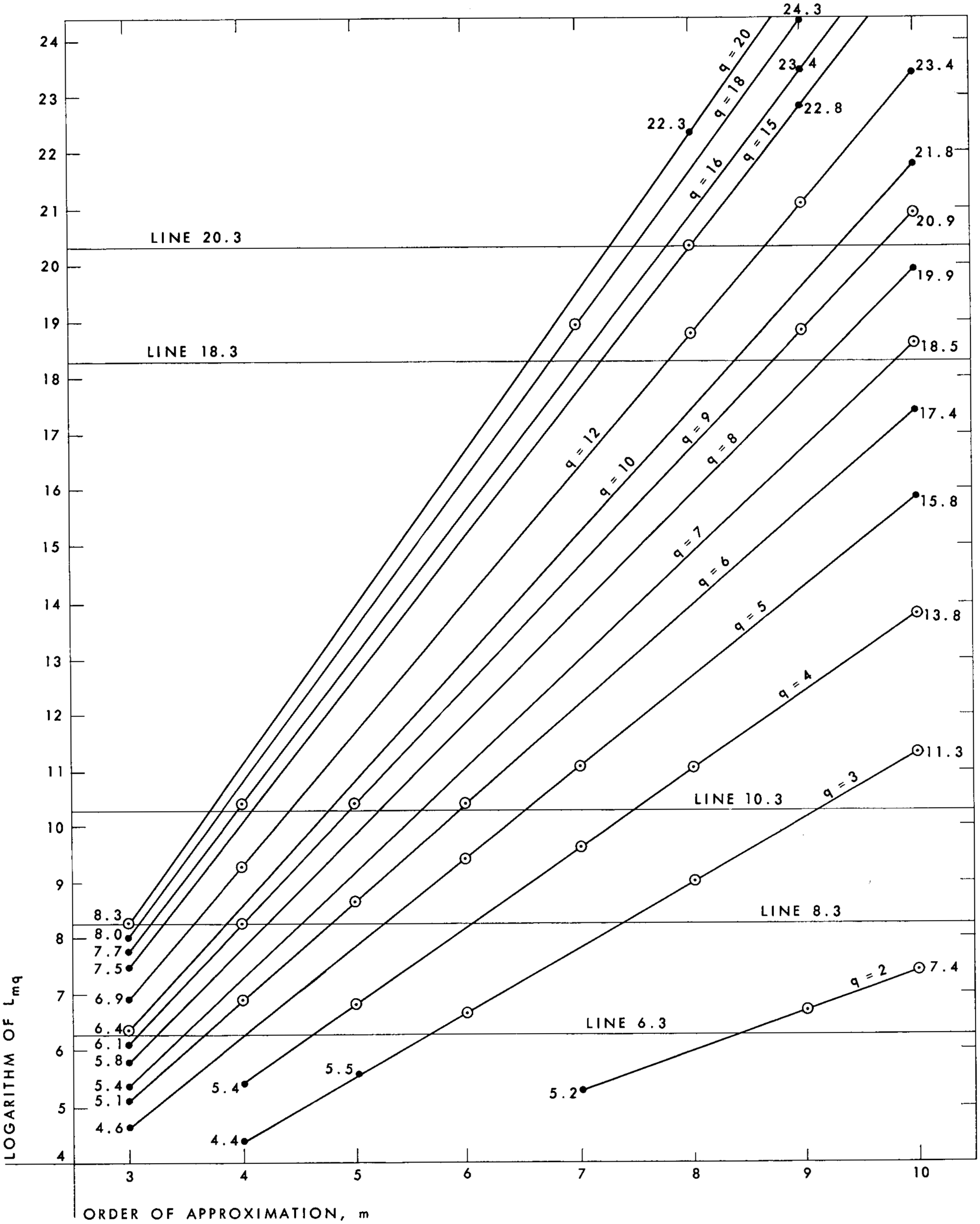

Figure 1 **Graphs of L (m, q).**

multiplications (and/or divisions) if they are given the following forms:

$$(m \geqq 1) \quad K_{2m}(t) = \lambda_m \cdot t \cdot \{ t^2 + B_{0,2m} + \sum_{s=1}^{m-1} \frac{-A_{s,2m}|}{|t^2 + B_{s,2m}|} \}^{-1} \tag{8}$$

with $\lambda_m = p_{m-1,2m}/q_{m,2m} = [(2m)!!/(2m-1)!!]^2 - 1$, and

$$(m \geqq 2) \quad K_{2m+1}(t) = t \cdot \{ A_{0,2m+1} + \sum_{s=1}^{m} \frac{-A_{s,2m+1}|}{|t^2 + B_{s,2m+1}} \}. \tag{9}$$

There are n constants in $K_n(t)$ and its computation necessitates $[n/2]+2$ multiplications. The rational numbers A_{sn}, B_{sn} are stored; they should be computed to the degree of accuracy expected of the final result. Thus, for instance, in a floating-point double-precision computation of Arctan N, these constants are to be stored with eighteen correct significant digits.

Other forms, obtained in decomposing the algebraical fractions $P_m(t^2)/Q_m(t^2)$ into simple fractions of the variable $z = t^2$, could also be used. They are equivalent to (8) and (9) since they allow the computation of $K_n(t)$ in the same number $[n/2]+2$ of multiplications and involve the same number n of constants. All $[n/2]$ roots $z_i = -\omega_{in}$, $1 \leqq i \leqq [n/2]$, of the equation $Q_n(z) = 0$ are simple, real and negative because $Q_n(z)$ has positive coefficients. Depending on the parity of n, the degree of $P_n(z)$ is equal to or one unit less than the degree $[n/2]$ of $Q_n(z)$. Thus, we obtain

$$K_{2m}(t) = t \cdot \sum_{s=1}^{m} \xi_{sm}(t^2 + \omega_{i,2m})^{-1} \tag{10}$$

$$K_{2m+1}(t) = t \{ \eta_{om} + \sum_{s=1}^{m} \eta_{sm} \cdot (t^2 + \omega_{i,2m+1})^{-1} \} \tag{11}$$

with $\eta_{om} = (2m)!!/(2m+1)!!$ and ξ_{sm}, $\eta_{sm} = P_n(-\omega_{sn})/Q'_n(-\omega_{sn})$, taking $n = 2m$ for ξ_{sm}, and $n = 2m+1$ for η_{sm}. *Example:* If $m = 4$, then $P_4(z) = 105 + 55z$; $Q_4(z) = 105 + 90z + 9z^2$ and $\omega_{s4} = 5 \pm 2(10/3)^{\frac{1}{2}}$; $\xi_{s2} = 5[11 \pm 17 \cdot (3/10)^{\frac{1}{2}}]/18$; $s = 1, 2$. Thus

$$K_4(t) = t \{ \xi_{12}(t^2 + \omega_{14})^{-1} + \xi_{22}(t^2 + \omega_{24})^{-1} \} = \lambda_2 t \{ t^2 + B_{04} - \frac{A_{14}}{t^2 + B_{14}} \}^{-1} \tag{12}$$

with $\lambda_2 = 55/9$, $B_{04} = 89/11$, $A_{14} = 1372/363$ and $B_{14} = 21/11$. In the sequel we use (8) and (9), but not (10) or (11).

The case $m = 3$ is an exception. In this case the forms (9) and (11) coincide and $K_3(t) = t \cdot \{4/9 + (25/27) \cdot (t^2 + 5/3)^{-1}\}$. To compute $K_3(t)$ three multiplications are needed, but this number can be reduced to two, computing $K_3(t)$ as a function of the variable $\tau = 4t/9$:

$$(\tau = 4t/9) \qquad K^*_3(t) = \tau + a \cdot (\tau + b/\tau)^{-1} \tag{13}$$

with $a = 100/243$ and $b = 80/243$. One multiplication is gained replacing t by τ because for $N \subset I_k$ it is necessary to compute first the argument z_k of K_m involved in (4).

Now, in approximating Arctan z_k by $K^*_3(z_k)$ it is necessary to compute not z_k but $t_k = 4z_k/9$, using (4*) with $\lambda_3 = 9/4$. In saving a multiplication, the number of constants is increased using three instead of two in each interval I_k. If q is not large this increase of PC by $[(q-1)/2]$ units is not important. It could be applied when $Dg = 6$ correct digits only are required, because then $m = 3$ is combined with $q = 10$ and $[(q-1)/2] = 4$. For $Dg = 8$, $m = 3$ one has $q = 20$ so that, using K^*_3 instead of K_3, the number PC is increased by 9. If the value $m = 3$ is used for $Dg = 8$, 10, 18, 20 the computation of Arctan N is achieved in only three multiplications.

It is important that the same device can be applied to $K_{2m}(t)$ (but not to K_{2m+1}). Defining τ by $\tau = t/\lambda_m$ where $\lambda_m = [(2m)!!/(2m-1)!!]^2 - 1$, we obtain for (8) the following equivalent form:

$$K^*_{2m}(t) = \tau[\tau^2 + B^*_{o,2m} + \sum_{s=1}^{m-1} \frac{-A^*_{s,2m}|}{|\tau^2 + B^*_{s,2m}}]^{-1} \tag{14}$$

where
$B^*_{s,2m} = B_{s,2m} \cdot \lambda_m^{-2}$ and $A^*_{s,2m} = A_{s,2m} \cdot \lambda_m^{-4}$. If $N \subset I_o$, then $\tau = t_o = N \cdot \lambda_m^{-1}$, but if $N \subset I_k$ and $1 \leqq k \leqq [(q-1)/2]$, then $\tau = t_k = z_k/\lambda_m$ is computed by (4*).

In the last interval I_γ the relation

Arctan $N = \pi/2 -$ Arctan (N^{-1})

is to be used for q even. In such a case the convergents $K_n(N^{-1})$ could be computed as follows:

$$K_{2m}(N^{-1}) = N \cdot [N^2 + b_o + \sum_{s=1}^{m-1} \frac{-a_s|}{|N^2 + b_s}]^{-1} \tag{15}$$

$$K_{2m+1}(N^{-1}) = [1 + \sum_{s=1}^{m} \frac{-c_s|}{|N^2 + d_s}]/N \tag{16}$$

with $c_1 = 1/3$ and

$(s \geqq 1) \qquad a_s = s^2[1 + (16s^2 - 16s + 3)^{-1}]/(16s^2 - 1)$

$(s \geqq 0) \qquad b_s = [1 + (16s^2 + 8s - 3)^{-1}]/2$

$(s \geqq 2) \qquad c_s = (4s^2 - 6s + 2)^2/[(4s-5)(4s-3)^2(4s-1)]$

$(s \geqq 1) \qquad d_s = (2s-1)^2/(16s^2 - 16s + 3) + 4s^2/(16s^2 - 1)$.

The forms (15) and (16) necessitate $m+1$ and $m+2$ multiplications respectively, so that $K_n(N^{-1})$ is computable in $[n/2] + [3 - (-1)^n]/2$ multiplications. For instance, $[5/2]+2 = 4$ multiplications suffice in

$$K_5(N^{-1}) = [1 - \frac{1/3|}{|N^2 + 0.6} - \frac{12/175|}{|N^2 + 23/45}]/N.$$

Expressions (15) or (16) will not be used, forming instead $t_o = 1/N$ and applying (8) or (9).

• 5. *Choice of combinations*

It is always possible to reduce M by increasing PC. In order to choose the most economical combinations from among those listed in Table 2, we studied the reduced values of PC (and PC^*). The results of this study are condensed in Table 3 which gives also the numbers $M = [m/2] + 3$ and $M^* = M - 1$, M^* and PC^* denoting the values of M and PC, when K^*_3 or K^*_{2p} (m even, $m = 2p$) are applied instead of K_3 and K_{2p}. Thus, Table 3 gives the best combinations (m, q) involving smaller numbers M and PC for $Dg = 6$, 8, 10, 18 and 20.

Table 3 **Best combinations (m, q)**

Case	Value of Dg (Accuracy)	m	q	M	PC	M*	PC*
1	six	3	10	—	—	3	19
2	six	4	6	—	—	4	11
3	six	5	4	5	7	—	—
4	eight	4	9	—	—	4	17
5	eight	5	6	5	9	—	—
6	ten	5	9	5	14	—	—
7	ten	8	4	—	—	6	11
8	eighteen	8	12	—	—	6	18
9	eighteen	8	12	7	15	—	—
10	twenty	8	15	—	—	6	30
11	twenty	9	12	7	16	—	—

It is possible to compute Arctan N with $Dg=6$, 8, or 10 in three multiplications, using (13) for $q=10$, 20 and 45, respectively. Likewise $Dg=18$ or 20 is obtained, using (14) with $m=6$ and for $q=27$ or 45, respectively. Naturally, by increasing q it becomes necessary to store more and more constants.

• 6. *Examples of R-approximations*

Example 1

Initially, the combination (5, 10) which yields $Dg=10$ correct digits in five multiplications will be studied. The upper bound for PC gives $PC=27$, but this upper bound can be reduced to $PC=17$ as follows.

The upper bound $B_{5,10}=20(\frac{1}{2}\tan 9°)^{11}\approx 10^{-10.8}$ was computed for the first interval $0<N\leqq \tan 9°$, but the relative error is much smaller for $N\geqq 1$ and decreases, when N increases. This suggests the use of larger intervals for $N>1$. Instead of dividing the range $(0,\pi/2)$ of θ = Arctan N into $[q/2]+1=6$ intervals i_k as described above, the following five intervals: $I_0=(0;\ \tan 9°)$, I_1 = (tan 9°; tan 27°), I_2 = (tan 27°; 1), I_3 = (1; tan 67°.5) and I_4 = (tan 67°.5; ∞) will be used. The corresponding constants used in (4) for $1\leqq k\leqq 4$, are:

k	1	2	3	4
$\alpha_k=$	tan 72°	tan 54°	tan 33°.75	tan 11°.25
$\beta_k=$	Sec^2 72°	Sec^2 54°	Sec^2 33°.75	Sec^2 11°.25

It is seen that the use of equation (15) is avoided and five constants are thus saved. Decreasing the number of intervals one location constant is omitted. One more constant can be omitted among four location constants a_k: $\tan 9°=\sqrt{5}+1-\sqrt{5+(5)^{\frac{1}{2}}}$, $\tan 27°=\sqrt{5}-1-\sqrt{5-2(5)^{\frac{1}{2}}}$, tan 45° and $\tan 67°.5=\sqrt{2}+1$, namely $a_3=1$. The four constants needed in (3) are $n_1=\pi/10$, $n_2=2n_1$, $n_3=5\pi/16$ and $n_4=7\pi/16$. Again, it suffices to store n_1 only because $n_3=2n_1+n_1+n_1/8$ and $n_4=4n_1+n_1/4+n_1/8$. Therefore in all, 17 instead of 27 constants will be used. The five constants in the expression of $K_5(t)$

$$K_5(t)=t\left\{A-\frac{B|}{|t^2+C}-\frac{D|}{|t^2+E}\right\} \tag{17}$$

are: $A=64/225$; $B=1{,}309/675$; $C=8{,}743/2{,}805$; $D=551{,}124/874{,}225$ and $E=1{,}449/935$.

To facilitate the computation of eight constants α_k, $\beta_k=1+a_k^2$ the following exact expressions of α_k^2 are given: $\alpha_1^2=5+2\sqrt{5}$; $\alpha_2^2=1+0.4\sqrt{5}$; $\alpha_3^2=[(2-\sqrt{2})^{\frac{1}{2}}-1]\sqrt{2}+1$; $\alpha_4^2=[(2+\sqrt{2})^{\frac{1}{2}}-1]\sqrt{2}-1$.

It remains to prove that with our choice of I_3 and I_4 the relative error $R_{5,10}$ does not exceed $10^{-10.3}$. Since Arctan $N\geqq\pi/4$ if $N\geqq 1$, it is sufficient to check $R_{5,10}$ at the left end of I_3, where the value of $|t_3|$ is $T=\tan 11°.25$, $N=1$ and Arctan $N=\pi/4$. The absolute error $E_m(t)$ verifies the inequality

$$E_m(t)\leqq E_m(T)\leqq [m!/(2m-1)!!]^2 T^{2m+1}/(2m+1).$$

Therefore Log $E_5(T)\leqq -10.54863$ and Log $R_5=$ Log $(4E_5/\pi)\leqq -10.44372$ or $R_{5,10}\leqq 3.6\times 10^{-11}$. A direct computation of R_5 shows indeed that $R_{5,10}=3.175\times 10^{-11}$.

Therefore the combination (5, 10) yields $Dg=10$ correct digits in $M=5$ multiplications the number PC of stored constants being equal to 17 and $K_5(z_k)$ being computed by (9).

Example 2

Sometimes a single multi-precision subroutine is desirable which allows the computation of Arctan N with the first 6, 8, 10, 18 or 20 correct significant digits. Such routines are possible as will be shown in this example, in which $q=9$. To locate N in one of five intervals

$$I_0=[0;\ \tan(\pi/18)],\quad I_k=[\tan((2k-1)\,\pi/18);\ \tan((2k+1)\,\pi/18)]$$

four constants $a_k=\tan[(2k-1)\,\pi/18]$, $1\leqq k\leqq 4$ are needed. Using (3) with $n_k=k\pi/9$, only one of these four constants, namely $n_2=2\pi/9$ can be stored because $n_1=n_2/2$, $n_3=n_2+n_2/2$ and $n_4=2n_2$. Moreover, in (4) there are four more constants α_k, $\beta_k=1+\alpha^2_k$, since $\alpha_k=a_{5-k}$, $1\leqq k\leqq 4$.

To these nine constants are added the constants involved in $K_m(t)$. Figure 1 shows that for $q=9$ the value $m=Dg/2$ insures exactly Dg correct digits. Thus, it is necessary to use five different convergents K_m involving 3, 4, 5, 9 and 10 constants since the number of constants in a K_m is equal to $m=Dg/2$. Adding these 31 constants to 9 a total of $PC=40$ precomputed and stored constants are obtained. This is not much for a subroutine which allows floating and fixed point, single and double precision computations. The number of multiplications M is equal to $[m/2]+3=[Dg/4]+3$, that is to 4, 5, 5, 7 and 8 for $Dg=6$, 8, 10, 18 and 20, respectively.

The explicit expressions (8) and (9) of $K_m(t)$ for $m=3$, 4 and 5 were already given in (12), (13) and (17). To obtain those for $m=9$ and $m=10$ it is sufficient to apply Euclid's algorithm to the quotients

$$P_9\cdot Q_9^{-1}=\Big(\sum_{s=0}^{4} p_{s,9}\cdot z^s\Big)\Big(\sum_{s=0}^{4} q_{s,9}\cdot z^s\Big)^{-1} \qquad \text{where } (z=t^2)$$

and

$$P_{10}\cdot Q_{10}^{-1}=\Big(\sum_{s=0}^{4} p_{s,10}\cdot z^s\Big)\Big(\sum_{s=0}^{5} q_{s,10}\cdot z^s\Big)^{-1},$$

using the numerical values of coefficients p's and q's (see Table 1). Another way would consist in retransforming the expressions (8) and (9) for $m=9$ and 10 into quotients of polynomials and solving the equations for their coefficients A's and B's obtained by identifying these quotients to $tP_m(t^2)\cdot Q_m^{-1}(t^2)$; $m=9$; 10.

• 7. *Detailed description of best combinations*

The eleven best combinations listed in Table 3 will now be described in order to facilitate their application.

Case 1: $m=3$, $q=10$, $M^*=3$, $PC^*=19$, $Dg=6$.

The same intervals and constants a_1 a_2, a_4, n_1 defined in Example 1 are used here. Since (13) is used, to form $\tau_1=4N/9$ when $N\subset I_1$ the constant 4/9 is stored. In the four intervals I_k, $1\leqq k\leqq 4$, twelve constants are needed: $\gamma_k=\text{cotan}\,\theta_k$ with $\theta_1=\pi/10$, $\theta_2=\pi/5$, $\theta_3=56°15'$, $\theta_4=78°45'$; $\alpha^*_k=4\gamma_k/9$ and $\beta^*_k=4(1+\gamma_k^2)/9$. Finally, in (13) two constants are used. In all there are nineteen stored constants.

Case 2: $m=4$, $q=6$, $M^*=4$, $PC^*=11$, $Dg=6$.

Four intervals: 0° — 15° — 45° — 75° — 90°. Since $a_1=2-\sqrt{3}$ and $a_2=2+\sqrt{3}$, only one location constant, $\sqrt{3}$, is stored. Storing $n_2=\pi/3$, we have $n_1=n_2/2$ and $\pi/2=n_2+n_2/2$, so that it suffices to store n_2. Using (5) and the form (14), we have to form N/λ_2, if $N\subset I_1$, and λ_2/N, if $N\subset I_4$, so that both $\lambda_2=55/9$ and $\lambda_2^{-1}=9/55$ are stored. In I_2 and I_3 the six constants $\gamma_1=\sqrt{3}$, $\gamma_2=\sqrt{3}/3$, $\alpha^*_1=9\sqrt{3}/55$, $\alpha^*_2=3\sqrt{3}/55$, $\beta^*_1=4.\lambda_2^{-1}$ and $\beta^*_2=4\lambda_2^{-1}/3$ are used. There are four to be stored: γ_2, α^*_1, α^*_2 and β^*_2. Finally, in (14) there are three constants to store: $B^*_{04}=B_{04}\cdot(9/55)^2=7{,}209/33{,}275$; $B^*_{14}=B_{14}\cdot(9/55)^2=1{,}701/33{,}275$ and $A^*_{14}=A_{14}\times(9/55)^4=3{,}000{,}564/1{,}107{,}225{,}625$.

Case 3: $m=5$, $q=4$, $M=5$, $PC=7$, $Dg=6$.

Three intervals: $i_1=(0;\ 22°30')$, $i_2=(22°\ 30';\ 67°30')$ and $i_3=(67°30';90°)$. Location constants: $a_1=\tan 22°.5=\sqrt{2}-1$ and $a_2=\tan 67°.5=\sqrt{2}+1$, to store $\sqrt{2}$, $n_1=\pi/4$ so that $\pi/2=2n_1$; in (4), applied in I_2 only, one has $\alpha_1=\text{Cotan}\,45°=1$ and $\beta_1=2$—nothing to store. Finally, there are five constants in K_5 (see (9)). In all, seven constants are to be stored.

Case 4: $m=4$, $q=9$, $M^*=4$, $PC^*=17$, $Dg=8$.

Five intervals: 0° — 10° — 30° — 50° — 70° — 90°; four location constants $a_k=\tan\,(20°\cdot k-10°)$; among the four constants $n_k=k\pi/9$ only one, n_2, to store since $n_1=n_2/2$, $n_3=n_2+n_2/2$ and $n_4=2n_2$; in (4*) $\gamma_k=\text{Cotan}\,(k\pi/9)=a_{5-k}$ are already stored as a_k, but $\alpha^*_k=\gamma_k\cdot\lambda_2^{-1}=9\gamma_k/55$ and $\beta^*_k=9(1+\gamma_k^2)/55$, $1\leqq k\leqq 4$, are stored as well as $\lambda_2^{-1}=9/55$; adding finally the three constants involved in K^*_4 (see Case 2), gives a total of 17 constants.

Case 5: $m=5$, $q=6$, $M=5$, $PC=9$, $Dg=8$.

Four intervals: 0° — 15° — 45° — 75° — 90° and only one location constant, namely $\sqrt{3}$, since $\tan 45°=1$ while $\tan 15°=2-\sqrt{3}$, $\tan 75°=2+\sqrt{3}$; to store also $n_1=\pi/6$ while $n_2=2n_1$ and $\pi/2=2n_1+n_1$; $\alpha_1=\text{Cotan}\,30°=\sqrt{3}$, $\alpha_2=\text{Cotan}\,60°=\sqrt{3}/3$, $\beta_1=4$, $\beta_2=4/3$ and thus it suffices to store α_2 and β_2; finally in K_5 (see (9)) there are five constants. In all, there are nine constants to store.

Case 6: $m=5$, $q=9$, $M=5$, $PC=14$, $Dg=10$.

Same intervals and same constants a_k, $1\leqq k\leqq 4$, and n_2 to store as in Case 4; in (4) $\alpha_k=a_{5-k}$, thus only four $\beta_k=1+\alpha_k^2$ to store; adding to these nine constants five involved in K_5, a total of 14 constants are obtained.

Case 7: $m=8$, $q=4$, $M^*=6$, $PC^*=11$, $Dg=10$.

Same intervals and same two constants $\sqrt{2}$, $n_1=\pi/4$ to store as in Case 3; but since (4*) is applied it is also required to store $\lambda_4=[(8!!/7!!)^2-1]=15{,}159/1{,}225$ and $\lambda_4^{-1}=1{,}225/15{,}159$; now $\gamma_1=1$, $\alpha^*_1=\lambda_4^{-1}\cdot\gamma_1=\lambda_4^{-1}$ and $\beta^*_1=\lambda_4^{-1}(1+\gamma_1^2)=2\lambda_4^{-1}$ so that there is nothing to store using (4*) in I_2; finally in K^*_8 there are seven constants $B^*_{s8}=B_{s8}\cdot\lambda_4^{-2}$, $0\leqq s\leqq 3$ and $A^*_{s8}=A_{s8}\cdot\lambda_4^{-4}$, $1\leqq s\leqq 3$. The coefficients A_{s8}, B_{s8} are deducible from the expressions of $P_8/9$ and $Q_8/9$ where $p_{08}/9=q_{08}/9=15!!/9$ should be used since Table 1 gives $p_{s8}/9$ and $q_{s8}/9$.

Case 8: $m=8$, $q=12$, $M^*=6$, $PC^*=18$, $Dg=18$.

Seven intervals: 0 — 7°.5 — 22°.5 — 37°.5 — 52°.5 — 67°.5 — 82°.5 — 90°, so that $\pi/2$ is necessary, as well as λ_4 and λ_4^{-1}, to form $t_7=\lambda_4/N$, if $N\subset I_7$, and $t_1=N/\lambda_4$, if $N\subset I_1$. Among $n_k=k\pi/12$, $1\leqq k\leqq 5$, it suffices to store $n_2=\pi/6$ since $n_1=n_2/2$, $n_3=n_2+n_2/2$, $n_4=2n_2$, $n_5=2n_2+n_2/2$ and $\pi/2=2n_2+n_2$. There are six location constants $a_k=\tan\,[(2k-1)\pi/24]$, $1\leqq k\leqq 6$ and they can be expressed in terms of four: $\sqrt{2}$, $\sqrt{3}$, $p=2\sqrt{2+\sqrt{3}}$ and $q=2\sqrt{2-\sqrt{3}}$ since $a_1=p-2-\sqrt{3}$, $a_2=\sqrt{2}-1$, $a_3=q-2+\sqrt{3}$, $a_4=q+2-\sqrt{3}$, $a_5=\sqrt{2}+1$ and $a_6=p+2+\sqrt{3}$. In five interior intervals 15 constants α^*_k, β^*_k, γ_k, $1\leqq k\leqq 5$ are used, but only four among them are to be stored: $\gamma_4=1/\sqrt{3}$, $\alpha^*_2=\sqrt{3}/\lambda_4$, $\alpha^*_4=\lambda_4^{-1}/\sqrt{3}$ and $\beta^*_4=4\lambda_4^{-1}/3$, because the eleven others are as follows: $\gamma_1=2+\sqrt{3}$, $\gamma_2=\sqrt{3}$, $\gamma_3=1$, $\gamma_5=2-\sqrt{3}$, $\alpha^*_1=2\lambda_4^{-1}+\alpha^*_2$, $\alpha^*_3=\lambda_4^{-1}$, $\alpha^*_5=2\lambda_4^{-1}-\alpha^*_2$, $\beta^*_1=8\lambda_4^{-1}+4\alpha^*_2$, $\beta^*_2=4\lambda_4^{-1}$, $\beta^*_3=2\lambda_4^{-1}$ and $\beta^*_5=8\lambda_4^{-1}-4\alpha^*_2$. Finally, there are seven constants involved in K^*_8, so that the total number of stored constants amounts to 18.

Case 9: $m=8$, $q=12$, $M=7$, $PC=15$, $Dg=18$.

If the form (8) of K_8 is used instead of the form (14) of K^*_8, there are eight constants in K_8, λ_4 included, so that λ_4^{-1} is no longer necessary and instead of α^*_k, β^*_k, $\alpha_k=\gamma_k$ and $\beta_k=1+\gamma_k^2$ will be used in the interior intervals, storing only $\gamma_4=1/\sqrt{3}$ and $\beta_4=4/3$ since $\beta_1=8+4\sqrt{3}$, $\beta_2=4$, $\beta_3=2$ and $\beta_5=8-4\sqrt{3}$. Thus, three constants are saved in comparison to Case 8 and $PC=15$.

Case 10: $m=8$, $q=15$, $M^*=6$, $PC^*=30$, $Dg=20$.

Here there are eight intervals, 0° — 6° — 18° — 30° — 42° — 54° — 66° — 78° — 90° and seven location constants $a_k=\tan\,[(2k-1)\,\pi/30]$ to store, $1\leqq k\leqq 7$. Among $n_k=k\pi/15$ only $n_2=2\pi/15$ is to be stored: $n_1=n_2/2$, $n_3=n_2+n_2/2$, $n_4=2n_2$, $n_5=2n_2+n_2/2$, $n_6=2n_2+n_2$, $n_7=4n_2-n_2/2$. Among 21 constants γ_k, α^*_k, β^*_k the constants $\gamma_k=a_{8-k}$ are already stored, so that fourteen constants $\alpha^*_k=\gamma_k/\lambda_4$,

$\beta^*_k=(1+\gamma_k^2)/\lambda_4$ will be stored. One also needs λ_4^{-1} and seven constants in K^*_8. The total is therefore: $PC=30$.

Case 11: $m=9$, $q=12$, $M=7$, $PC=16$, $Dg=20$.

Same intervals and same constants $\sqrt{2}$, $\sqrt{3}$, p, q, n_2 as in Case 8 except that α^*_k, β^*_k are not needed now, but only $\alpha_k=\gamma_k$ and $\beta_k=1+\gamma_k^2$, $1\leqq k\leqq 5$. It suffices to store $\alpha_4=\sqrt{3}/3$ and $\beta_4=4/3$. Adding to these seven constants the nine constants involved in K_9 the total of 16 constants is obtained. Thus twenty correct significant digits can be obtained in seven multiplications using in the subroutine only sixteen constants.

Polynomial approximations

• *8. Study of the relative error R_n.*

The Tchebychev polynomial $T_{2m+1}(x)$ verifies in $-1\leqq x\leqq 1$ not only the inequality $|T_{2m+1}(x)|\leqq 1$, but also

$$|T_{2m+1}(x)|\leqq(2m+1)\cdot|x|. \qquad (18)$$

This inequality gives an upper bound for the relative error R_n made in approximating Arctan N, $N=x\cdot\tan 2\theta$, by the polynomial P_{n-1} of degree $2n-1$

$$P_{n-1}=2\sum_{m=o}^{n-1}(-1)^m\cdot\tan^{2m+1}\theta\cdot\ T_{2m+1}(x)/(2m+1). \qquad (|x|\leqq 1)\quad (19)$$

With the aid of (18)

$$2\left|\sum_{m=n}^{\infty}(-1)^m\tan^{2m+1}\theta\cdot T_{2m+1}(x)/(2m+1)\right|\leqq|x|\cdot\tan\ 2\theta\cdot(\tan\ \theta)^{2n}$$

so that

$$|R_n|\leqq|x|\cdot\tan\ 2\theta(\tan\ \theta)^{2n}/N=\tan^{2n}\theta.$$

Subdividing the range $(0,\infty)$ of N into intervals as explained in Section 2, we choose $\theta=\pi/4q$ so that the order of magnitude of $[\tan(\pi/4q)]^{2n}$ depends on the two parameters n and q.

To insure an accuracy characterized by first Dg correct significant digits the integers n and q should be chosen so as to verify the condition

$$2n\cdot|\text{Log}\tan(\pi/4q)|>Dg+0.3.$$

As for R-approximations there are many combinations (n, q) verifying this condition for the same value of Dg. In them the number $M=n+1$ does not depend on q, but the number PC is a function of both parameters n and q.

Omitting the details of a long comparative study of all possible combinations (n, q) for various values of Dg (it is quite similar to the study of combinations (m, q) for R-approximations), we will simply state the final results obtained for $Dg=6$, 8, 10, 18 and 20.

The eleven best cases listed in Table 4 were retained. In them $M=n+1$ and $PC=n+2[q/2]$, $3\leqq n\leqq 9$ while the parameter q takes four values only; $q=5$, 6, 9 and 12:

Table 4 **Best combinations (n, p)**

$q=12$					$q=9$				
Case	*Dg*	*n*	*M*	*PC*	*Case*	*Dg*	*n*	*M*	*PC*
1	6	3	4	9	6	6	3	4	11
2	8	4	5	10	7	8	4	5	12
3	10	5	6	11	8	10	5	6	13
4	18	8	9	14	$q=6$				
5	20	9	10	15	10	8	5	6	8
$q=5$					11	10	6	7	9
9	6	4	5	7					

Substituting in (19) the explicit expression of $2(-1)^m T_{2m+1}(x)/(2m+1)$, namely

$$2(-1)^m T_{2m+1}(x)/(2m+1)=\sum_{s=o}^{m}(-1)^s\binom{m+s}{m-s}(2x)^{2s+1}/(2s+1),$$

grouping together the like terms and replacing $2x$ by $t\cdot(1-\tan^2\theta)$. Cotan θ, $P_{n-1}(t)$ takes the following form

$$P_{n-1}(t)=\sum_{s=0}^{n-1}(-1)^s A_{ns}\cdot t^{2s+1}/(2s+1) \qquad (20)$$

where

$$A_{ns}=(1-\tan^2\theta)^{2s+1}\cdot\sum_{j=0}^{n-s-1}\binom{2s+j}{j}\cdot\tan^{2j}\theta.$$

In particular for $s=0$:

$$A_{no}=(1-\tan^2\theta)\sum_{j=0}^{n-1}\tan^{2j}\theta=1-\tan^{2n}\theta,$$

which shows that the value of A_{no} can be rounded off to one rejecting $\tan^{2n}\theta$. It can be neglected because $|R_n|\leqq\tan^{2n}\theta$ and the first term of our approximation is $N\cdot A_{no}$. Thus, $A_{no}\approx 1$ need not to be stored and the polynomial P_{n-1} has $n-1$ coefficients to store.

Here $t=x\cdot\tan 2\theta$ is equal to N, if $N\subset I_o$, but if $N\subset I_k$ then $t=z_k$ is computed by (4). To illustrate this transformation of P_{n-1}, consider Case 1: $n=3$, $q=12$ and $\theta=3°.75$. Now one has $s=0$, 1, 2 and

$$A_{3s}=(1-\tan^2\theta)^{2s+1}\cdot\sum_{j=o}^{2-s}\binom{2s+j}{j}\cdot\tan^{2j}\theta$$

so that $A_{30}=1-\tan^6\theta$; $A_{31}=(1+3\tan^2\theta)(1-\tan^2\theta)^3$ and $A_{32}=(1-\tan^2\theta)^5$. Since $\tan\theta=\tan(\pi/48)=0.065\ 543\ 4628\ldots$, it is found that $A_{30}=1-793\times 10^{-10}$; $A_{31}=0.999\ 889\ 90136\ldots$; $A_{32}=0.978\ 704\ 0328\ldots$. Thus, for $N\leqq\tan 7°.5$, we obtain the approximation:

$$\text{Arctan } N\approx N[d_0-N^2(d_1-d_2\cdot N^2)] \qquad (21)$$

with $d_0=A_{30}=0.999\ 999\ 9207$, $d_1=A_{31}/3=0.333\ 296\ 6338$ and $d_2=A_{32}/5=0.195\ 740\ 8066$. Applying (21) to $N^*=\tan 7°.5=0.131\ 652\ 497\ldots$ one should obtain first six correct digits in the true value of Arctan $N^*=\pi/24=0.130\ 899\ 6938\ldots$. Computing the right hand member of (21) for $N=N^*$ we find much better approximation, namely 0.130 899 6948 . ., so that the relative error is equal to 7.64×10^{-9} and eight digits are correct instead of six. The reason for it is simple: the upper bound $\tan^{2n}\theta$ of the relative error was obtained with the aid of (18) and this inequality greatly exaggerates the true value of $T_{2m+1}(x)$ for $x \approx 1$, though for small x it gives a very reasonable estimate.

This suggests that for small values of N no more than six correct digits can be obtained. Indeed, applying (21) to $N = \tan 0°.5 = 008\ 726\ 8678\ldots$, one expects the value of $\pi/360 = 0.008\ 726\ 6462\ldots$ and the right-hand member of (21) yields the number 0.008 726 6450 with first six correct significant digits, the relative error being equal to $12 \times 360 \times 10^{-10}/\pi = 1.4 \times 10^{-7}$.

• 9. *Description of eleven cases*

The location constants a_k, as well as the constants n_k, α_k and β_k depend only on q and therefore their number and values are the same both for R- and P-approximations, provided the value of q is the same. Therefore, in describing the cases, except Case 9, of P-approximations it is sufficient to refer to the corresponding cases of Section 7 to define all the constants except the $n-1$ coefficients of $P_{n-1}(t)$ (the first is always equal to one).

Case 1

See Case 9, Section 7. To the seven constants of Case 9 which are not coefficients of K_8 are added rounded-off d_{21} and d_{22} of the example ($d_{20} \approx 1$):

$d_{21} \approx 0.333\ 2966$; $d_{22} \approx 0.195\ 7408$

Cases 2–5

In these cases, the seven constants a_k, n_k, α_k, β_k are the same as in Case 1, since the value of $q = 12$ does not change. The $n-1$ coefficients $d_{n-1,j}$ of $P_{n-1}(t)$, $1 \leqq j \leqq n-1$ are added to the constants in each case.

Cases 6–8

Since $q = 9$, there are the same nine constants a_k, n_2, β_k as in Case 6, Section 7. Adding to the constants $d_{n-1,j}$; $1 \leqq j \leqq n-1$; for $n = 3$, 4 and 5, one obtains $PC = 11$, 12 and 13 of Table 4.

Case 9

Here $q = 5$, $n = 4$, and $2\theta = 18°$. There are three intervals $0°$ — $18°$ — $54°$ — $90°$ and four constants to store: $a_1 = \tan 18° = (1 - 0.4\sqrt{5})^{\frac{1}{2}}$; $a_2 = \tan 54° = (1 + 0.4\sqrt{5})^{\frac{1}{2}}$; $n_1 = \pi/5$ and $0.4\sqrt{5}$ since $n_2 = 2n_1$, $\alpha_1 = \cot 36° = a_2$, $\alpha_2 = \cot 72° = a_1$, $\beta_1 = 2 + 0.4\sqrt{5}$ and $\beta_2 = 2 - 0.4\sqrt{5}$. Adding to these four constants the coefficients d_{31}, d_{32} and d_{33} a total of seven stored constants is obtained. Thus, six correct digits are obtained in five multiplications, if $PC = 7$.

Cases 10 and 11

These cases correspond to Case 5, Section 7. Four stored constants: $\sqrt{3}$, $\pi/6$, $\sqrt{3}/3$ and $4/3$. Adding to them four, if $n = 5$, and five, if $n = 6$, coefficients of $P_{n-1}(t)$ one has $PC = 8$ and 9 as in Table 4.

If it is desired to decrease by one the values $M = 9$ and 10 necessary for obtaining $Dg = 18$ and 20, it will be necessary to increase the number PC of stored constants. Choosing $q = 15$ (see Case 10 of Table 3) one can use $2\theta = 6°$, so that the upper bound $2n \cdot \text{Log} \tan 3°$ of the logarithm of relative error is equal to $-17.92 < -17.3$ and $-20.49 < -20.3$ for $n = 7$ and $n = 8$, respectively. Therefore, $Dg = 17$ and $Dg = 20$ if $q = 15$, $n = 7$ and 8, respectively. This gives the two cases listed in the abstract, when $Dg = 17$ and 20 are obtained in eight and nine multiplications, the number of precomputed constants being equal to 21 and 22, respectively.

• 10. *Conclusion*

To compare our results with known approximations to Arctan x the following three formulae are chosen which seem to be the best among the known approximations which involve no tables of values of Arctan x stored in the subroutine:

(L)* $\quad \text{Arctan}\ x/x \approx (1 + a_2x^2 + a_4x^4 + a_6x^6)/(1 + a_3x^2 + a_5x^4 + a_7x^6 + a_9x^8)$

(M)† $\quad \text{Arctan}\ x/x \approx (b_0 + b_2x^2 + b_4x^4 + b_6x^6)/(1 + b_1x^2 + b_3x^4 + b_5x^6)$

(H)‡ $\quad \text{Arctan}\ x/x \approx c_0 - c_1x^2 + c_2x^4 - c_3x^6 + c_4x^8 - c_5x^{10} + c_6x^{12} - c_7x^{14}$

where $a_2 = 5/3$; $a_3 = 2$; $a_4 = 47/60$; $a_5 = 5/4$; $a_6 = 19/210$; $a_7 = 1/4$; $a_9 = 1/128$; $b_0 = 1 - 19 \times 10^{-10}$; $b_1 = 1.45356\ 71346$; $b_2 = 1.12023\ 40143$; $b_3 = 0.56503\ 09796$; $b_4 = 0.28050\ 45407$; $b_5 = 0.04901\ 75912$; $b_6 = 0.00856\ 11889$; $c_0 = 0.99999\ 93329$; $c_1 = 0.33329\ 85605$; $c_2 = 0.19946\ 53599$; $c_3 = 13908\ 53351$; $c_4 = 0.09642\ 00441$; $c_5 = 0.05590\ 98861$; $c_6 = 0.02186\ 12288$; $c_7 = 0.00405\ 40580$.

The formula (H) will be compared to our P-approximations. Those (L) and (M) necessitate, in the form in which they are given by their authors, eight multiplications, but replacing them by the equivalent continued fractions it is possible to reduce the number of multiplications. The upper bound of errors in (L) is not mentioned by Dr. C. Lanczos. We computed it for real x and found $|x|^9/8{,}000$ for small $|x|$ and 1.4×10^{-5} for $|x| = 1$ insofar as absolute error is concerned. It belongs to the same type as our R-approximation and could be used in a reduced range only. Similar to our K_8 it could give Arctan x in seven multiplications, but it has an insufficient accuracy: for $x = 1$ only the first four digits are correct, while Case 7 gives ten correct digits in six multiplications against seven necessitated by (L). For $x = 0.1$ Lanczos' method (L) gives six correct digits, while our Case 6 yields ten in five multiplications only.

The approximation (M) is much better: its range of validity is $0 \leqq x \leqq 1$ with the same upper bound 6.10^{-10} for the absolute error in the whole range. It gives eight correct digits and for many values of x even nine. Thus, for $x = 0.057$, the correct value of Arctan 0.057 is 0.056 938 389 06 and the formula (M) gives 0.056 938 388 98 . . so that the absolute and relative errors are equal to 8×10^{-11} and 1.4×10^{-9}, respectively, and $Dg = 8$. For $x = 0.1$, Arctan $0.1 = 0.099\ 668\ 652\ 49$ and (M) yields the approximation 0.099 668 652 52, so that again $Dg = 8$. Our Case 6, ($m = 5$; $q = 12$), gives 0.099 668 652 49 so that $Dg = 10$ in five multiplications.

*C. Lanczos. *Applied Analysis*, p. 492. Prentice Hall, 1956.
†Dr. Hans J. Maehly, Institute for Advanced Study, Princeton, N. J.
‡C. Hastings. *Approximations for Digital Computers*, p. 137. Princeton Univ. Press, 1955.

It is possible to give to (M) another form which involves also only five multiplications. For $0 \leqq x \leqq 1$ it is:

$$(0 \leqq x \leqq 1) \quad \text{Arctan } x = x. \left\{ B_0 + \frac{A_1|}{|x^2+B_1} - \frac{A_2|}{|x^2+B_2} - \frac{A_3|}{|x^2+B_3} \right\} \quad (22)$$

with

$B_0 = 0.17465\ 54388$; $A_1 = 3.709\ 256\ 262$;

$B_1 = 6.762\ 139\ 240$; $A_2 = 7.106\ 760\ 045$;

$B_2 = 3.316\ 335\ 425$; $A_3 = 0.264\ 768\ 6202$.

$B_3 = 1.448\ 631\ 538$.

Since the form (22) holds only for $0 \leqq x \leqq 1$, the values of $x \geqq 1$ necessitate another form equivalent to (M), namely

$$(x \geqq 1) \quad \text{Arctan } x = \pi/2 - \left(B^*_0 - \frac{A^*_1|}{|x^2+B^*_1} - \frac{A^*_2|}{|x^2+B^*_2} - \frac{A^*_3|}{|x^2+B^*_3} \right) / x \quad (23)$$

with

$B^*_0 = 0.999\ 999\ 9981$; $A^*_1 = 0.333\ 333\ 1177$;

$B^*_1 = 0.59998\ 72689$; $A^*_2 = 0.06847\ 53582$;

$B^*_2 = 0.50597\ 40184$; $A^*_3 = 0.05451\ 02420$.

$B^*_3 = 0.34760\ 58473$.

We transformed (M) into the forms (22) and (23) in order to save three multiplications. Using them it is possible to compute Arctan N for $0 \leqq N < \infty$ in five multiplications, the number of stored constants being equal to $PC = 14$. Since no subdivisions of the ranges (0; 1) and (1, ∞) are involved, the logical part of the corresponding program is very short, which also saves time.

The book of C. Hastings contains six P-approximations to Arctan x in the range (0; 1) (sheets 8–13, pp. 132–137). Their accuracy and number of operations and of stored constants are:

Sheet	*8*	*9*	*10*	*11*	*12*	*13*
Dg	2	3	3	4	5	6
M	4	5	6	7	8	9
PC	3	4	5	6	7	8

We consider only the last one with $Dg = 6$ (see formula (H)). This approximation belongs to the same type as (M) and it holds in the interval (0; 1). Here are some numerical results. For $x = 0.1$ formula (H) yields 0.099 668 615 . . so that the relative error is 3.7×10^{-7}. For $x = 1$ it gives 0.785 398 126 which corresponds to a relative error 4.7×10^{-8}, the first seven digits being correct.

Comparing now (H) with our Cases 1, 6 and 9 since they have the same accuracy of $Dg = 6$:

	(H)	*Case 1*	*Case 6*	*Case 9*
Number of multiplications	9	4	4	5
Number of stored constants	8	9	11	7

It is to be noted that in nine multiplications our Case 4 yields 18 correct digits instead of six, using six more constants ($PC = 14$) than in (H).

Received July 10, 1957