ON THE MONOTONICITY OF SOME COMPUTED FUNCTIONS

W. Kahan

Mathematics Dept. and
Elect. Eng. and Computer Science Dept.,
University of California at Berkeley.

Mar. 21, 1985

Abstract:
Techniques are introduced to help decide whether roundoff errors will abrogate the monotonicity properties of a function when it is computed. Those techniques are applied to several expressions, among them
$z/(1+z)$, $z + z/(1+z)$, $2y - y^2$, $w + 1/w$, $t - t/(1+4/t^2)$, ... , that have turned up during the calculation of certain elementary transcendental functions. Within appropriate ranges of their variables, the expressions are shown to remain monotonic when computed in binary arithmetic rounded as specified in proposed IEEE standards p754 and p854 . Because these standards are being adopted so widely, the conclusions from this error-analysis will be appreciated widely enough to justify its propagation.

ON THE MONOTONICITY OF SOME COMPUTED FUNCTIONS

W. Kahan

Introduction

A program F(x) intended to compute some elementary function f(x) might well be expected to do so fairly accurately. And then should f(x) be monotonic, say f(x) ≤ f(y) whenever x < y , contrary accidents F(x) > F(y) will occasion surprise. In the face of roundoff such aberrations are hard to extirpate. Fortunately some elementary functions *can* be calculated accurately, economically and monotonically despite roundoff provided arithmetic is carried out carefully enough. Specifically, when computed in *Binary* arithmetic rounded according to the proposed IEEE standards p754 and p854 **(¹)**, the following expressions will be proved monotonic:

z/(1+z) for -1 ≤ z ≤ 0 and 1/2 ≤ z ≤ 1 ,
z + z/(1+z) for -1 ≤ z < ∞ ,
(z-1)/(z+1) for |z| ≤ 3 ,
2y - y² for -∞ ≤ y ≤ 1/2 ,
w + 1/w for 1/2 ≤ |w| < 1/√2 and 1 ≤ |w| ,
t - t/(1+4/t²) for |t| ≤ 2/(1+√2) , and
3/4 + ((1-2q) + q/4)/(4+q) for 4/15 < q < 5/7 .

These deserve attention because they arise during the calculation of diverse elementary transcendental functions **(²)**. The ranges of values are significant too; for instance, roundoff destroys the monotonicity of w + 1/w at many values w between 0 and 1 , so this expression has to be avoided as a means of computing sin(θ) = 2/(w + 1/w) from w = tan(θ/2) .

Notation:

I, J, K, L, M, N, i, j, k, l, m, n are integers.
N = number of significant bits carried during arithmetic;
N = 24 for IEEE single, 53 for double, ...
u, v, w, x, y, z are real variables.
[x] = x rounded to N significant bits (explained below).
1ulp(x) = one *u*nit in the *l*ast *p*lace of [x]
= 2^{k+1-N} if $2^k \le |x| < 2^{k+1}$, normally.
x′ = Nextafter(x, +∞) if x is representable exactly,
= x + 1ulp(x) if x > 0 too.

Every number x representable exactly as a floating-point number with N significant bits has a value $x = \pm 2^n M$ where $0 \le M < 2^N$. Except when x = 0 or when x is *Subnormal* (n has the minimum value allowed for the chosen floating-point format), the *Normal* representation of x is determined uniquely by $2^{N-1} \le M \le 2^N-1$. And [x] = x . But when x is not representable exactly, when [x] ≠ x , then [x] is an exactly representable value closest to x and so |[x] - x| ≤ 1ulp(x)/2 . To fix [x] uniquely when x lies just midway between two exactly representable values, a tie-breaking rule must be invoked. The IEEE standard breaks the tie by "rounding to nearest even"; this means that $x = 2^n(M + 1/2)$, with $2^{N-1} \le M \le 2^N-1$, rounds to $[x] = 2^n M$ when M is even, to $[x] = 2^n(M+1)$ when M is odd. (This tie-breaking rule avoids the statistical bias inherent in another rule, widely used by earlier computers like the DEC VAX **(³)**, that always rounds midway cases up in magnitude to $[x] = 2^n(M+1)$.) Whether monotonicity might be affected by the choice of tie-breaking rule, or by the choice of radix (we have chosen binary, radix 2 , instead of decimal, radix 10), are interesting questions not to be answered here.

Easy Decisions:
Obviously $x < y$ implies $[x] \leq [y]$, so $[z]$ is a monotonic function of z. It soon follows that examples like $[1/[1+[1/z]]]$ and $[z/[\sqrt{[1-[z^2]]}]]$ are monotonic functions of z too, though ostensibly algebraically equivalent expressions $[z/[1+z]]$ and $[z/[\sqrt{[[1-z][1+z]]}]]$ respectively fail to be monotonic at a host of valid values z despite that the latter two expressions may be more accurate in the face of roundoff. The connection between monotonicity and accuracy is weak, but valuable none the less as an important consequence of the discrete nature of representable numbers; it amounts to this:

Suppose that $f(x)$ is a monotone function of x; for the sake of definiteness suppose $f(x)$ is increasing. And suppose $F(x)$ is an approximation to $f(x)$. Finally, suppose the uncertainty in $F(x)$ is a known function $\varepsilon(x) \geq |F(x)-f(x)|$. We do not assume that either $F(x)$ or $\varepsilon(x)$ is representable; their provenance is irrelevant. How small must $\varepsilon(x)$ be to imply that $F(x)$ is a monotonic function when its argument x is restricted to representable values? If, for any consecutive representable numbers x and x', we find $\varepsilon(x)+\varepsilon(x') \leq f(x')-f(x)$, then
$$F(x')-F(x) \geq (f(x')-\varepsilon(x')) - (f(x)+\varepsilon(x)) \geq 0 ,$$
so the monotonicity of f will not be violated by F nor by its rounded value $[F]$. In effect, so long as f increases fast enough compared with the uncertainty ε in its approximation F, then F will be nondecreasing too.

Values that differ by less than 1ulp round to values that differ by no more than 1ulp; i.e., if $0 < v \leq w < v + 1\text{ulp}(v)$, then $[v] \leq [w] \leq [v]'$. Similarly, if $0 < v \leq w < v + 2\text{ulp}(v)$ then $[v] \leq [w] \leq [v]"$. These inferences are somewhat delicate; a slight weakening of their hypotheses can vitiate them. For instance, when v and w are consecutive midway cases we can find $[v] < v < [v]' < w = v + 1\text{ulp}(v) < [w] = [v]"$ because of the way the IEEE standard rounds midway cases to nearest even. (On a DEC VAX consecutive midway cases that do not straddle powers of 2 satisfy $v < [v] < w = v + 1\text{ulp}(v) < [w] = [v]'$.) Another instance: if v and w straddle the negative of a power of 2 then $[v] < v < [v]' < [v]" = [w] < w < v + 1\text{ulp}(v) < 0$ can happen. The reader should check the tedious details in this paragraph if only to confirm that the notation is understood.

Monotonicity of $z/(1+z)$:
This function $f(z) := z/(1+z) = 1/(1+1/z)$ is *increasing* at all z except $z = -1$. The computed value of $f(z)$ is $[F(z)]$ where $F(z) := z/[1+z]$; these functions *decrease* at representable arguments z distributed in a surprisingly complicated way for so ostensibly simple a function as $f(z)$. Here is the picture:

$-\infty < z \leq -2^{N+1}$: $[1+z] = z$ in this range so $F(z) = [F(z)] = 1$.
~~~~~~~~~~~~~~~
$-2^{N+1} < z < -2^N$: For $z = -2^{N+1}+2, -2^{N+1}+4, -2^{N+1}+6, -2^{N+1}+8, \ldots, -2^N-8, -2^N-6, -2^N-4, -2^N-2$ in turn, $[1+z]$ takes the values $-2^{N+1}+4, -2^{N+1}+4, -2^{N+1}+8, -2^{N+1}+8, \ldots, -2^N-8, -2^N-4, -2^N-4, -2^N$ respectively. Consequently $[F(z)] = [z/[1+z]]$ takes respectively the values $1', 1, 1', 1, \ldots, 1, 1', 1, 1'$,
~~~~~~~~~~~~~~~~~

where $1' = 1 + 2^{1-N}$. Monotonicity is lost on this interval. (On a DEC VAX, which rounds midway cases up, $[1+z] = z$ in this interval, so $[F(z)] = 1$, which is less accurate on average but still monotonic.)

$-2^N \le z \le 0$: $[1+z] = 1+z$ exactly when $z \le -1/2$ in this range, so $F(z) = f(z)$, and therefore $F(z)$ and $[F(z)]$ are, like $f(z)$, monotonic except at $z = -1$. And when $-1/2 < z \le 0$ then $F(z) = -|z|/[1-|z|]$ is obviously monotonic, as is $[F(z)]$.

In subsequent intervals the essential observation is that no failure of monotonicity, i. e. $[F(z)] > [F(z')]$, can occur unless $0 < [1+z] < [1+z']$ occurs too.

$0 < z < 1/2$: Within this interval exist scatterings of points z at which $[F(z)]$ decreases instead of increasing as $f(z)$ does. The least such z is $z = 2^{1-N}+2^{-N}-2^{2-2N}$, for which $z' = 2^{1-N}+2^{-N}$ and $[F(z)] = 2^{1-N}+2^{-N}-2^{3-2N} = [F(z')]'$. (On a DEC VAX the least point of decrease is $z = 2^{-N}-2^{-2N}$.) The largest z at which $[F(z)] = [F(z')]' > [F(z')]$ turns out to be

$$z = 1/2 - 3/4^{\text{integer part of } (N+1)/2} .$$

The details are tedious. Other such places z are confined within certain subintervals $z_k < z < 2^{1-k}$, where $z_k = [1/(2^k-1)]$ for $k = 2, 3, 4, \ldots, N$; also z must satisfy $[1+z'] = [1+z]'$.

$1/2 \le z \le 1$: $F(z)$, and therefore also $[F(z)]$, is monotonic throughout this interval. To see why, first let $\psi := 1\text{ulp}(z)$; and classify z as *even* or *odd* according as its least significant bit is 0 or 1 . If z is even, $[1+z] = 1+z$; if odd, $[1+z] = 1 + z \pm \psi$, depending upon whether the second-last bit of z is 1 or 0 in accordance with the way the IEEE standard rounds the midway cases. Since $1/2 \le z < z' = z+\psi \le 1$, we find when z is even that

$$F(z) = z/[1+z] = z/(1+z) < (z+\psi)/(1+z+2\psi) = z'/(1+z'+\psi) \le z'/[1+z'] = F(z') .$$

And when z is odd,

$$F(z) = z/[1+z] \le z/(1+z-\psi) \le (z+\psi)/(1+z+\psi) = z'/(1+z') = z'/[1+z'] = F(z') .$$

$1 < z < 2^N$: $[F(z)]$ fails to be monotonic at many of the points z where $[1+z'] > [1+z]$; these points all lie in subintervals of the form $2^k-1 < z < 2^k$ for $k = 1, 2, 3, \ldots$. Abundant though these points may be, yet they are too rare to be found by random sampling. An economical way to find all of them in the range $1 < z < 2$ will be described but not explained: Let $\zeta = 1\text{ulp}(z) = 2^{1-N}$, and let n run through small odd integers $1, 3, 5, 7, 9, \ldots$ in succession. Whenever

$$(n - 4 + \sqrt{(n(n+8/\zeta))})/8 < \text{integer } m < (n - 2 + \sqrt{((n-2)^2+8n/z)})/8 ,$$

then $z = 1 + (4m+1)\zeta$ satisfies $[F(z)] > [F(z')]$. Whenever

$$(n - 2 + \sqrt{((n+2)^2+8n/\zeta)})/4 < \text{integer } j < (n + \sqrt{(n(n+8/\zeta))})/4 ,$$

then $z = 1 + 2j\zeta$ satisfies $[F(z)] > [F(z')]$ provided j is odd. (Monotonicity fails for DEC VAX rounding when j is even.)

$2^N \le z < 2^{N+1}$: Successive values of $[F(z)]$ oscillate between 1 and $1 - 2^{1-N}$ or $1 - 2^{-N}$. ($[F(z)]$ remains monotonic on a DEC VAX but less accurate on average.)

$2^{N+1} \le z < \infty$: $F(z) = 1$.

Monotonicity of z + z/(1+z) :
This function g(z) := z + z/(1+z) is interesting because it is used to calculate sinh(x) = sign(x) g($e^{|x|}$-1)/2 accurately from a subroutine that calculates exp(x)-1 relatively accurately. And if that latter subroutine is monotonic, then so is the computed value of sinh(x) , as we shall see when we verify for all z ≥ 0 that G(z) := z + [z/[1+z]] and [G(z)] are both monotonic. The function F(z) above will figure in the proof:

Since G(z) = z + [F(z)] , any interval on which F(z) is non-decreasing is an interval on which G(z) is increasing; among such intervals are -1 < z ≤ 0 and 1/2 ≤ z ≤ 1 . Elsewhere the proof is more complicated; there we shall deduce G(z) ≤ G(z′) from [F(z)] - [F(z′)] ≤ 1ulp(z) = z′ - z , which will follow from two facts: First, 0 < F(z) ≤ z , so 1ulp(z)/1ulp(F(z)) is a positive integer (a power of 2). Second, we shall demonstrate that F(z) - F(z′) < 1ulp(z) , so [F(z)] - [F(z′)] ≤ 1ulp(z) .

0 < z < 1/2 : In this range let u := 1ulp(z) and ψ := 2^{-N} .
~~~~~~~~~~~~~ Then u > ψz ; and if F(z) > F(z′) then, as we have observed above, [1+z′] = [1+z]′ = [1+z] + 2ψ . Consequently
0 < F(z) - F(z′) = z/[1+z] - z′/[1+z′] = z/[1+z] - (z+u)/([1+z]+2ψ)
= (2ψz - u[1+z])/([1+z]([1+z]+2ψ))
< (2 - [1+z])u/$[1+z]^2$ ≤ u as claimed.

1 < z < ∞ : In this range let u := 1ulp(z) and v := 1ulp([1+z])
~~~~~~~~~~ so that u ≤ v ≤ 2u . Now F(z) > F(z′) implies that [1+z] < [1+z′] ≤ [1+z] + 2v , whereupon
0 < F(z) - F(z′) = z/[1+z] - z′/[1+z′] ≤ z/[1+z] - (z+u)/([1+z]+2v)
= (2vz - [1+z]u)/([1+z]([1+z]+2v))
≤ (4z - [1+z])u/([1+z]([1+z]+2v)) < (9/16)u .

This completes the proof that G(z) and [G(z)] are monotonic.

Monotonicity of (z-1)/(z+1) :
Let b(z) := (z-1)/(z+1) = 1/b(-z) and B(z) := [z-1]/[z+1] = 1/B(-z) . These functions arise during argument reduction for the function arctan. Given a subprogram that calculates arctan(x) accurately enough and monotonically for |x| < √2-1 = tan(π/8) , we can use it to calculate
arctan(x) := sign(x)π/2 - arctan(1/x) for |x| > √2+1 , but
:= sign(x)(π/4 + arctan(b(|x|))) for √2-1 < |x| < √2+1 .
Of course, monotonicity must be checked as |x| passes the thresholds √2±1 since it may fail if arctan(√2-1) is computed too big. Monotonicity need not be checked for other arguments x since [B(z)] , the computed value of b(z) , is monotonic for all pertinent z = |x| ; a proof is outlined below.

In fact, B(z) is monotone increasing at every z except -1 in |z| ≤ 3 . This is obvious for -1 < z ≤ 1 , and becomes obvious for 2 ≤ |z| ≤ 3 when it is realized that [z±1] = z±1 exactly in this range. For 1 < |z| ≤ 2 we find that at least one of [z±1] = z±1 exactly, and the rounding error in the other is easily proved incapable of reversing monotonicity. (B(z) fails to be monotonic at many z in 3 < |z| < 4 , beyond our concern.)

Monotonicity of 2y - y² :
Given a subprogram that calculates arctan(z) monotonically and accurately enough for $-\infty \le z \le +\infty$, we may then calculate both

$$\arccos(x) = 2\arctan\sqrt{(1-x)/(1+x)} \quad \text{and}$$
$$\arcsin(x) = \arctan(x/\sqrt{(1-x^2)})$$

monotonically for $-1 \le x \le 1$. But the last formula for arcsin is not so accurate as we might like; when $|x|$ is slightly less than 1 the expression $1 - [x^2]$ suffers cancellation and comes out accurate to as few as N/2 significant bits, which leads to a calculated arcsin accurate to as few as 3N/4 significant bits. A better procedure for arcsin is as follows:

```
        If  |x| < 1/2  then     r := 1 - x²
                       else  {  y := 1 - |x|  ... exactly ;
                                r := 2y - y²  };
        arcsin(x) := arctan( x/√r ) .
```

Computed this way, r matches $1 - x^2$ accurately to within $+(5/8)\text{ulp}(r)$; consequently arcsin(x) is accurate to within less than 2.5 ulps for all $|x| \le 1$. And this computation preserves monotonicity, as shall now be proved.

Monotonicity is obvious for $|x| \le 1/2$, so suppose $|x| > 1/2$, whence $0 \le y = 1-|x| < 1/2$. Indeed $2^{-n-1} \le y < y+2^{-n}\psi = y' \le 2^{-n}$ for some $n = 1, 2, 3, \ldots$ and $\psi := 2^{-N}$. Let $R(y) := 2y - y^2$ so that $[R(y)]$ is the value calculated for r. Since $y^2 < 2^{-2n}$ so $|[y^2] - y^2| < 2^{-2n-1}\psi$, and similarly for $[(y')^2]$. Then

$$\begin{aligned} R(y') - R(y) &= 2(y' - y) + [y^2] - [(y')^2] \\ &> 2^{1-n}\psi + y^2 - 2^{-2n-1}\psi - (y')^2 - 2^{-2n-1}\psi \\ &= 2^{1-n}\psi(1 - y - 2^{-n}) \\ &> 0 , \text{ confirming monotonicity.} \end{aligned}$$

Monotonicity of w + 1/w :
This function $c(w) := w + 1/w$ is increasing for all $w > 1$. It is interesting because it provides both $\cosh(x) = c(e^{|x|})/2$ and $\sin(\theta) = 2/c(\cot(\theta/2))$, for $|\theta| \le \pi/2$, as functions that inherit their monotonicity from subprograms that evaluate e^x and $\cot(\theta/2)$ monotonically. That inheritance is not jeopardized by roundoff because, as we shall show, both $C(w) := w + [1/w]$ and $[C(w)]$ are nondecreasing for all representable $w \ge 1$. However, the formula $\sin(\theta) = 2/c(\tan(\theta/2))$ does jeopardize monotonicity because $[C(w)]$ increases at some arguments w in the interval $0 < w < 1$ whereas c(w) is decreasing therein. Proofs follow:

$2 \le w \le \infty$: In this range we may assume $2^n \le w < 2^{n+1}$ for some $n = 1, 2, 3, \ldots$. Then $u := 1\text{ulp}(w) = 2^{n+1-N}$ and $2^{-n} \ge 1/w > 1/w' = 1/(w+u) \ge 2^{-1-n}$, so $2v := 1\text{ulp}(1/w') = 2^{-n-N}$. Now

$$\begin{aligned} C(w')-C(w) &= w'+[1/w'] - w-[1/w] = u + [1/(w+u)] - [1/w] \\ &> u + (1/(w+u) - v) - (1/w + v) = u-2v - u/(w^2+wu) \\ &> u - 2v - u/4 > 0 \text{ as claimed.} \end{aligned}$$

$1 \le w < 2$: For use in this interval we introduce temporarily $\{x\}$:= x rounded to N+1 significant bits, just as $[x]$ = x rounded to N significant bits. Then set $D(w) := \{c(w)-1\}$ and observe that, because $1 \le D(w) \le 3/2$, the rounded value D(w) is obtained by rounding off bits past the N^{th} after the binary point. Because $0 \le w-1 < 1$, the fraction w-1 is representable exactly in N-1 bits after the point, and so the bits rounded off w-1 + 1/w to get D(w) are just the bits of 1/w lying beyond the N^{th}. And $1/2 < 1/w \le 1$. Evidently

$D(w) = w-1 + [1/w] = C(w)-1$. As the rounded value of a monotonic function, $D(w)$ must be monotonic too, and therefore so must be $C(w) = 1 + D(w)$, as claimed. (Proofs this easy are unusual.)

$0 < w < 1$: $c(w)$ is decreasing throughout this interval, but $[C(w)]$ increases at a scattering of arguments w in the interval. We shall see that they are scattered unevenly. For each $n = 1, 2, 3, \ldots$ suppose $2^{-n} \le w < w' = w+u \le 2^{1-n}$ where $u := 1ulp(w) = 2^{1-n-N}$. Then $2^{n-1} \le 1/w' < 1/w \le 2^n$ and $v := 1ulp(1/w') = 2^{n-N} \ge 2u$. If now $C(w) < C(w')$ then $0 \le [1/w] - [1/w'] < w'-w = u < v$, implying $[1/w] = [1/w']$. Consequently $v > 1/w - 1/w' = u/(ww')$, whence follows $(w')^2 > ww' > u/v = 2^{1-2n}$, which means $w' > 2^{-n}\sqrt{2}$. Therefore monotonicity can fail only in subintervals where $2^{-n}\sqrt{2} < w < 2^{1-n}$. Further detailed analysis reveals that successive failures in those subintervals are separated on average by roughly 2^{2-N} for $1 \le n \le N/2$. Moreover, $[2/[C(w)]] > [2/[C(w')]]$ at many of those failures, so calculating $\sin(\theta) = 2/c(\tan(\theta/2))$ will not inherit monotonicity from $\tan(\theta/2)$ for all $|\theta| < \pi/4$; some other way has to be found to calculate $\sin(\theta)$.

Trigonometric functions:
Suppose a subprogram is available to calculate $T(\theta) := 2\tan(\theta/2)$ accurately enough and monotonically for all $|\theta| \le \pi/4$. Programs that calculate all trigonometric functions everywhere can be built out of *calls* upon this one subprogram $T(\theta)$. Such programs are readily portable from one computer to another provided both have binary floating-point arithmetic; only subprogram $T(\theta)$ need be much altered to accommodate different precisions. For instance, here is a procedure to calculate $\tan(\theta)$ for all $|\theta| \le \pi/2$:

```
      If  |θ| ≤ π/8  then  tan(θ) := T(2θ)/2 ;
      if  π/8 ≤ |θ| ≤ 3π/8  then { t := T(2|θ| - π/2) ;
                                   tan(θ) := sign(θ)(2+t)/(2-t) };
      if  3π/8 ≤ |θ| ≤ π/2  then  tan(θ) := 2sign(θ)/T(π - 2|θ|) .
```

This procedure's $\tan(\theta)$ inherits from $T(\theta)$ its accuracy and its monotonicity except possibly when θ crosses one of the thresholds $\pm\pi/8$ and $\pm 3\pi/8$, where some adjustments may be necessary to preserve monotonicity. Those adjustments can be sometimes as simple as deciding which of the procedure's "$\le$" signs to replace with "$<$" signs; but if $T(\pi/4)$ is much too big the necessary adjustments may entail replacing $T(\theta)$ by a more accurate subprogram.

$\sin(\theta)$ and $\cos(\theta)$ can be calculated from $t := T(\theta)$ fairly accurately for all $|\theta| \le \pi/4$ by using the following procedure:

```
    t := T(θ) ;   q := t² ;   sin(θ) := t - t/(1+4/q) ;
    if  q ≤ 4/15  then  cos(θ) := 1 - 2/(1+4/q)   ...  ≥ 7/8  ...
                  else  cos(θ) := 3/4 + ((1-2q) + q/4)/(4+q) .
```

These expressions would be monotonic functions of t , and hence of θ , for all $|t| \le T(\pi/4) = 2/(1 + \sqrt{2})$ if roundoff did not intervene. Does roundoff destroy their monotonicity? No. ...

Monotonicity of $t - t/(1+4/t^2)$:
This function $s(t) := t - t/(1+4/t^2) = 4t/(4+t^2)$ is increasing when $0 \le t \le 2/(1+\sqrt{2})$ because $s'(t) = 4(4-t^2)/(4+t^2)^2 > 0.6$, although $s''(t) \le 0$. Among simple expressions algebraically

equivalent to $s(t)$, including $t-t^3/(4+t^2)$ and $t-t^2/(t+4/t)$ too, the particular expression chosen above and below for $S(t)$ suffers less from roundoff than the others and is in consequence provably monotonic despite roundoff, whereas the others are not.

Let $S(t) := t - [t/[1+[4/[t^2]]]]$, so that $[S(t)]$ is the value calculated for $\sin(\theta) = s(t)$. Let $\psi := 2^{-N}$; and write, say, "$[t^2] = t^2(1\pm\psi)$" to mean that $[t^2]$ lies between $t^2(1-\psi)$ and $t^2(1+\psi)$, as is the case for binary arithmetic rounded to N sig. bits. This notation will facilitate an error-analysis whose goal is to infer that $S(t)$ is monotonic from inequalities of the form $\Delta s(t) > \varepsilon s(t)+\varepsilon s(t')$ where $\Delta s(t) := s(t') - s(t) > s'(t)\,(t'-t)$ and $\varepsilon s(t)$ is a bound for the contribution of roundoff to $S(t)$; $\varepsilon s(t) > |S(t)-s(t)|$. Terms of order ψ^2 will be ignored during the error-analysis because they don't matter. Negative values of t can be skipped over because $S(-t) = -S(t)$, so the proof deals only with t in the interval $0 < t \le 2/(1+\sqrt{2}) = 0.82843$, and that interval is dealt with in three overlapping pieces:

$0 < t < 0.55$: Now $[t^2] = t^2(1\pm\psi)$ and $[4/[t^2]] = (4/t^2)(1\pm\psi)^2$ and $[t/[1+[4/[t^2]]]] = (t/(1+[4/[t^2]]))(1\pm\psi)^2$, whence $S(t) = s(t) \pm \varepsilon s(t)$ with $\varepsilon s(t) = 2\psi t^3(8+t^2)/(4+t^2)^2$. And $\Delta s(t) > s'(t)\,\psi t$; therefore $\Delta s(t) > 2\varepsilon s(t)$ whenever $4\psi t(4-t^2)/(4+t^2)^2 > 4\psi t^3(8+t^2)/(4+t^2)^2$, which is true for all t under consideration now.

$0.54 < t < 0.78$: $[4/[t^2]] = (4/t^2)(1\pm\psi)^2$ again; moreover $[1 + [4/[t^2]]] = 1 + [4/[t^2]] + 8\psi < 15$, and $[\,t/[1 + [4/[t^2]]]\,] = t/[1 + [4/[t^2]]] \pm \psi/16 < 0.11$. Therefore $\varepsilon s(t) = (\psi/16)t^3(132 + 129t^2)/(4+t^2)^2$ this time. And $\Delta s(t) > s'(t)\psi$ now, so $\Delta s(t) > 2\varepsilon s(t)$ whenever $4\psi(4-t^2)/(4+t^2)^2 > (\psi/8)t^3(132+129t^2)/(4+t^2)^2$, which is true for all t under consideration now.

$0.77 < t < 0.83$: Now $0.59 < [t^2] = t^2 \pm \psi/2 < 0.69$, and $5.8 < [4/[t^2]] = 4/[t^2] \pm 4\psi < 6.8$. Then $[1 + [4/[t^2]]] = 1 + [4/[t^2]]$ exactly because it lies between 6.8 and 7.8; the absence of a rounding error here is what makes the proof work. The third rounding error is committed when we find $0.09 < [\,t/(1+[4/[t^2]])\,] = t/(1+[4/[t^2]]) \pm \psi/16 < 0.122$, and then $|S(t)-s(t)| < \varepsilon s(t) = 2\psi t(1+2t^4)/(4+t^2)^2 + \psi/16$. Since $\Delta s(t) > s'(t)\psi$ again, $\Delta s(t) > 2\varepsilon s(t)$ for all t in question because $4\psi(4-t^2)/(4+t^2)^2 > 4\psi t(1+2t^4)/(4+t^2)^2 + \psi/8$. Here ends the proof that $S(t)$ is monotonic.

As a byproduct of the proof we find that $\varepsilon s(t) < 0.29\,\mathrm{ulp}(S(t))$.

Monotonicity of $3/4 + ((1-2t^2) + t^2/4)/(4+t^2)$:
Let $h(t) := (4-t^2)/(4+t^2)$. This formula could be used to compute $\cos(\theta) = h(2\tan\theta/2)$, but it is not quite accurate enough. The error $[[4-[t^2]]/[4+[t^2]]] - h(t)$ can approach 1.5 ulps, and it destroys the identity $\arccos(\cos(\arccos x)) = \arccos x$ when x is slightly less than 1 **(²)**. A better procedure to compute $\cos\theta$ was given above. It rearranges $h(t) = 1 - 2/(1+4/t^2)$ to retain monotonicity and achieve better accuracy, within 0.86 ulps when $|t| < 2/\sqrt{15}$ and better than that when $|t|$ is very tiny. (Other rearrangements, like $1 - 2t/(t+4/t)$ and $1 - 2t^2/(4+t^2)$, are

comparably accurate and faster, but not monotonic.) And when $2/\sqrt{15} < |t| < 2/(1+\sqrt{2})$ the procedure uses another rearrangement $h(t) = 3/4 + ((1-2t^2) + t^2/4)/(4+t^2)$ to achieve accuracy within 0.84 ulps and monotonicity. Here is the proof of monotonicity:

Let $q := [t^2]$; it increases monotonically between roughly $4/15 = 0.2666$ and $4/(3+2\sqrt{2}) = 0.6863$. Throughout that range $[[1-[2q]] + [q/4]] = [1 - 7q/4]$ because $q/4$, $2q$ and $1-2q$ are all computed exactly. Therefore only $[1 - 7q/4]/[4+q]$ need be proved monotonic. It is obviously monotone nonincreasing while $q < 4/7 = 0.5714...$. Otherwise, while $4/7 < q < 0.6863$, the numerator $[1 - 7q/4]$ decreases through $0 > [1-7q/4] > -0.2011$ in steps of at least 6 ulps, whereas the denominator $[4+q]$ can increase by at most an ulp when q increases by an ulp. Therefore monotonicity is confirmed again.

Acknowledgements:
The foregoing work was undertaken as part of an ongoing informal Elementary Functions Project in which the other participants are currently Alex Z.-S. Liu, Stuart McDonald, Peter P. Tang and Dr. Kwok Choi Ng. Their work has been supported respectively by Motorola, Zilog, ELXSI and National Semiconductor Corporations. This author's work has been supported in part by the U. S. Office of Naval Research and the U. S. Air Force Office of Scientific Research. I am deeply grateful to all of them for their help.

Footnotes:
(¹) Draft 1.0 of p854 has been published, to invite public comment, in the IEEE magazine *MICRO* **4** no. 4 (Aug. 1984) pp.86-100. It contains specifications for both Decimal and Binary arithmetic, with precisions determined by the implementer subject to mild constraints; in Binary the number N of significant bits must exceed 17 . Draft 10.1 of p754 specifies only Binary arithmetic, and further restricts N . An earlier draft 8.0 of p754, published in IEEE magazine *COMPUTER* in March of 1981, has been superseded by draft 10.1, which is expected to be adopted officially in mid 1985.

(²) "Elementary Functions from Kernels", by W. Kahan, will, when it appears, contain formulas that derive economically all of the elementary transcendental functions each via several algebraic operations upon just a few programs that deliver exp, log, tan and arctan within restricted domains.

(³) "VAX" is a trademark of Digital Equipment Corp. The VAX line provides Binary floating-point arithmetic in four formats: F has $N=24$; D has $N=56$; G has $N=53$; H has $N=113$.

Author's address: Prof. W. Kahan,
Elect. Eng. and Computer Science Dept.,
University of California,
Berkeley, California 94720

This manuscript was prepared on an IBM PC and printed from a special character font downloaded on an EPSON FX-80.