SHARP ULP ROUNDING ERROR BOUND
FOR THE HYPOTENUSE FUNCTION

ABRAHAM ZIV

ABSTRACT. The hypotenuse function, $z = \sqrt{x^2 + y^2}$, is sometimes included in math library packages. Assuming that it is being computed by a straightforward algorithm, in a binary floating point environment, with round to nearest rounding mode, a sharp roundoff error bound is derived, for arbitrary precision. For IEEE single precision, or higher, the bound implies that $|\bar{z} - z| < 1.222\, ulp(z)$ and $|\bar{z} - z| < 1.222\, ulp(\bar{z})$. Numerical experiments indicate that this bound is sharp and cannot be improved.

1. INTRODUCTION

The hypotenuse function, $z = \sqrt{x^2 + y^2}$, is sometimes included in math libraries. Often error bounds for math library subroutines are given in ULP's (units in the last place) rather than as relative error bounds. In this paper we find, by a rigorous error analysis, a ULP bound. Namely, a bound of the form $|\bar{z} - z| < \alpha \times ulp(z)$ or $|\bar{z} - z| < \alpha \times ulp(\bar{z})$, where $\bar{z}$ is the approximation computed for z and α is a constant. The algorithm analyzed is the straightforward one:

$$x_1 = x, \quad x_2 = y, \quad x_3 = x_1 \times x_1 = x^2, \quad x_4 = x_2 \times x_2 = y^2,$$

$$x_5 = x_3 + x_4 = x^2 + y^2, \quad x_6 = \sqrt{x_5} = \sqrt{x^2 + y^2} = z.$$

The floating point arithmetic assumed is binary with round to nearest rounding mode. The precision (number of binary digits of the mantissa) assumed is $p \in \{3, 4, 5, \dots\}$. It is assumed that neither overflow nor underflow occurs during the computations.

Usually, the function $ulp(x)$ is discussed only for values of x which are machine numbers (see, e.g., Higham [1], Section 2.1). It is very easy, though, to generalize its definition for general real numbers. The following definition, for instance, is meaningful for all real x: Define the mantissa $m(x)$ and the exponent $e(x)$, for $x \neq 0$, by the relations $|x| = m \times 2^e$, $e \in \{0, \pm 1, \pm 2, \dots\}$, $\frac{1}{2} \leq m < 1$. Then, for a given precision $p \in \{1, 2, 3, \dots\}$, define $ulp(x)$ by $ulp(x) = 2^{e-p}$, $ulp(0) = 0$.

The main part of the error analysis is performed by the method described in [5]. Namely, we compute linearized relative error bounds l_i for x_i ($i = 1, \dots, 6$) and convert the approximate bound $l = l_6$ into an exact bound by Theorem 3.1 of [5]:

Theorem (Ziv [5]). *Let δ be equal to $(2n_\pm + 1)B$, where $n_\pm$ is the total number of $\pm$ operations in the algorithm and $B = \max l_i$. If $\delta < 1$, then the accumulated relative error is bounded by $l/(1 - \delta)$, where l is the accumulated linearized error bound.* $\square$

The analysis for a constant bound, independent of x and y, for the relative error is very simple: Let l_i denote approximate linearized bounds of the relative errors $|\bar{x}_i - x_i|/|x_i|$, where $\bar{x}_i$ are the approximate values computed for x_i. One gets $l_1 = l_2 = 0$, $l_3 = l_4 = 2^{-p}$, $l_5 = (x_3 l_3 + x_4 l_4)/x_5 + 2^{-p}$, $l_6 = \frac{1}{2}l_5 + 2^{-p}$; so $l_5 = 2^{1-p}$, $l = l_6 = 2^{1-p}$, and from the theorem one gets the bound $|\bar{z} - z|/z \le 2^{1-p}/(1 - \delta)$, where $\delta = 6 \times 2^{-p}$.

This bound for the relative error can be translated into a ULP bound by the relation $ulp(z)/z > 2^{-p}$, which implies

$$|\bar{z} - z| < 2\, ulp(z)/(1 - \delta), \qquad \delta = 6 \times 2^{-p}.$$

Our purpose is to improve (reduce) the value of the factor 2, in the numerator, as much as possible. In order to do this we assume, for the local roundoff errors, the tighter bound $\frac{1}{2}ulp$ instead of the relative error bound 2^{-p} used in the last analysis.

In §2 the main result of this paper, formula (2.1), is stated and proved. In §3 number experiments are discussed. Their purpose was to find out how close to the theoretical bound the actual accumulated roundoff error may be. It turns out that the theoretical bound is quite sharp and practically cannot be improved.

It is worth mentioning that often the algorithm used for the hypotenuse function is not the one we analyze here, but the following: $z = a\sqrt{1 + (b/a)^2}$, where $a = \max\{|x|, |y|\}$, $b = \min\{|x|, |y|\}$. This algorithm has the advantage of not overflowing, unless z is too large. It also does not suffer accuracy loss if $(b/a)^2$ underflows. On the other hand, this algorithm is somewhat less accurate and somewhat slower than the algorithm discussed here. The intermediate overflow/underflow problem of our algorithm may be solved without any loss of accuracy by scaling; that is, by computing $2^e\sqrt{(x2^{-e})^2 + (y2^{-e})^2}$. The integer e may be chosen equal to (or close to) $\max\{e(x), e(y)\}$. Such scaling needs to be applied only when $|x|$ and $|y|$ have extreme values. Thus, for instance, with IEEE double precision arithmetic (see [1] or [2]), it suffices to apply scaling only in the rare cases where $e(x)$ or $e(y)$ falls outside of the range $[-500, 500]$.

2. ULP ERROR ANALYSIS

In this section we prove the main results of this paper:

(2.1)
$$\boxed{\begin{array}{c} |\bar{z} - z| < \alpha\, ulp(z) \quad \text{and} \quad |\bar{z} - z| < \alpha\, ulp(\bar{z}), \\[2mm] \text{where } \alpha = \dfrac{5\sqrt{3} + 6}{12(1 - 7 \times 2^{-p})}, \quad p \in \{3, 4, \dots\}. \end{array}}$$

This result implies that for $p \ge 15$

$$|\bar{z} - z| < 1.222\, ulp(z), \qquad |\bar{z} - z| < 1.222\, ulp(\bar{z}).$$

Let $a > 0$ be a real number and assume that it is rounded to the nearest machine number $\bar{a}$. Usually $ulp(\bar{a}) = ulp(a)$. The only exceptional case is when a is slightly smaller than an integral power of 2, say 2^n, and is rounded upward into $\bar{a} = 2^n$. In this exceptional case $ulp(\bar{a}) = 2\, ulp(a)$. In all cases, including the exceptional,

$|\overline{a} - a| \leq \frac{1}{2}ulp(a) \leq \frac{1}{2}ulp(\overline{a})$. In view of this, the linearized bounds, l_i on the relative errors $|\overline{x}_i - x_i|/|x_i|$ $(i = 1, 2, \ldots, 6)$ may be chosen as

$$l_1 = l_2 = 0, \qquad l_3 = \frac{1}{2}ulp(x_3)/x_3, \qquad l_4 = \frac{1}{2}ulp(x_4)/x_4,$$

$$l_5 = (x_3 l_3 + x_4 l_4)/x_5 + \frac{1}{2}ulp(\overline{x}_3 + \overline{x}_4)/(\overline{x}_3 + \overline{x}_4),$$

$$l_6 = \frac{1}{2}l_5 + \frac{1}{2}ulp(\sqrt{\overline{x}_5})/\sqrt{\overline{x}_5}.$$

We get

$$l = l_6 = \frac{1}{4}\frac{ulp(x_3) + ulp(x_4)}{x_5} + \frac{1}{4}\frac{ulp(\overline{x}_3 + \overline{x}_4)}{\overline{x}_3 + \overline{x}_4} + \frac{1}{2}\frac{ulp(\sqrt{\overline{x}_5})}{\sqrt{\overline{x}_5}}.$$

In order to convert the linearized bound $l = l_6$ into a true bound we have to divide it by $1 - \delta$. The value $\delta = 6 \times 2^{-p}$, used earlier, may be used here too.

Noting that

$$\overline{x}_3 + \overline{x}_4 \geq x_3(1 - 2^{-p}) + x_4(1 - 2^{-p}) = x_5(1 - 2^{-p}),$$

$$\sqrt{\overline{x}_5} \geq \sqrt{(\overline{x}_3 + \overline{x}_4)(1 - 2^{-p})} \geq x_6(1 - 2^{-p}),$$

one gets

(2.2)
$$|\overline{z} - z| \leq \frac{zl}{(1 - \delta)}$$
$$< \left(\frac{1}{4}\frac{ulp(x^2) + ulp(y^2) + ulp(\overline{x}_3 + \overline{x}_4)}{\sqrt{x^2 + y^2}} + \frac{1}{2}ulp(\sqrt{\overline{x}_5})\right)\Big/(1 - 7 \times 2^{-p}).$$

The rest of the discussion is split into four cases:

- Case I: $ulp(\overline{x}_3 + \overline{x}_4) \leq ulp(x_3 + x_4)$, $ulp(\sqrt{\overline{x}_5}) \leq ulp(\sqrt{x_5})$, $ulp(\overline{z}) \geq ulp(z)$.
- Case II: $ulp(\overline{x}_3 + \overline{x}_4) > ulp(x_3 + x_4)$.
- Case III: $ulp(\sqrt{\overline{x}_5}) > ulp(\sqrt{x_5})$.
- Case IV: $ulp(\overline{z}) < ulp(z)$.

These four cases cover all possibilities. Case I is the principal one; Cases II, III, and IV are very rare. They cannot be ignored, though.

Case I. In this case we deduce from (2.2) that

(2.3)
$$|\overline{z} - z| < \frac{\frac{1}{4}A(x, y) + \frac{1}{2}}{1 - 7 \times 2^{-p}}ulp(z), \quad \text{where } A(x, y) = \frac{ulp(x^2) + ulp(y^2) + ulp(x^2 + y^2)}{\sqrt{x^2 + y^2}\, ulp(\sqrt{x^2 + y^2})}.$$

In order to prove (2.1) we shall show that $\max A(x, y) = 5/\sqrt{3}$. Since $ulp(z) \leq ulp(\overline{z})$, both inequalities of (2.1) are thus proved simultaneously.

Because of symmetry we may assume that $0 \leq y \leq x$ and $x > 0$. Note also that the problem may be scaled by multiplying both x and y by the same integral power of 2. In what follows we shall use scaling often.

Starting from some point (x, y), we shall change x and y continuously in such a manner as to increase $A(x, y)$, up to a local maximum. $\max A(x, y)$ is equal to the largest of the local maxima.

We start by increasing y and decreasing x, keeping $x^2 + y^2$ constant. We may continue with this up to a point where either $y^2 = x^2$ or $y^2 < x^2 = 2^n$ $(n = 0, \pm 1, \pm 2, \ldots)$.

If $y^2 = x^2$ we have $A = 4\,ulp(x^2)/(\sqrt{2}\,x\,ulp(\sqrt{2}\,x))$, and we may decrease x up to a point where x^2 is an even power of 2. Scaling the problem, we may assume that $x^2 = 1$, so $A = 2\sqrt{2} < 5/\sqrt{3}$.

If $y^2 < x^2 = 2^n$ we may scale the problem to get $\frac{1}{2} \le x^2 < 2$, so either $y^2 < x^2 = 1$ or $y^2 < x^2 = \frac{1}{2}$.

If $y^2 < x^2 = 1$ we get $A = (2^{1-p} + ulp(y^2) + 2^{1-p})/(\sqrt{1+y^2} \times 2^{1-p})$, and we may decrease y until $y^2 = 2^{-n}$ $(n = 1, 2, \dots)$; so $A = (2 + 2^{-n})/\sqrt{1 + 2^{-n}} < 2 + 2^{-1} = 2.5 < 5/\sqrt{3}$.

If $y^2 < x^2 = \frac{1}{2}$ we get $A = (2^{-p} + ulp(y^2) + 2^{-p})/(\sqrt{\frac{1}{2} + y^2} \times 2^{-p})$. We decrease y until $y^2 = 2^{-n}$ $(n = 2, 3, \dots)$; so $A = (2 + 2^{1-n})/\sqrt{\frac{1}{2} + 2^{-n}}$. This expression decreases with n, so its largest value is attained when $n = 2$, i.e., $A = 5/\sqrt{3}$. This completes the proof of (2.1) for Case I.

Case II. In this case $\overline{x}_3 + \overline{x}_4 > x_3 + x_4 = x_5$, and an integral power of 2 must lie in the interval $(x_5, \overline{x}_3 + \overline{x}_4]$. Scaling the problem so that $\frac{1}{2} \le x_5 < 2$, we find that there are two possibilities: either $x_5 < 1 \le \overline{x}_3 + \overline{x}_4$ or $x_5 < 2 \le \overline{x}_3 + \overline{x}_4$.

If $x_3 + x_4 = x_5 < 1 \le \overline{x}_3 + \overline{x}_4$, then either $x_3 < \frac{1}{2}$ or $x_4 < \frac{1}{2}$, and we have $1 \le \overline{x}_3 + \overline{x}_4 \le x_3 + x_4 + \frac{1}{2}ulp(\frac{1}{2}) + \frac{1}{2}ulp(\frac{1}{4}) = x_5 + \frac{3}{8}ulp(1) < 1 + \frac{3}{8}ulp(1)$. It follows that $\overline{x}_3 + \overline{x}_4$ rounds into $\overline{x}_5 = 1$, so $\overline{x}_6 = 1$ too. As for $x_6 = \sqrt{x_5}$, we get

$$1 - \tfrac{3}{8}ulp(1) \le x_5 < 1 \Rightarrow 1 - \tfrac{3}{8}ulp(1) < x_6 = z < 1$$
$$\Rightarrow 0 < \overline{x}_6 - x_6 < \tfrac{3}{8}ulp(1) = \tfrac{3}{8}ulp(\overline{z}).$$

Hence $|\overline{z} - z| < \frac{3}{8}ulp(\overline{z}) = \frac{3}{4}ulp(z)$, which is compatible with (2.1).

If $x_3 + x_4 = x_5 < 2 \le \overline{x}_3 + \overline{x}_4$, then either $x_3 < 1$ or $x_4 < 1$, and we have $2 \le \overline{x}_3 + \overline{x}_4 \le x_3 + x_4 + \frac{1}{2}ulp(1) + \frac{1}{2}ulp(\frac{1}{2}) = x_5 + \frac{3}{8}ulp(2) < 2 + \frac{3}{8}ulp(2)$, from which it follows that $\overline{x}_3 + \overline{x}_4$ rounds into $\overline{x}_5 = 2$; so $\sqrt{2} - \frac{1}{2}ulp(1) \le \overline{z} = \overline{x}_6 \le \sqrt{2} + \frac{1}{2}ulp(1)$. Also

$$2 - \tfrac{3}{8}ulp(2) \le x_5 < 2 \Leftrightarrow 2(1 - \tfrac{3}{8}ulp(1)) \le x_5 < 2$$
$$\Rightarrow \sqrt{2}(1 - \tfrac{3}{8}ulp(1)) < z = x_6 < \sqrt{2}.$$

Hence $|\overline{z} - z| < [(4 + 3\sqrt{2})/8]ulp(1)$. Since $ulp(\overline{z}) = ulp(z) = ulp(1)$, this is also compatible with (2.1). This completes the proof of (2.1) in Case II.

Case III. In this case $\sqrt{\overline{x}_5} > \sqrt{x_5} = z$, and an integral power of 2 must lie in the interval $(\sqrt{x_5}, \sqrt{\overline{x}_5}]$. Scaling to get $1/\sqrt{2} \le \sqrt{\overline{x}_5} < \sqrt{2}$, we have $\sqrt{x_5} < 1 \le \sqrt{\overline{x}_5} \Leftrightarrow x_5 < 1 \le \overline{x}_5$. Now there are two possibilities: either $\overline{x}_3 + \overline{x}_4 \ge 1$ or $\overline{x}_3 + \overline{x}_4 < 1$. In the first case $ulp(\overline{x}_3 + \overline{x}_4) > ulp(x_5)$. This falls in Case II, which was already discussed. We are left, then, with the case $x_5 < 1 \le \overline{x}_5$, $\overline{x}_3 + \overline{x}_4 < 1$. In this case $\overline{x}_3 + \overline{x}_4$ must round upward to $\overline{x}_5 = 1$, so $\overline{x}_6 = \overline{z} = 1$ too. Hence $0 < 1 - (\overline{x}_3 + \overline{x}_4) \le \frac{1}{2}ulp(\frac{1}{2})$. Also, since $x_5 = x_3 + x_4 < 1$, we have either $x_3 < \frac{1}{2}$ or $x_4 < \frac{1}{2}$; so $\overline{x}_3 + \overline{x}_4 - x_5 \le \frac{1}{2}ulp(\frac{1}{2}) + \frac{1}{2}ulp(\frac{1}{4}) = \frac{3}{8}ulp(1)$. Combining the two

results, we get

$$0 < 1 - x_5 \leq \tfrac{3}{8}ulp(1) + \tfrac{1}{2}ulp(\tfrac{1}{2}) = \tfrac{5}{8}ulp(1)$$

$$\Rightarrow x_5 \geq 1 - \tfrac{5}{8}ulp(1) \Rightarrow x_6 \geq \sqrt{1 - \tfrac{5}{8}ulp(1)}$$

$$\Rightarrow 0 < \overline{x}_6 - x_6 = 1 - x_6 \leq 1 - \sqrt{1 - \tfrac{5}{8}ulp(1)} < \tfrac{5}{16}ulp(1)/(1 - \tfrac{5}{8}ulp(1))$$

$$= \tfrac{5}{16}ulp(1)/(1 - 1.25 \times 2^{-p}).$$

Since $ulp(z) = ulp(\tfrac{1}{2})$ and $ulp(\overline{z}) = ulp(1)$, we have

$$0 < \overline{z} - z \leq \frac{5ulp(\overline{z})}{16(1 - 1.25 \times 2^{-p})} = \frac{10ulp(z)}{16(1 - 1.25 \times 2^{-p})},$$

which is consistent with (2.1). This completes the proof of (2.1) in Case III.

Case IV. In this case $\overline{z} < z$, and an integral power of 2 must lie in the interval $(\overline{z}, z]$. Scaling so that $1/\sqrt{2} \leq z < \sqrt{2}$, we have

$$\overline{z} < 1 \leq z \Rightarrow \sqrt{\overline{x}_5} < 1 \leq \sqrt{x_5} \Rightarrow \overline{x}_5 < 1 \leq x_5$$

$$\Rightarrow \overline{x}_3 + \overline{x}_4 \leq 1 - \tfrac{1}{2}ulp(\tfrac{1}{2}) < 1 \leq x_3 + x_4.$$

Hence either $\overline{x}_3 < \tfrac{1}{2}$ or $\overline{x}_4 < \tfrac{1}{2}$, and we have

$$1 \leq x_3 + x_4 \leq \overline{x}_3 + \overline{x}_4 + \tfrac{1}{2}ulp(\tfrac{1}{4}) + \tfrac{1}{2}ulp(\tfrac{1}{2}) \leq 1 + \tfrac{1}{2}ulp(\tfrac{1}{4})$$

$$\Rightarrow 1 \leq x_5 \leq 1 + \tfrac{1}{4}ulp(\tfrac{1}{2}) \Rightarrow 1 \leq z = \sqrt{x_5} < 1 + \tfrac{1}{8}ulp(\tfrac{1}{2}).$$

Also,

$$1 - \tfrac{3}{4}ulp(\tfrac{1}{2}) \leq \overline{x}_3 + \overline{x}_4 \Rightarrow 1 - ulp(\tfrac{1}{2}) \leq \overline{x}_5 < 1$$

$$\Rightarrow \overline{x}_5 = 1 - ulp(\tfrac{1}{2}) \Rightarrow 1 - ulp(\tfrac{1}{2}) < \sqrt{\overline{x}_5} < 1 - \tfrac{1}{2}ulp(\tfrac{1}{2})$$

$$\Rightarrow \overline{z} = \overline{x}_6 = 1 - ulp(\tfrac{1}{2}) \Rightarrow 0 < z - \overline{z} < ulp(\tfrac{1}{2}) + \tfrac{1}{8}ulp(\tfrac{1}{2})$$

$$= \tfrac{9}{8}ulp(\overline{z}) = \tfrac{9}{16}ulp(z),$$

which is consistent with (2.1). This completes the proof of (2.1) in Case IV.   □

3. NUMERICAL TESTS

The purpose of the numerical tests was to find out how close to the theoretical bound the actual accumulated roundoff error can be.

Given a pair (x, y) of IEEE single precision numbers (see [1] or [2]), the algorithm was performed first in IEEE single precision arithmetic ($p = 24$) and then in IEEE double precision arithmetic ($p = 53$). The signed difference between the two results was taken to be a very good estimate of the actual accumulated error in the single precision computation. This kind of computation was repeated for many single precision pairs (x, y), and the maximal and minimal errors, expressed in single precision ULP's, were printed out.

Reading carefully the error analysis in §2, Case I, we concluded that the maximal accumulated error is likely to occur for some pair (x, y) where x is slightly larger than $1/\sqrt{2}$ and y is slightly larger than $\tfrac{1}{2}$. In view of this, the error was computed for the 9×10^6 (x, y) combinations formed from the single precision values of $x = (1 + 2^{-12})/\sqrt{2} + m \times 2^{-24}$, $y = (1 + 2^{-12})/2 + n \times 2^{-24}$, $m, n \in \{0, 1, \ldots, 2999\}$.

The smallest error found was $-1.213571u$ and the largest $+1.213855u$, where $u = 2^{-24} = ulp(z) = ulp(\bar{z})$. Comparing this with the theoretical bound, $1.222u$, the largest actual error found is less than 0.7% away from this bound. This means that the bound is quite sharp and practically cannot be improved.

REFERENCES

1. Nicholas J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, Philadelphia, PA, 1996. MR **97a:**65047
2. *IEEE standard for binary floating point arithmetic. An American national standard*, ANSI/IEEE Std 754-1985.
3. Pat H. Sterbenz, *Floating-point computation*, Prentice-Hall, Englewood Cliffs, NJ, 1974. MR **50:**1556
4. J. H. Wilkinson, *Rounding errors in algebraic processes*, Prentice-Hall, Englewood Cliffs, NJ, 1963. MR **28:**4661
5. Abraham Ziv, *Converting approximate error bounds into exact ones*, Math. Comp. **64** (1995), 265–277. MR **95c:**65074

IBM ISRAEL, SCIENCE AND TECHNOLOGY, MATAM–ADVANCED TECHNOLOGY CENTER, HAIFA 31905, ISRAEL
    E-mail address: ziv@haifasc3.vnet.ibm.com