Lars Eldén & Linde Wittmeyer-Koch

It should be emphasized that the values of these parameters are only *theoretical*. In practice, one has to take into account the time for memory accesses. The measured values are $r_\infty = 70$, $n_{1/2} = 53$ for vector multiplication, and $r_\infty = 148$, $n_{1/2} = 60$ for the SAXPY operation (from R.W. Hockney, C.R. Jesshope, *Parallel Computers* 2, Adam Hilger, Bristol, 1988).

Exercises

1. Show that if u is correctly rounded to s significant digits, then we have

$$\frac{|\Delta u|}{|u|} \leq \frac{1}{2}\beta^{-s+1},$$

where β is the base of the position system.

2. How accurately do we need to know an approximation of π to be able to compute $\sqrt{\pi}$ with four correct decimals?

3. Derive the error propagation formula for division.

4. Let $y = \log x$. Derive the error propagation formula for this function. Use the result to give an error propagation formula for $f(x_1, x_2, x_3) = x_1^{\alpha_1} x_2^{\alpha_2} x_3^{\alpha_3}$. This technique is sometimes called logarithmic differentiation.

5. Compute the focal distance f of a lens using the formula

$$\frac{1}{f} = \frac{1}{a} + \frac{1}{b},$$

where $a = 32 \pm 1$ mm and $b = 46 \pm 1$ mm. Give an error estimate.

6. When I lie on the beach, I can just see the top of a factory chimney across the water. On my road map, I find that the factory is on the other side of the bay 25 ± 1 km away. I recall that the radius of the earth is 6366 ± 10 km. Compute the height of the chimney and estimate the error.

Hint: Elementary geometry gives

$$h = \frac{r(1 - \cos\alpha)}{\cos\alpha}, \quad \alpha = \frac{a}{r},$$

where a is the distance to the factory and r is the radius of the earth.

7. Let f be a function from R^n to R^m, and assume that we want to compute $f(\bar{a})$, where the vector $\bar{a}$ is an approximation of a. Show that the general error propagation formula applied to each component in f leads to

$$\Delta f \approx J\,\Delta a,$$

where J is an $m \times n$ matrix with elements

$$(J)_{ij} = \frac{\partial f_i}{\partial a_j}.$$

8. Use Taylor expansion to avoid cancellation in the following expression
a) $e^x - e^{-x}$, $\quad x$ close to 0;
use a reformulation to avoid cancellation in the following expressions
b) $\sin x - \cos x$, $\quad x$ close to $\pi/4$,
c) $1 - \cos x$, $\quad x$ close to 0,
d) $(\sqrt{1+x^2} - \sqrt{1-x^2})^{-1}$, $\quad x$ close to 0.

9. Show that if f is a normalized floating point number in a floating point system (β, t, L, U), then $r \le |f| \le R$, where

$$r = \beta^L,$$

$$R = \beta^U(\beta - \beta^{-t}).$$

10. Show that $\mathrm{fl}[1 + \mathrm{x}] = 1$ for all $x \in [0, \mu]$ and that $\mathrm{fl}[1 + \mathrm{x}] > 1$ for $x > \mu$ (μ is the unit roundoff of the floating point system).

11. Show that the computation of $\mathrm{sq} := \sqrt{x_1^2 + x_2^2}$ can give overflow even if the result sq can be represented in the floating point system (e.g., take $x_1 = x_2 = 0.8 \cdot 10^5$ in the system $(10, 4, -9, 9)$). Rewrite the computation so that overflow is avoided for all data x_1, x_2 such that the result sq can be represented.

12. Assume that $n\mu < 0.1$ and $|\epsilon_i| \le \mu$, $i = 1, 2, \ldots, n$. Show that

$$|(1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_n)| \le 1 + 1.06n\mu.$$

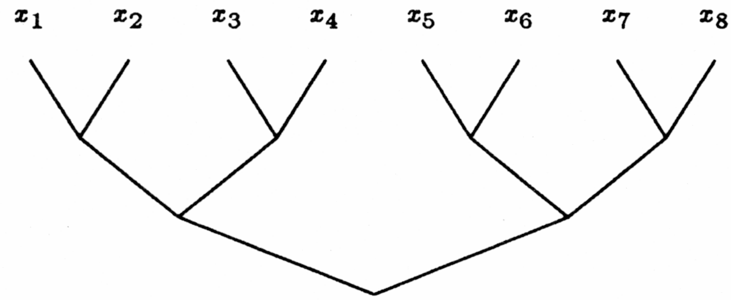
Hint: Use $(1+x)^n \le e^{nx}$ and make a series expansion.

13. Let $S_n = \sum_{i=1}^{n} x_i y_i$. Show that

$$\left|\hat{S}_n - S_n\right| \leq \sum_{i=1}^{n} |(n-i+2)x_i y_i| \; 1.06\mu.$$

(Cf. Theorem 2.7.2.)

14. Assume that $n = 2^k$, and that we compute the sum $S_n = \sum_{i=1}^{n} x_i$ in the order illustrated in the figure.



Derive the forward and backward error estimates (see Theorems 2.7.2 and 2.7.3) for this computation.

15. The second degree polynomial $p(x) = ax^2 + bx + c$ is evaluated in a floating point system using Horner's scheme (see Chapter 4). Show that the computed value $\hat{p}(x)$ satisfies

$$|\hat{p}(x) - p(x)| \leq (4|ax^2| + 3|bx| + |c|)\mu,$$

where μ is the unit roundoff. (Terms that are $O(\mu^2)$ can be discarded.)

References

The historical development of the representation of numbers is a fascinating chapter in the cultural history of mankind. A nice survey is given in

D. E. Knuth, *The art of computer programming, Volume 2 /Seminumerical algorithms*, Second edition, Addison Wesley, Reading, Massachusetts, 1981.

One is tempted to believe that the binary number system is a fruit of the development of computers. As a matter of fact, several mathematicians in the 17th and 18th centuries used binary representation for number theoretical research. Knuth's book gives a good presentation of floating point