THE UNIVERSITY OF TEXAS AT DALLAS

ERIK JONSSON SCHOOL OF ENGINEERING & COMPUTER SCIENCE

FLOATING-POINT COMPUTATION

Notes prepared for EE 6481

by

Professor Cyrus D. Cantrell

May–August 2005

REFERENCES

1. "What Every Computer Scientist Should Know About Floating-Point Arithmetic", by David Goldberg, *ACM Computing Surveys* **23**, 5–48 (1991).
2. *Matrix Computations*, Second Edition, by Gene H. Golub and Charles F. Van Loan (Johns Hopkins University Press, 1989).
3. *Numerical Computing with IEEE Floating Point Arithmetic*, by Michael L. Overton (Society for Industrial & Applied Mathematics, 2001).
4. "A Survey of Error Analysis", by W. Kahan, in *Information Processing 71*, edited by C. L. Frieman (North-Holland, 1972), pp. 1214–1239.
5. *Rounding Errors in Algebraic Processes*, by J. H. Wilkinson (Dover, 1994).
6. *Modern Mathematical Methods for Physicists and Engineers*, by C. D. Cantrell (Cambridge University Press, 2000). See especially Chapters 1, 3 and 6.

FLOATING-POINT COMPUTATION (2)

- A consequence of catastrophic cancellation:
  Huge relative errors in summing an alternating series

  Example: Computation of $\sin x$ by truncating the infinite series

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}$$

  - If $x = 40.0$, and if 80 terms are kept, the truncated series gives $\sin(40.0) = 5.2344314\text{E}+08$ in single precision
  - The correct result is $\sin(40.0) = 0.7451131$
  - The theory of convergence of alternating series says that the error in truncating a series is less than the absolute value of the first neglected term (-6.82×10^{-31} in this case)
  - **WHAT'S GOING ON HERE??? HOW CAN THIS BE???**

Computation of the sine series for large argument

Angle in radians: 40
Number of terms: 80 (Note: Not all terms are shown)
The sine series evaluates the sine as 5.2344314E+08
The value of the sine function is 0.7451131

2n+1	Value of term in sine series
1	40.0000000
3	-10666.6669922
5	853333.3750000
7	-32507938.0000000
9	722398592.0000000
11	-10507616256.0000000
13	107770421248.0000000
15	-821107949568.0000000
17	4830046715904.0000000
19	-22596710039552.0000000
21	86082704113664.0000000
23	-272198266781696.0000000
25	725862022381568.0000000
27	-1654386371592192.0000000
29	3259874540519424.0000000
31	-5608386548727808.0000000
33	8497555214172160.0000000
35	-11425284096000000.0000000
37	13724065483194368.0000000
39	-14816805021286400.0000000
41	14455420030550016.0000000
43	-12806573495681024.0000000
45	10348746248290304.0000000
47	-7658646632660992.0000000
49	5209963907514368.0000000
51	-3268997051056128.0000000
53	1897821225615360.0000000
55	-1022395288649728.0000000
57	512478853201920.0000000
59	-239616057671680.0000000
61	104750192263168.0000000
63	-42908426174464.0000000
65	16503240916992.0000000
67	-5971322077184.0000000
69	2036256473088.0000000
71	-655535308800.0000000
73	199554121728.0000000
75	-57529114624.0000000
77	15729081344.0000000
79	-4084149760.0000000
81	1008432064.0000000
83	-237068960.0000000
85	53124696.0000000
87	-11360534.0000000
89	2320844.5000000
91	-453400.6250000
93	84787.3984375
95	-15191.4707031
97	2610.2182617
99	-430.4627075
101	68.1921158
103	-10.3852453
105	1.5216477
107	-0.2146567
109	0.0291752
111	-0.0038231

ERROR ANALYSIS OF SUMMATION OF SERIES (1)

- The algorithm for summation in natural order is

$$s_{k+1} = s_k + x_{k+1}$$

  where s_k is the sum of k terms

- There are 2 major sources of error:

  If all terms are positive, then errors accumulate because small terms (x_{k+1}) are repeatedly added to large terms (s_k)

  If the terms alternate in sign, then the mathematical (not numerical) identity

$$s_{k+1} = s_k + x_{k+1} = s_{k-1} + x_{k+1} + x_k$$

  (where x_k is opposite in sign to x_{k+1}) shows that cancellation of significant digits must occur if $|x_{k+1}| \approx |x_k|$

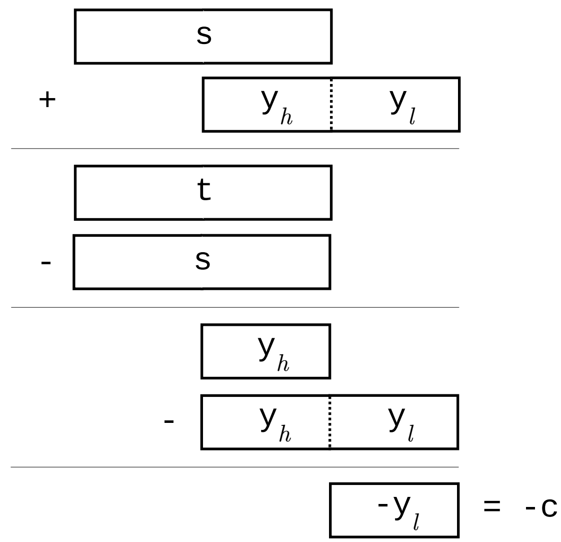
KAHAN'S SUMMATION METHOD (1)

- W. Kahan invented a way to recover the bits that are "lost" as a result of shifting before adding
- FORTRAN segment that implements Kahan's algorithm:

```
      s=0.
      c=0.
      do 100 j=1,N
        y=c+x(j)
        t=s+y
        c=(s-t)+y
100   s=t
      s=s+c
```

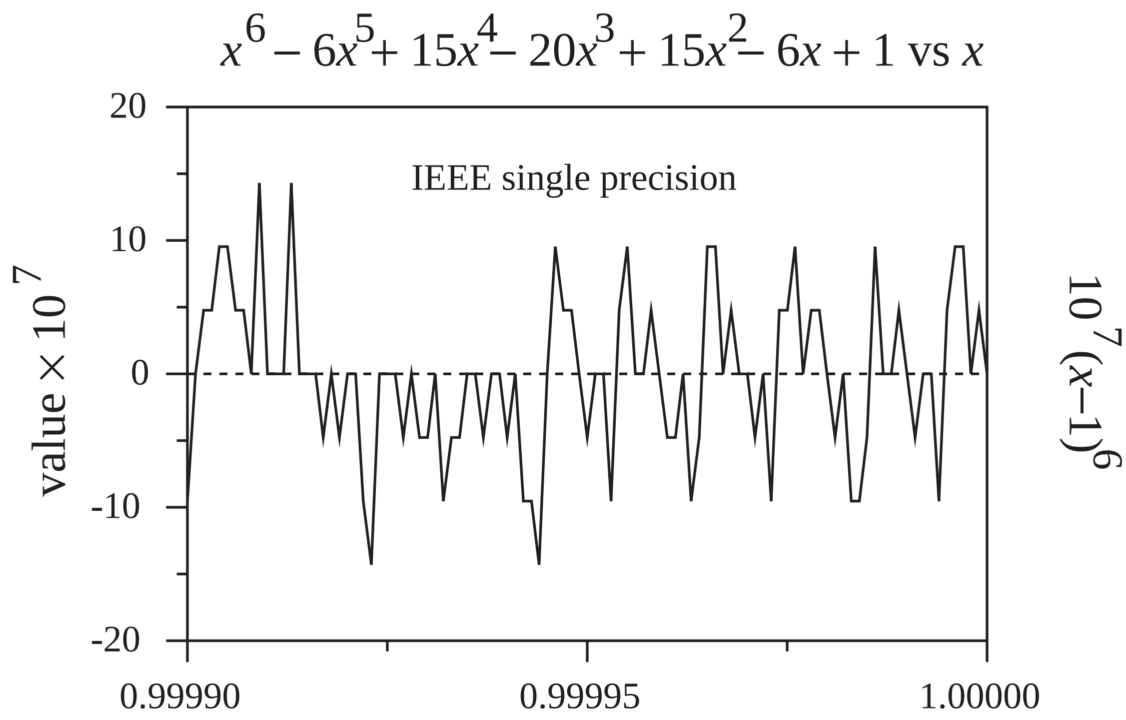
KAHAN'S SUMMATION METHOD (2)

- Picture that shows what Kahan's algorithm does:

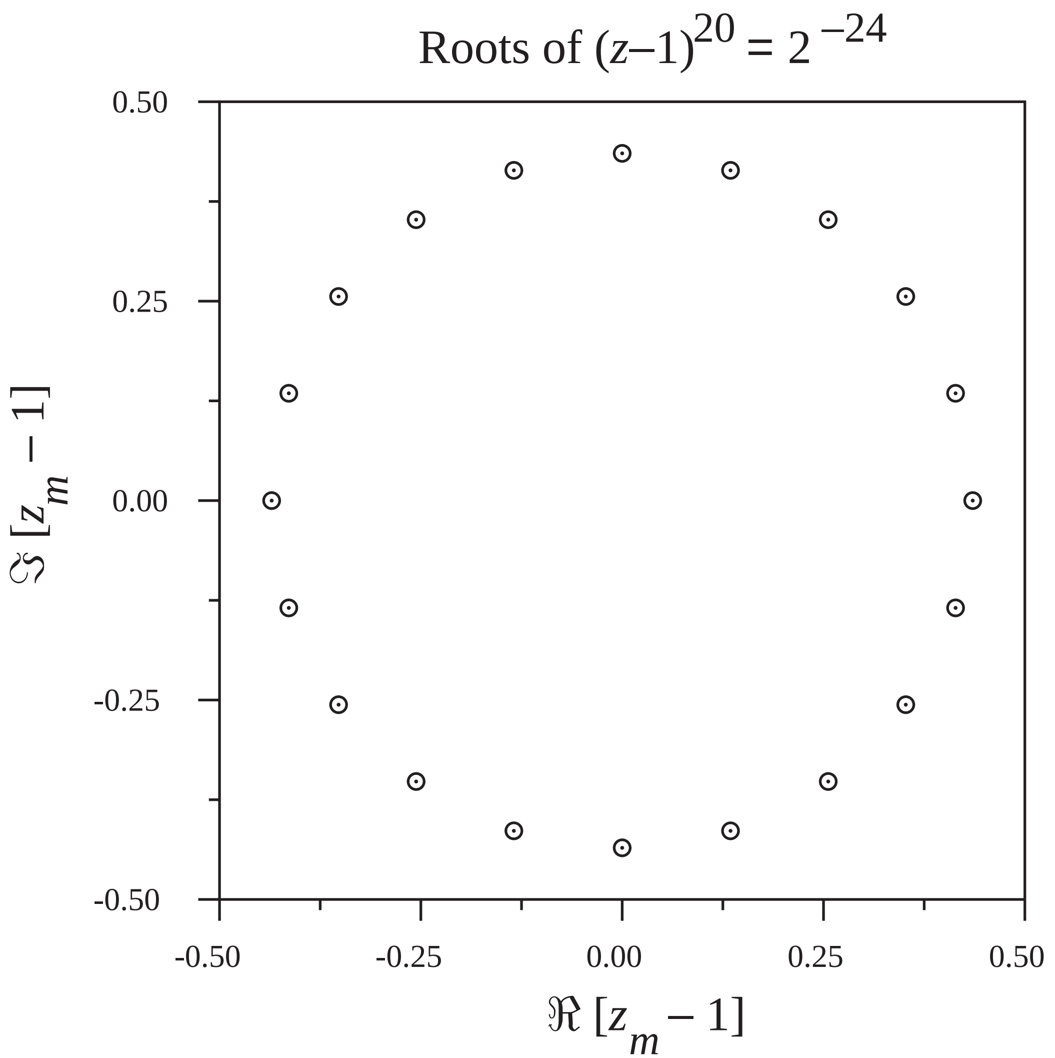


FLOATING-POINT COMPUTATION (3)

- Another consequence of catastrophic cancellation:
  Huge relative errors in evaluating a polynomial



ROOTS OF A POLYNOMIAL PERTURBED IN 1 LSB



FLOATING-POINT COMPUTATION (4)

- Yet another consequence of catastrophic cancellation:
  Ill-conditioned systems of linear equations

  Ill-conditioned systems often have high dimensionality, but here's an ill-conditioned system of 2 linear equations in 2 unknowns:

$$\mathbf{A}\mathbf{x} = \begin{pmatrix} 888,445 & 887,112 \\ 887,112 & 885,781 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

  The determinant of the coefficient matrix is 1

  But a hand calculator can't produce a correct solution!

  - The correct solution is
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 885,781 \\ -887,112 \end{pmatrix}$$
  - Using the division key on an HP-28 yields
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1,279,847.88527 \\ -1,281,771.0215 \end{pmatrix}$$

FLOATING-POINT COMPUTATION (5)

- Elementary analysis of Nievergelt's system of equations:

The coefficient matrix $\mathbf{A}$ is of the form

$$\mathbf{A} = \begin{pmatrix} \frac{1}{2}\alpha^2 + \alpha + 1 & \frac{1}{2}\alpha^2 \\ \frac{1}{2}\alpha^2 & \frac{1}{2}\alpha^2 - \alpha + 1 \end{pmatrix}$$

where $\alpha = 1332$

$\det[\mathbf{A}] = (\frac{1}{2}\alpha^2 + 1)^2 - \alpha^2 - (\frac{1}{2}\alpha^2)^2 = 1$

The inverse of $\mathbf{A}$ is

$$\mathbf{A}^{-1} = \begin{pmatrix} \frac{1}{2}\alpha^2 - \alpha + 1 & -\frac{1}{2}\alpha^2 \\ -\frac{1}{2}\alpha^2 & \frac{1}{2}\alpha^2 + \alpha + 1 \end{pmatrix}$$

The correct solution to the problem $\mathbf{Ax} = (1, 0)^T$ is

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\alpha^2 - \alpha + 1 \\ -\frac{1}{2}\alpha^2 \end{pmatrix}$$