

Floating Point Arithmetic

OR

YOU CAN'T ALWAYS COUNT ON YOUR COMPUTER

Derek O'Connor

University College, Dublin

March 4, 2005

1 LINKS FOR FLOATING POINT ARITHMETIC

Here are some links to people and papers mentioned in the talk above.

1.1 People

1. Richard Brent, Professor of Computing Science, Fellow of St Hugh's College, Oxford.
<http://web.comlab.ox.ac.uk/oucl/people/richard.brent.html>
2. James Demmel, Professor of Mathematics and Computer Science, University of California at Berkeley.
<http://www.cs.berkeley.edu/~demmel/>
3. Richard Fateman, Professor of Computer Science, University of California at Berkeley
<http://http.cs.berkeley.edu/~fateman/>
4. Gene Golub, Professor of Computer Science, Stanford.
<http://sccm.stanford.edu/faculty/nf-golub.html>
5. Nicholas Higham, Professor of Applied Mathematics, University of Manchester.
<http://www.ma.man.ac.uk/~higham/>
6. William Kahan, Professor of Computer Science, University of California at Berkeley.
<http://www.cs.berkeley.edu/~wkahan/>
7. G. W. (Pete) Stewart, Professor of Computer Science, University of Maryland.
<http://www.cs.umd.edu/~stewart/>
8. L. N. Trefethen, Professor of Numerical Analysis, Oxford University .
<http://web.comlab.ox.ac.uk/oucl/work/nick.trefethen/>
9. Charles F. Van Loan, Professor of Engineering, Department of Computer Science, Cornell University.
<http://www.cs.cornell.edu/cv/default.htm>

1.2 Web Links

- 1. IEEE Standard.**
<http://grouper.ieee.org/groups/754/>
Kahan's paper on IEEE Standard
<http://www.cs.berkeley.edu/~wkahan/ieee754status/why-ieee.pdf>
- 2. Decimal Arithmetic.** This is a proposal for a decimal IEEE standard.
<http://www2.hursley.ibm.com/decimal/>
- 3. Paranoia Test.** This is Kahan's test (written just after IBM PCs came out) to see if the machine-compiler system does IEEE arithmetic correctly. There are versions for Basic, C, Fortran, Modula, and Pascal available below.
<http://www.netlib.org/paranoia/index.html>
- 4. Round-off Error.** These are the latest papers and slides by Kahan on the difficulty of assessing round-off error.
These are slides that give the flavour of Kahan's style. A great rant.
<http://www.cs.berkeley.edu/~wkahan/Mindless1.pdf>
This is a long paper on which the slides above are based
<http://www.cs.berkeley.edu/~wkahan/Mindless1.pdf>
The need for extra precision
<http://www.cs.berkeley.edu/~wkahan/Qdrtcs.pdf>
- 5. Java Programmers Beware.** Here is Kahan on language-compiler systems that do not adhere to the IEEE standard.
<http://www.cs.berkeley.edu/~wkahan/JAVAhurt.pdf>
- 6. IEEE Floating Point Arithmetic.** This is an excellent introduction. It is now part of Overton's book.
<http://www.derekroconnor.net/NA/Readings/Overton--IEEE2up.pdf>
- 7. Calculating Elementary Functions.** This paper by Ng shows why computing $\sin(x)$ or $\cos(x)$ is difficult.
<http://www.derekroconnor.net/NA/Readings/Ng--TrigArgReduce2up.pdf>
Here are my (naïve) efforts at computing $\sin(x)$.
<http://www.derekroconnor.net/NA/LE/LE-2004-1-sol.pdf>
- 8. Cleve Moler's Book.** *Numerical Computing with Matlab*, SIAM, 2004. Moler developed Matlab around 1980 and knows what he is talking about. This is an excellent introductory textbook which is downloadable free from the link below. Be sure to get the Matlab programs that come with the book. (ncm.zip)
<http://www.mathworks.com/moler/>
- 9. Interactive Modules for FP Arithmetic.** The following modules illustrate the structure and behavior of finite-precision, floating-point number systems. Other modules at this site may be of interest.
http://www.cse.uiuc.edu/eot/modules/floating_point/
- 10. Spreadsheet Problems.** In the hierarchy of numerical software, spreadsheets should be viewed as executive toys. The site below discusses the numerical, statistical, and other problems with popular spreadsheets. Here is a quotation from this site :

I know there are many spreadsheets in financial companies that take all night to compute. These are complicated and commonly fail. When such spreadsheets are replaced by code more suited to the task, it is not unusual for the computation time to be cut to a few minutes and the process much easier to understand.

http://www.burns-stat.com/pages/Tutor/spreadsheet_addiction.html

2 TEXTBOOKS

2.1 Modern Texts

1. Coleman, T.F., and Van Loan, C. : *Handbook for Matrix Computations*, SIAM, 1988.
2. Demmel, James W. : *Applied Numerical Linear Algebra*, SIAM, 1997
3. Gill, P., Murray, W., and Wright, M : *Numerical Linear Algebra and Optimization, Vol 1*, Addison-Wesley, 1991.
4. Golub, Gene H, and Van Loan, Charles F. : *Matrix Computations 3rd Ed*, The Johns Hopkins University Press, 1996.
5. Hager, William W. : *Applied Numerical Linear Algebra*, Prentice-Hall, 1988. (Out of print by P-H, but available from the author)
6. Higham, N.J. : *Accuracy and Stability of Numerical Algorithms 2nd Ed*, SIAM, 2002.
7. Kahaner, D., Moler, C., and Nash, S. : *Numerical Methods and Software*, Prentice-Hall, 1989.
8. Moler, Cleve : *Numerical Computing with MATLAB*, SIAM, 2004. This book is available in PDF form at www.mathworks.com/moler/ (Moler is the inventor of MATLAB).
9. Stewart, G.W. : *Afternotes on Numerical Analysis*, SIAM, 1996.
10. Stewart, G.W. : *Matrix Computations, Volume 1 : Basic Decompositions*, SIAM, 1998.
11. Trefethen, Lloyd, N. and Bau III, David, *Numerical Linear Algebra*, SIAM, 1997.
12. Van Loan, Charles , F : *Introduction to Scientific Computing : A Matrix-Vector Approach using Matlab*, Prentice-Hall, 2000.
13. Watkins, D.S. : *Fundamentals of Matrix Computations*, Wiley, 1991.

2.2 Classic Texts

1. Brent, Richard P. : *Algorithms for Minimization without Derivatives*, Prentice-Hall, 1973.
2. Davis, Philip J. : *Interpolation and Approximation*, Blaisdell, 1963, (Reprinted by Dover, 1975).
3. Demidovich, B.P., & Maron, I.A. : *Computational Mathematics*, MIR Publishers, Moscow, 1973 (Translated and revised from 1970 Russian edition).
4. Forsythe, G.E. and Moler C.B. : *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, 1967.
5. Forsythe, G.E., Malcolm, M., and Moler, C.B. : *Computer Methods for Mathematical Computations*, Prentice-Hall, 1977.
6. Householder, Alston S. : *Principles of Numerical Analysis*, McGraw-Hill, 1953, (Reprinted by Dover, 1974).
7. Householder, Alston S. : *The Theory of Matrices in Numerical Analysis*, Blaisdell, 1964, (Reprinted by Dover, 1975).
8. Lanczos, Cornelius. *Applied Analysis*, Prentice-Hall, 1956. (Reprinted by Dover, 1988).
9. Varga, Richard S. : *Matrix Iterative Analysis*, Prentice-Hall, 1962.
10. Whittaker, Sir E.T. and Robinson, G. : *The Calculus of Observations : An Introduction to Numerical Analysis*, Dover, 1967, a republication of the 4th ed., 1944. First edition, Blackie & Sons, 1924. From the Preface of the First Edition, 1924:

[Numerical Analysis] is now included in the syllabus for the Open Competitive Examination for appointments in the Home and Indian Civil Services, the Colonial Service, etc. The present volume represents courses of lectures given at different times during the years 1913–1923 by Professor Whittaker to undergraduate and graduate students in the Mathematical Laboratory of the University of Edinburgh, ... etc.
11. Wilkinson, James H. : *Rounding Errors in Algebraic Processes*, Notes on Applied Science 32, Her Majesty's Stationary Office, 1963. Also Prentice-Hall, 1963.
12. Wilkinson, James H. : *The Algebraic Eigenvalue Problem*, Oxford University Press, 1965.

3 PAPERS

3.1 Online

3.2 Offline

3.3 Books