Lecture Notes in Computer Science

Edited by G. Goos and J. Hartmanis

235

Accurate Scientific Computations

Symposium, Bad Neuenahr, FRG, March 12–14, 1985
Proceedings

Edited by Willard L. Miranker and Richard A. Toupin

Springer-Verlag
Berlin Heidelberg New York London Paris Tokyo

Editors

Willard L. Miranker
Mathematical Sciences Department, IBM Research Center
Yorktown Heights, N.Y. 10598, USA

Richard A. Toupin
Department of Mechanical Engineering
Division of Applied Mechanics, Stanford University
Stanford, CA 94305, USA

CR Subject Classifications (1985): G.1, G.4, I.1

ISBN 3-540-16798-6 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-16798-6 Springer-Verlag New York Berlin Heidelberg

Library of Congress Cataloging-in-Publication Data. Accurate scientific computations. (Lecture notes in computer science; 235) 1. Mathematics--Data processing--Congresses. 2. Numerical calculations--Congresses. I. Miranker, Willard L. II. Toupin, Richard A., 1926-. III. Series.
QA76.95.A23 1986 510'.28'5 86-20364
ISBN 0-387-16798-6 (U.S.)

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2145/3140-543210

Accurate Scientific Computations

Preface

The theme of the symposium is the "accuracy" of certain kinds of mathematical results obtained with the aid of a computing machine. The symposium is entitled, "Accurate Scientific Computations", even though, as remarked by Stepleman in his introduction to Scientific Computing [1] , "when different people use the words scientific computing, it is not only possible but probable that each has in mind something different."

No less than in mathematics, physics, chemistry, or any other branch of science, "scientific computing" cannot be defined independently of examples. This symposium brings together three quite different kinds of work, concepts of accuracy, and notions of scientific computation. A shared aspect of the work in the 12 papers presented at the symposium (9 of which are collected here), and its panel discussion, is the use of present day computing machines to address mathematical problems and questions. We are careful here to avoid using the term "numerical questions" so as not to exclude one of the three kinds of work represented in these papers; viz., Computer Algebra.

An alternative title for this symposium might be **Applications of Computing Machines in Mathematics.** Computing machines have come to be widely used as instruments of simulation and empiricism in what today is called "Scientific Computing". Important and useful as these applications of computers are in the various sciences and fields of engineering, they were not the dominant theme of this symposium. Rather it was algorithms which deliver precise results, both analytic and numerical. To express an indefinite integral of a rational function of elementary functions as a similar object, if and when it

---

1 North-Holland Publishing Co. (1983)

exists, is an example of the former. An algorithm which computes an error bound is an example of the latter. Another example of the latter is an algorithm which computes the probability that a real number lies in a prescribed interval. Some of the papers deal also with the efficiency of the implementations of such algorithms.

Scientific Computing has come to mean more narrowly the construction of solutions, or approximations of solutions of systems of differential or algebraic equations, or other constructive, finite, algorithmic processes of algebra and analysis. If we combine this narrower definition of "Scientific Computing" with the quest for mathematical theorems strictly proven, or computation automatically validated with the aid of a computing machine we arrive at the title of the symposium and a unifying concept for the results presented in the papers collected here. They address the idea of **Accurate Scientific Computation** in three quite different ways which we can illustrate with the important special and pervasive case of a "problem" in "Scientific Computing"; viz., "solving" a system of linear equations $Ax = B$.

To embrace all three concepts of accuracy in one simple and familiar example, we must narrow the problem even further and consider the case when the coefficient matrix A and the vector B are prescribed rational numbers with numerators and denominators of reasonable length. In this case, if the system is consistent, there exist rational solutions $x = (x^1, ..., x^n)$ and algorithms to compute each and every rational number x^i, $i = 1, 2, ..., n$. If the size of the system is not too large, it is a feasible task to compute and display the numerator and denominator of each and every x^i. A computer algebra "system" might implement such a display. It is one concept of an "accurate scientific computation". Of course, if the dimension of the system exceeds ten or twenty, then, in general, the numerators and denominators in this definition and representation of the "solution" may be very large integers indeed. The computation may be rather extensive and time consuming even on a large computer. But when A and B have small representations and the dimension of the linear system is small, there could be useful insight and purpose in this sort of accurate scientific computation. In particular, the precise integer rank of the matrix A could be determined in this way.

A second definition of the problem of "solving" the same system of linear equations $Ax = B$ is to construct (compute) a **floating-point** or other approximation $\tilde{x}$ to the rational solution x of the system if it be consistent, **and** to compute an upper bound on some norm of the difference $|x - \tilde{x}|$, **and** to require that this bound on the "error" of $\tilde{x}$ be less than some prescribed value. This approach is termed **validated computation.**

A third definition of the same "problem" is to compute a floating-point or other approximation $\tilde{x}$ to the same system of equations, **and** to compute (exhibit) a lower bound on the **probability** that the difference $|x - \tilde{x}|$ be not greater than some prescribed value.

Thus we have before us at least three quite different concepts of **Accurate Scientific Computing**, each of which is represented in the lectures and results collected here.

Basic to scientific computation is the evaluation of the elementary functions. In separate lectures by S. Gal and by F. Gustavson (abstract only) methods are described for computing very accurate values of the elementary scalar functions (sin, cos, log, square root, etc.) which, at the same time, are very fast. The speed and efficiency of the new algorithms exploit the architectural changes in computing machines which have occurred in the last two decades since the pioneering work of Kuki on this problem. Moreover, the new algorithms bound the relative error of the computed value of the function for all values of the argument. The bound guarantees the accuracy or significance of all but the last digit in the function value, and even the last for more than 99.9% of the arguments.

A review of concepts and results of a new and systematic theory of computer arithmetic is presented by U. Kulisch. The new arithmetic broadens the arithmetical base of numerical analysis considerably. In addition to the usual floating-point operations, the new arithmetic provides the arithmetic operations in the linear spaces and their interval correspondents, which are most commonly used in computation, with maximum accuracy on computers directly. This changes the interplay between computation and numerical analysis in a qualitative way. Floating-point arithmetic is defined concisely and axiomati-

cally. The subsequent lectures by S. Rump, W. Ames, W. L. Miranker, and F. Stummel show aspects of this.

New computational methods to deal with the limitations inherent in floating-point arithmetic are presented by S. Rump. The mathematical basis is an inclusion theory, the assumptions of which can be verified by a digital computation. For this verification the new well-defined computer arithmetic is used. The algorithms based on the inclusion theory have the following properties:

- results are automatically verified to be correct, or when a rare exception occurs, an error message is delivered.
- the results are of high accuracy; the error of every component of the result is of the magnitude of the relative rounding error unit.
- the solution of the given problem is imputed to exist and to be unique within the computed error bounds.
- the computing time is of the same order as a comparable (purely) floating-point algorithm which does not provide these features.

The approach has thus far been developed for some standard problems of numerical analysis such as systems of linear and non-linear equations, eigenproblems, zeros of polynomials, and linear and convex programming. When data of a given problem is specified with tolerances, every problem included within the tolerances is solved and an inclusion of its solution is computed. The key property of the algorithms is that the error is controlled automatically. These concepts and this "validation" approach to scientific computation are collected in a subroutine library called ACRITH. The presentations of W. Ames and of W. Miranker develop other possibilities for the exploitation of ACRITH. In the latter presentation, methods for directly exploiting the new computer arithmetic are given as well.

W. Ames describes software for solving the finite difference equations corresponding to boundary value problems of elliptic partial differential equations.

The routines, programmed in VS Fortran, employ the ACRITH Subroutine Library, and provide the user a choice of any one of eleven classical algorithms for solving a system of linear equations. Each algorithm can be executed with traditional computer arithmetic, or with ACRITH. This permits the user to observe the advantages of using ACRITH. Illustrative data is presented.

W. Miranker shows that good arithmetic can improve algorithmic performance. Compared to results obtained with conventional floating-point arithmetic, the computations are either more accurate or, for a given accuracy, the algorithms converge in fewer steps to within the specified error tolerance. Two approaches are presented. First: the high performance linear system solver of ACRITH is used in the areas of regularization (harmonic continuation) and stiff ordinary differential equations. Second: the routine use of a highly accurate inner product (a basic constituent of the new floating-point arithmetic) is shown to result in acceleration of eigenelement calculations (QR-algorithm), the conjugate gradient algorithm and a separating hyperplane algorithm (pattern recognition). Not all algorithms are susceptable of improvment by such means and some speculations are offered.

Schauer and Toupin present a method for computing a bound on the error of an approximation to the solution of a restricted class of systems of linear equations. These systems include those arising from discretization of certain boundary-value problems of elliptic partial differential equations. They also present empirical evidence for the existence of a critical precision P(A) of floating-point arithmetic used in the conjugate gradient algorithm for constructing an approximation to the solution of a system (A) of linear equations. The critical precision P(A) depends on the system. If the precision of the floating-point arithmetic is less that P(A), then the residual fails to diminish monotonically as it would were the precision infinite. If the precision of the arithmetic exceeds P(A), they observe that the approximate residual diminishes montonically to zero in a number of "steps" of the algorithm not greater that the dimension of the system (were the precision infinite, the number of "steps" would be the number of distinct eigenvalues of the matrix (A)). Moreover, for each digit of precision in excess of P(A), one more significant digit in the approximate solution is obtained. For the large sparse systems investigated, the number of steps is a small fraction of the dimension if the

precision is greater than P(A). The critical precision P(A) is an emprically determined "condition number" for a system of linear equations. It may be less than or greater than any of the precisions provided by the floating-point arithmetic units of a particular machine.

F. Stummel presents a new method for the derivation of exact representations of errors and residuals of the computed solutions of linear algebraic systems under data perturbations and rounding errors of floating-point arithmetic. These representations yield both strict and first-order optimal componentwise a posteriori error and residual bounds which can be computed numerically together with the solutions of the linear systems. Numerical examples of large linear systems arising in difference approximations of elliptic boundary value problems, in finite element and boundary integral methods show that the bounds so obtained constitute realistic measures of the possible maximal errors and residuals.

Some problems of algebra and analysis, such as obtaining explicit formulas for the derivative or integral of special classes of functions are finite computational tasks. One objective of "computer" algebra is to discover and implement such algorithms. Approximations to real numbers, such as provided by floating-point arithmetic, runs counter to the spirit of this work in computer algebra or "symbol manipulation". On the other hand, the results delivered by these algorithms, though finite, may be bewilderingly long. Taking integration as an example, J. Davenport shows how results of such computer algebra systems might be combined with numerical integration schemes to speed and enhance the accuracy of the latter.

The two lectures of B. Trager and G. E. Collins (abstracts only) also concern finite computational tasks in algebra for which no approximations to real numbers are invoked.

The problem of computing a bound on the error of an approximate solution of a system of linear equations, the value of an elementary function, or the root of a polynomial using only finite approximations to real numbers is not a trivial one, as placed in evidence by several of the papers presented at the symposium. What can one hope to do with the same question if applied to

the floating-point approximations to solutions of large systems of non-linear equations in many variables such as those computed daily by the national and international weather bureaus? Indeed, it would seem a hopeless task if approached in the same spirit and with the same ideas that have been found effective for the elementary and fundamental sub-tasks of such large and complex computations involving billions of round-off errors. R. Alt and J. Vignes present an alternative question and means to address it. They replace the problem of computing error bounds on approximations by the problem of computing probabilistic estimates of the error of an approximation. In practice, their approach resembles the familiar scheme of computing two approximations with a large and complicated program; one using floating-point arithmetic with double the precision of the other. One gains some confidence, in this way, with the significance of common high-order digits in the two approximations. Alt and Vignes propose that one perturb the program and intermediate floating-point results in such large and complex computations in a systematic way, and infer the probability that the common leading digits of a small sample of approximations computed in this way are significant.

These lectures and the exchange of views of the panelists and participants during the panel discussion point to a continuing evolution and broadening of the concepts, objectives, and methods of Scientific Computation. The papers collected here provide evidence of the interplay between the discovery of algorithms for new and old mathematical tasks and the evolution of computer architectures. The theme of the work presented in these papers is "accuracy"; different concepts and definitions of it, ways to achieve it efficiently, and algorithms to prove or "validate" it. We foresee a gradual evolution of the objectives of Scientific Computing wherein the quest for "accuracy" competes in a more balanced way with the quest for "speed". We believe that the concepts, results, and methods described in the papers of this symposium will seed and influence such an evolution of the subject. In summary, these are:

- Efficient algorithms for evaluation of elementary functions having specified and guaranteed accuracy based on the non-standard Accurate Tables Method.

- Axioms for "computer arithmetic", including directed roundings (interval arithmetic).

- The theory of and techniques for computing inclusions.

- Computer architectures which implement essential primitives for achieving accuracy and proving it, such as high precision inner products, and variable precision floating-point arithmetic.

- Probabilistic algorithms for estimating the accuracy of complex and extensive floating-point computations.

- A synergism of the concepts and methods of "computer" algebra (exact) computations, and those which invoke approximations to real numbers and functions of them.

Yorktown Heights, NY — *W. L. Miranker*

Heidelberg, FRG — *R. A. Toupin*

Acknowledgements

The symposium on **Accurate Scientific Computations** was held March 12-14, 1985 at the Steigenberger Kurhotel, Bad Neuenahr, Federal Republic of Germany.

It was sponsored and organized by IBM Deutschland GmbH Scientific Programs, Prof. Dr. C. Hackl, with the assistance of Ms. E. Rohm.

The Scientific Chairman was Dr. S. Winograd, Director Mathematical Sciences Department, IBM Research Center, Yorktown Heights, N.Y. Dr. H. Bleher, Dr. W. L. Miranker, and Dr. R. A. Toupin were associate organizers. The sessions were chaired by Dr. S. Winograd, Prof. Dr. L. Collatz, Prof. Dr. R. Loos, and Dr. A. Blaser.

There was a panel discussion chaired by Prof. Dr. P. Henrici. The panel members were Prof. F. W. J. Olver, Prof. Dr. H. J. Stetter, and Prof. Dr. H. Werner.

Table of Contents

Computing Elementary Functions:
A New Approach for Achieving High Accuracy and Good Performance
S. Gal 1

Fast Elementary Function Algorithms for 370 Machines (Abstract)
F. G. Gustavson 17

A New Arithmetic for Scientific Computation
U. Kulisch 18

New Results on Verified Inclusions
S. M. Rump 31

Accurate Elliptic Differential Equation Solver
W. F. Ames and R. C. Nicklas 70

Case Studies for Augmented Floating-Point
W. L. Miranker, M. Mascagni and S. Rump 86

Strict Optimal Error and Residual Estimates for the Solution of Linear Algebraic Systems by Elimination Methods in High-Accuracy Arithmetic
F. Stummel 119

Solving Large Sparse Linear Systems with Guaranteed Accuracy
U. Schauer and R. A. Toupin 142

Symbolic and Numeric Manipulation of Integrals
J. H. Davenport 168

Computer Algebra and Exact Solutions to Systems of Polynomial Equations (Abstract)
B. M. Trager 181

The Euclidean Algorithm for Gaussian Integers (Abstract)
G. E. Collins 182

An Efficient Stochastic Method for Round-Off Error Analysis
J. Vignes and R. Alt 183