JUL 28 1965

LA-RC-7192

CONF-650505-4
MASTER

IFIP CONGRESS 65

ANALYSIS OF INHERENT ERRORS IN MATRIX DECOMPOSITION
USING UNNORMALIZED ARITHMETIC

N. Metropolis*

Los Alamos Scientific Laboratory
Los Alamos, New Mexico (U.S.A.)

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:
A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or
B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.
If a person acting on behalf of the Commission is employed by the Commission, in whole or in part, the use of any information, apparatus, method, or process disclosed in this report, in whole or in part, is authorized by the Commission, and the Commission is not liable for any damages resulting from the use of any information, apparatus, method, or process disclosed in this report. In no event shall the Commission be liable for any damages resulting from the use of any information, apparatus, method, or process disclosed in this report, in whole or in part, by any person other than the Commission, its employees, or contractors of the Commission, or employees of such contractors, or any person who has been granted access to, or provides access to, any information pursuant to his employment or contract with the Commission, or provides access to, any information pursuant to his employment or contract with the Commission.

Introduction: The propagation of inherent errors in numerical calculation is a central problem of error analysis. Its study is essential in any approach to an error-monitoring scheme. The realization of such a scheme can be accomplished in two stages: the first is one of logical design of the arithmetic processor with sufficient flexibility such that the result of any basic arithmetic operation ($\pm$, $\times$, $+$) reflects the precision of the participating operands. An unnormalized representation of computer numbers with an appropriate set of adjustment rules (cf. below) for these operations provides a simple and direct approach [1]. The second stage is one of analysis. For each algorithm of interest, the appropriate one must be selected from the set of mathematically equivalent algorithms. This selection must (1) take correlative effects into account, (2) allow varying precision and magnitude of input numbers, (3) provide optimal precision of output numbers, (4) permit an immediate interpretation of such precision, namely by mere inspection of the results.

*Permanent address: University of Chicago, Chicago, Illinois.

APPROVED FOR PUBLIC RELEASE

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

An algorithm is described for the reduction of a square matrix to triangular form; more specifically, a strategy is given for the selection of pivot elements for each stage of reduction by Gaussian elimination.

Summary of unnormalized arithmetic. Representation of numbers in an unnormalized format provides a simple means of specifying the precision of initial operands. A number x is represented in binary form as $x = 2^e \cdot f = (e, f)$ where the exponent e is an integer and the fractional part f satisfies $0 \leq |f| < 1$. The operation $(e+s, 2^{-s}f) \rightarrow (e, f)$ is called adjustment. The precision of each input operand is reflected in its particular adjustment by arranging that the least significant bit resides in some specified position, say k , (at the option of the programmer) within the format of available bits, say p . That is, if $x^T = (e, f^T)$ is the true value of $x = (e, f)$, then

$$\frac{1}{2} \cdot 2^{-(k+1)} \leq |f - f^T| < 2^{-(k+1)} \quad (1)$$

is satisfied. The minimum choice of k is such that it accommodates the most precise input of interest; once specified, k is regarded as fixed throughout a given procedure. If $k < p$, the remaining $p-k$ bits on the right serve as guard bits to separate the effect of rounding errors from those of inherent errors.

The number of leading digits m in a fractional part is defined by $\frac{1}{2} \leq 2^m |f| < 1$.

Let $(e_1, f_1) * (e_2, f_2) \rightarrow (e_0, f_0)$ represent symbolically the four basic arithmetic operations; an m-value is associated with each f. The adjustment rule for addition and subtraction is $e_0 = \max(e_1, e_2)$. (If f_0 overflows, a subsequent adjustment by one is made.) The corresponding rule for multiplication and division is $m_0 = \max(m_1, m_2)$; for details cf. [2].

Strategy for the selection of pivot elements. The reduction of a square matrix of order n to triangular form is achieved in n-1 stages. The simplest approach takes the current diagonal element a_{rr} in the r^{th} step ($r = 1, 2, \dots, n-1$) as the pivot element and transforms the eligible elements a_{ij} ($r < i, j \leq n$) according to

$$a_{ij} - \frac{a_{ir}}{a_{rr}} a_{rj} \rightarrow a_{ij}. \quad (2)$$

It is advantageous in minimizing rounding errors to adopt the strategy of selecting as a pivot for each stage the element of maximum magnitude in the r^{th} column--sometimes called partial pivoting [3]. However, if the eligible elements at successive stages are not represented with the same precision (the fact that the original elements of the matrix may have equal precision is irrelevant), the question arises whether the maximal element is still the best strategy. Consider an instance where the maximal element is much less precisely known than, say, the next largest in magnitude. The advantage of greatest magnitude can be dissipated by the loss of precision inflicted upon the transformed elements as a result of

selecting a relatively imprecisely known element as pivot.

The consequence of selecting a pivot based on the most precisely represented element is also unsatisfactory. If the pivot thus selected has a relatively small magnitude, the transformed elements are amplified and in subsequent transformations they are deamplified by cancellation. The disturbing feature of this latter effect is that meaningful bits have been discarded.

It seems intuitively clear that a simple, uncorrelated scan of eligible elements can not lead to an optimal strategy, for the reduction of a matrix to triangular form is a global affair and correlations must be taken into account. Consistent with this notion we take the precision with which the associated determinant is evaluated as a measure of the strategy used in the triangularization process. The following strategy is proposed:

- (1) At the r^{th} stage of reduction, one seeks the r elements that solve the "assignment problem" for the matrix whose eligible elements are $z_{ij} = \log a_{ij}$; namely one selects an element from each eligible row and column exclusively such that their sum is maximal [4].
- (2) The element with maximal magnitude from among these selected r elements is taken as pivot.

Several remarks may be made: (a) the element of maximal magnitude is often, but not always a member of set that solves the assignment problem; (b) if a pivot has relatively low precision and reduces

the precision of the transformed elements, this loss is unavoidable in the sense that another strategy would do better; (c) often the set of elements, depleted by the selected pivot, is the solution of the assignment problem for the subsequent stage; this fact results in considerable simplification.

The following example illustrates a case where the maximal element is not selected as pivot and what happens if it is taken as pivot. The exponents μ_{ij} are conveniently taken to be those corresponding to normalized fractional parts, i.e., $\mu_{ij} = e_{ij} - m_{ij}$. They have been so chosen that the fractional parts f_{ij} may be omitted for the present purpose. The m -values are relevant and are included. A column interchange has been made for the second case to have pivot elements on the diagonal.

$$x_{ij} = (\mu_{ij}, m_{ij})$$

Pivot strategy: assignment problem

$$\begin{pmatrix} 8,2 & 9,6 & 3,2 \\ 1,0 & 7,3 & 5,3 \\ 2,2 & 2,1 & 6,3 \end{pmatrix} \rightarrow \begin{pmatrix} 8,2 & 9,6 & 3,2 \\ 0 & 7,3 & 5,3 \\ 0 & 0 & 6,3 \end{pmatrix}$$

det = (21,3)

Pivot strategy: maximum magnitude

$$\begin{pmatrix} 9,6 & 8,2 & 3,2 \\ 7,3 & 1,0 & 5,3 \\ 2,1 & 2,2 & 6,3 \end{pmatrix} \rightarrow \begin{pmatrix} 9,6 & 8,2 & 3,2 \\ 0 & 6,6 & 5,3 \\ 0 & 0 & 6,3 \end{pmatrix}$$

det = (21,6)

REFERENCES

- [1] N. Metropolis and R. L. Ashenhurst, Significant digit computer arithmetic, IRE Transactions on Electronic Computers, vol. EC-7, No. 4 (Dec. 1958), 265-267.
- [2] R. L. Ashenhurst and N. Metropolis, Unnormalized floating-point arithmetic, J. Assoc. Comput. Mach. no. 3, 6 (July 1959) 415-28.
- [3] Wilkinson, J. H., Rounding Errors in Algebraic Processes, Prentice-Hall (1963).
- [4] Munkres, J., Algorithms for the assignment and transportation problems. J. Soc. Indust. Appl. Math. 5, 32 (1957).

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.