ERROR - FREE HESSENBERG TRANSFORMATION AND SYMMETRIZATION OF MATRICES

M. Sc. Thesis

by

JOSEPH SUNDAY AINA B.Sc. (Hons) A.B.U.

Ahmadu Bello University
Zaria

March 1976

DEDICATED TO

My Father Mathew A. Aina.
My Mother Abigail Ibironke Aina
and to my wife
Elizabeth Ibitayo Aina.

ACKNOWLEDGMENTS

The author wishes to express his profound gratitude to Dr. B. N. Datta his former supervisor who introduced him to this area of research and who also introduced him to Professor E.V. Krishnamurthy. He is greatly indebted to Professor E.V. Krishnamurthy, visiting Professor, Department of Computer Science, University of Lagos, who had given him the necessary attention to make this thesis a success.

He is also grateful to Professor Iya Abubakar, Vice-Chancellor of Ahmadu Bello University and Dr. A.R. Kokan, Acting Head of Mathematics Department, Ahmadu Bello University who made arrangements for him to complete this work at the Lagos University.

He also wishes to thank a number of his friends and colleagues, Messers J. Oladipo, M.S.O. Adewale, J. Oke, R. Adeseko, L. Awogbami and Mr and Mrs. W.A. Oni who provided him the necessary comfort in Lagos.

He wishes to express his gratitude also to the Department of Computer Science of the University of Lagos.

His thanks are due to the Polytechnic, Ibadan which provided the fund for him to carry out the research.

Finally, his appreciation to his wife who had carried out the family responsibilities while he has been away.

Last but not the least, to Mr. M. A. Olaniran who did an excellent typing of this work.

TABLE OF CONTENTS

DEDICATION ---------- ii

ACKNOWLEDGEMENTS ---------- iii

NOMENCLATURE AND DEFINITIONS ---------- vi

INTRODUCTION ---------- 1

CHAPTER 1: REDUCTION OF MATRIX TO THE HESSENBERG FORM AND ITS APPLICATIONS ---------- 6

1.1 Various methods for reduction to the Hessenberg form ---------- 6

1.2 Reduction of a matrix to a Hessenberg form ---------- 6

1.3 Applications of Hessenberg Matrices ---------- 13

CHAPTER 2: SYMMETRIZATION OF MATRIZES AND ITS APPLICATIONS ---------- 15

2.1 Symmetrizers ---------- 15

2.2 Applications ---------- 16

2.3 Methods for the construction of symmetrizers ---------- 20

CHAPTER 3: RESIDUE ARITHMETIC APPLICATIONS TO EXACT'S COMPUTATION OF A SYMMETRIZER AND TRANSFORMATION TO THE HESSENBERG FORM ---------- 28

3.1 Residue Arithmetic with a single modulus ---------- 28

3.2 Residue Arithmetic with multiple moduli ---------- 34

3.3 Applications to symmetrizers and Reduction of matrix to Hessenberg form ---------- 36

CHAPTER 4: P-ADIC ARITHMETIC APPLICATIONS TO EXACT COMPUTATION OF A SYMMETRIZER AND TRANSFORMATION TO THE HESSENBERG FORM ----- 45

INTRODUCTION ---------- 45

4.1 Properties of H(p, r) Codes ------ 47

4.2 P-Adic Number Representation ----- 49

..../ v...

4.3 Arithmetic Algorithms ---------- 51

4.4 Conversion Algorithms ---------- 56

4.5 Application of P-adic arithmetic to symmetrizer and deduction of a matrix to the Hessenberg form ---- 61

* REMARK ---------- 67

CONCLUDING REMARKS ---------- 68

BIBLIOGRAPHY ---------- 70

NOMENCLATURE AND DEFINITIONS

The definition and notations used in this Thesis are given in the table below: The notations have the meaning given below unless otherwise specified)

Symbol	Meaning
Capital letters	
A, M, X, I	Matrices
A^T	Transpose of A
det(A)	Determinant of A
λ	The eigenvalue of a matrix
$\phi(\lambda)$	Characteristic equation
$\phi(A)$	Characteristic polynomial of A
m(A)	The minimum polynomial
I	Identity Matrix
A^{-1}	The Inverse of A
GF(P)	Galois Fields
$\|x\|_m$	Residue of x modulo m
$a \equiv b \pmod p$	a is congruent to b modulo p
$\|m\|$	Absolute value of m
a^{-1}	The multiplicative inverse of a
GCD (a, b)	Greatest Common divisor of a and b
LCM (a, b)	Least Common Multiple of a and b
$\sim$	Equivalent to
Σ	Summation
Π	Product
ε	Belongs to

Symbol	Meaning
=	Equal to
≠	Unequal to
i.e.	That is
∴	Therefore
α, β, γ	p-adic Numbers
H(p, r)	The Hensel Code with base p and length r
Δ (J)	Right shifts by j digits
viz.	Namely
e.g.	For example
No.	Number
+ve	Positive
-ve	Negative

I N T R O D U C T I O N

The symmetrization and Hessenberg transformation of matrices are of great importance in a variety of scientific applications Datta (1972), (1973), (1974) Datta and Datta (1975a), (1975b), Howland, (1961, (1962), Howland and Farrel (1963), Howland and Senez (1970), Olga Taussky and Zassenhaus (1959), Wilkinson (1965), (1958) These applications vary from solution of non-symmetric matrix equations to the study of undamped small vibrations of a mechanical system about a position of stable equilibrium.

It is also an accepted fact that when a matrix is reduced to an almost triangular form, called "Hessenberg form", the eigenvalue problem becomes simpler to handle and the computation is easier, Wilkinson (1958, 1965).

In order that the processes of reduction of a matrix to a Hessenberg form and symmetrization be numerically stable and free from errors , (round off, commulative or truncation errors), it is extremely important that error-free computation methods are adopted for the reduction of a matrix to the Hessenberg form and its symmetrization. The modular arithmetic has been chosen here for this purpose. (For relevant details see Krishnamurthy (1975), Mahadeva Rao (1975), Young and Gregory (1973).

This dissertation is divided into four chapters. Chapter 1 deals with the method of reducing any given matrix to a Hessenberg form and its applications.

.../2...

The method Wilkinson (1958), Krishnamurthy and Sen (1975) is as follows: Given a general matrix A it is pre and post multiplied by a sequence of elementary transformation matrices

$$H_1, H_2, \ldots\ldots\ldots H_k \text{ and } H_k^{-1} H_{k-1}^{-1}$$

Such that

$$H_k^{-1} \ldots\ldots H^{-1} A H_1 H_2 \ldots\ldots\ldots H_{k-1} H_k = S:$$

here S is a lower Hessenberg matrix for which

$$S_{ij} = 0 \; j \geq i + 2.$$

Chapter 2 deals with the construction and applications of symmetrizers. The symmetrizers have the following applications:

(i) Solving the non-symmetric eigenvalue problem $Ay = \lambda y$. The introduction of a symmetrizer X converts it to the form $E\lambda = \lambda Z$ (Datta (1973) Krishnamurthy and Sen (1975)

(ii) It was also found out by Howland and Senez(1970) that the knowledge of a symmetrizer can help in the solution of the Lyapunov matrix equation $XA + A^T X = - I$

(iii) Howland and Farrell (1963) also showed that Lanozos minimized iterations is also greatly simplified when we are able to symmetrize a matrix.

(v) Solution of the general characteristic equation $Ay = By$.

We then describe a way of constructing a symmetrizer for the companion matrix Datta (1972).

Also we describe an algorithm for computing a symmetrizer for a Hessenberg matrix (lower or upper) Datta (1973)

Also a method for constructing a positive definite symmetrizer for a matrix which is in the rational canonical form (Datta 1972) is described.

In Chapter 3 we describe;

(a) The principles and algorithms of residue Arithmetic.

(b) Its applications to the exact computation of the reduction of a matrix to a Hessenberg form and the symmetrization of the Hessenberg form.

The first part of chapter 3 explains the basic concepts of the residue arithmetic. In this approach every rational number a/b ($b \neq 0$ $0 \leq a \leq p-1$, $0 < b \leq p-1$) is uniquely identified with $a.b^{-1}$ where the binary operation, the product modulo a prime number p and b^{-1} is in the Galois Field with elements 0, 1,2 ..., p-1. In this procedure the magnitude of p restricts the range of uniquely representable rationals. It therefore implies the choice of a larger prime for increasing the range. This however, increases the labour of computation. In order to minimize the labour of computation, a multiple prime approach is suggested (Young and Gregory, 1973).

Here a modulus M which is a product of the distinct primes $P_1, P_2, \ldots\ldots, P_S$ is chosen. In using this procedure one has to be careful in the choice of M as well as the prime numbers to be used in the computation of a particular algorithm. The concept of common factor is introduced here so as to enable us convert back the results into rational numbers.

In the case of multiple-prime method the matrix computations are to be performed for each P_i and the results are to be combined using the Chinese Remainder's theorem.

Chapter 4 deals with

(a) P-adic Arithmetic

(b) Its applications to the exact computation of the algorithms as in Chapter 3.

As mentioned earlier, in the multiple prime method each algorithm has to be worked out for different choices of primes: the use of p-adic arithmetic eliminates the necessity of using the multiple prime method. In the p-adic arithmetic, each rational number a/b ($0 < a, b < N$ where $N = \sqrt{(p^r - 1)/2}$ is represented by an r-digit sequence with each digit α_i an integer such that $0 \leq \alpha_i \leq p-1$ where p is a prime number. The algorithms for conversion of a rational p-adic form and vice-versa are also described (Krishnamurthy et al (1975)

.../5...

The latter part of the chapter deals with the application of the p-adic arithmetic for the reduction of a given general matrix to a Hessenberg form and symmetrization.

An up-to-date bibilography is provided at the end.

.../6...

CHAPTER ONE

1.1 Various methods are available for the reduction of a given matrix to an almost or near triangular form a called the Hessenberg form.

These methods include:

(a) Reduction by a sequence of unitary transformations, a method usually referred to as the method of Givens Wilkinson (1958), (1965)

(b) Reduction to the Hessenberg form of a general matrix by means of similarity transformations which are derived as the product of a number of simple elementary similarity operations. Wilkinson (1958), (1965) Krishnamurthy and Sen (1975)

In this dissertation we will describe the second method in detail as given by Wilkinson (1958).

1.2 Reduction of a Matrix to a Hessenberg form

Given a general matrix A it is pre and post multipled by a sequence of elementary similarity transformation matrices $H_1, H_2, \ldots\ldots\ldots H_k$

thus: $H_k^{-1} H_{k-1}^{-1} \ldots\ldots H_2^{-1} H_1^{-1} A H_1 H_2 \ldots\ldots H_{k-1} H_k = S$ (1.2.1)

Such that S is in the lower Hessenberg form, viz. a matrix for which $S_{ij} = 0; \quad j \geq i + 2$ (1.2.2)

For this purpose two types of matrices are used:

The first type consists of permutation matrices denoted by $I^{(pq)}$ for which

.../7...

$I_{ii}(pq) = 1$, $i \neq p, q$; $I^{(pq)}_{pq} = I^{(pq)}_{qp} = 1$; $I^{(pq)}_{ij} = 0$

otherwise (1.2.3)

Multiplication on the right by $I^{(pq)}$ interchanges the columns p and q and multiplication on the left interchanges the rows p and q. We have $I^{(pp)} = I$ the identity matrix and $I^{(pq)} I^{(pq)} = I$ so that the matrix is its own inverse. The second type of matrix is denoted by

$M^{p}_{(ij)} = 1$, $M^{p}_{(p,j)} = m_j$ for $j > p$, (1.2.4)

Otherwise $M^{p}_{(i,j)} = 0$

All transformation $M^{(p)}_{(p,j)}$ simply regarded as $M^{(p)}$ which are used, have elements M_j satisfying

$$|M_j| \leq 1$$

The second type of matris is typified by the matrix $M^{(3)}_{(3j)}$ shown below for n = 6

$$M^{3}_{(3j)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & M_4 & M_5 & M_6 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

We now describe the transformation to a lower Hessenberg form. It will be illustrated by showing the configuration of zero elements in a matrix of order five at each stage, It is to be understood that a_{ij} refers to the current value of the element in the (ij)

.../8...

position after all the transformations described up to that point have been applied. The steps are as follows:

(1) Select the largest in the absolute magnitude of the (n-1) elements a_{12}, a_{13}, ... a_{1n}. Suppose it is $a_{1\alpha}$, then interchange columns 2 and α.

(2) Calculate the (n-2) quantities m_{2j} given by $$m_{2j} = -\frac{a_{1j}}{a_{12}} \qquad j = 3 \text{ to } n \qquad (1.2.5)$$ we have $|m_{2j}| \leq 1$

(3) Multiply on the right by $M^{(2)}$ with the elements defined by $$m_j = m_{2j} \qquad (1.2.6)$$

This adds multiples of column 2 to columns 3, 4, ----, n and produces zeros for a_{13}, a_{14}, -----, a_{1n}

The configuration is now

$$\begin{bmatrix} x & x & 0 & 0 & 0 \\ x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \end{bmatrix}$$

4. Complete the similarity transformation by multiplying on the left by $I^{(2\alpha)-1}$ and $M^{(2)})^{-1}$ to produce the matrix $(M^{(2)})^{-1} \; (I^{(2\alpha)})^{-1} \; A \; I^{(2\alpha)} \; M^{(2)}$

This does not alter the configuration of the zeros. The next stage in the process has 4 steps similar to the above.

(1) Select the largest of the (n-2) elements a_{23}, a_{24}, ..., a_{2n}

.../9...

Suppose it is 2β, the interchange columns 3 and β

This is a multiplication on the right by $I^{(3\beta)}$

(2) Calculate the (n-3) quantities m_{3j} given by

$$m_{3j} = -\frac{a_{2j}}{a_{23}} \quad j = 4 \text{ to } n \qquad (1.2.7)$$

we have $(m_{3j}) < 1$

(3) Multiply on the right by $M^{(3)}$ with the elements m_j defined by $m_j = m_{3j}$

This adds multiples of column 3 to columns 4, 5, ...n and produces zero for $a_{24}, a_{25}, \ldots, a_{2n}$

The configuration is now

$$\begin{bmatrix} x & x & 0 & 0 & 0 \\ x & x & x & 0 & 0 \\ x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \end{bmatrix}$$

since the zeros introduced by the previous stage are not destroyed.

(4) Complete the similarity transformation by multiplying on the left by $(I^{(3\beta)})^{-1}$ and $(M^{(3)})^{-1}$. This does not alter the configuration of zeros. The transformed matrix is now

$$(M^{(3)})^{-1} (I^{(3\beta)})^{-1} (M^{(2)})^{-1} (I^{(2\alpha)})^{-1} \ldots\ldots,$$

$$A\, I^{(2\alpha)} M^{(2)} I^{(3\beta)} M^{(3)}$$

Continuing in this way it is clear that in (n-2) complete stages we produce a lower Hessenberg matrix.

.../10...

The idea of choosing largest element as well as interchanging columns how ever is to make all transformations numerically stable so that rounding errors are kept to a minimum.

In modular arithmetic, however, the interchanges of columns (or rows) as well as choosing the largest pivotal element are not required.

Example 1.2.1.

Reduce $A = \begin{bmatrix} 5 & 4 & 6 & 3 \\ 2 & 3 & 2 & 1 \\ 4 & 3 & 4 & 2 \\ 3 & 3 & 1 & 4 \end{bmatrix}$

to Hessenberg form.

Sol. $I^{(2\alpha)} = I^{(24)}$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$AI^{(24)} = \begin{bmatrix} 5 & 3 & 6 & 4 \\ 2 & 1 & 2 & 3 \\ 4 & 2 & 4 & 3 \\ 3 & 4 & 1 & 3 \end{bmatrix}$$

$$M^{(2)} = M_{2j} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -\frac{6}{8} & -\frac{4}{8} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (M^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{6}{8} & \frac{4}{8} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

.../11...

$$A^{I^{(24)}}{}_{M^{(2)}} =$$

$$= \begin{bmatrix} 5 & 8 & 6 & 4 \\ 2 & 1 & 2 & 3 \\ 4 & 2 & 4 & 3 \\ 3 & 4 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & \frac{3}{4} & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 8 & 9 & 6 \\ 2 & 1 & \frac{5}{2} & \frac{5}{2} \\ 4 & 2 & \frac{5}{2} & 2 \\ 3 & 4 & -2 & 1 \end{bmatrix}$$

$$(I^{(24)})^{-1} \quad A \quad I^{(24)}{}_{M^{(2)}}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5 & 8 & 0 & 0 \\ 2 & 1 & \frac{6}{4} & \frac{3}{2} \\ 4 & 2 & \frac{5}{2} & 2 \\ 3 & 4 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 8 & 0 & 0 \\ 3 & 4 & -2 & 1 \\ 4 & 2 & \frac{5}{2} & 2 \\ 2 & 1 & \frac{5}{4} & \frac{5}{2} \end{bmatrix}$$

Let $(M^{(2)})^{-1} \; (I^{(24)})^{-1} \; AI^{(24)}{}_{M^{(2)}} = A^{(1)}$

$$A^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{3}{4} & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 8 & 0 & 0 \\ 3 & 4 & -2 & 1 \\ 4 & 2 & \frac{5}{2} & 2 \\ 2 & 1 & \frac{5}{4} & \frac{5}{2} \end{bmatrix} = \begin{bmatrix} 5 & 8 & 0 & 0 \\ 7 & 6 & \frac{1}{2} & \frac{15}{4} \\ 1 & 2 & \frac{5}{2} & 2 \\ 2 & 1 & \frac{5}{4} & \frac{5}{2} \end{bmatrix}$$

$$I^{(38)} = I^{(34)} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

.../12...

$M^{(3)} = m_{34}$ $\qquad (M^{(3)})^{-1}$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{2}{15} \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{2}{15} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$A^{(1)} \; I^{(34)} M^{(3)}$

$$\begin{bmatrix} 5 & 3 & 0 & 0 \\ 7 & 6 & \frac{15}{4} & \frac{1}{2} \\ 4 & 2 & 2 & \frac{67}{30} \\ 2 & 1 & \frac{5}{2} & \frac{11}{12} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{-2}{15} \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 3 & 0 & 0 \\ 7 & 6 & \frac{15}{4} & 0 \\ 4 & 2 & 2 & \frac{67}{30} \\ 2 & 1 & \frac{5}{2} & \frac{11}{12} \end{bmatrix}$$

$(I^{(34)})^{-1} \; A^{(1)} I^{(34)} M^{(3)}$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 5 & 3 & 0 & 0 \\ 7 & 6 & \frac{15}{4} & 0 \\ 4 & 2 & 2 & \frac{67}{30} \\ 2 & 1 & \frac{5}{2} & \frac{11}{12} \end{bmatrix} = \begin{bmatrix} 5 & 3 & 0 & 0 \\ 7 & 6 & \frac{15}{4} & 0 \\ 2 & 1 & \frac{5}{2} & \frac{11}{12} \\ 4 & 2 & 2 & \frac{67}{30} \end{bmatrix}$$

$(M^{(3)})^{-1} \; (I^{(34)})^{-1} \; A^{(1)} I^{(34)} M^{(3)} = S$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \frac{2}{15} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 3 & 0 & 0 \\ 7 & 6 & \frac{15}{4} & 0 \\ 2 & 1 & \frac{5}{2} & \frac{11}{12} \\ 4 & 2 & 2 & \frac{67}{30} \end{bmatrix} = \begin{bmatrix} 5 & 3 & 0 & 0 \\ 7 & 6 & \frac{15}{4} & 0 \\ \frac{38}{15} & \frac{19}{15} & \frac{80}{30} & \frac{3169}{2530} \\ 4 & 2 & 2 & \frac{67}{30} \end{bmatrix}$$

.../13...

When the rows are not interchanged the matrix S

$$\begin{bmatrix} 5 & 4 & 0 & 0 \\ 14 & \frac{27}{2} & \frac{-41}{4} & 0 \\ \frac{344}{41} & \frac{302}{41} & \frac{-461}{32} & \frac{2186}{1681} \\ 3 & 3 & \frac{-7}{2} & \frac{123}{4} \end{bmatrix}$$

1.3 Applications of Hessenberg matrices.

There are many applications for a matrix that has been reduced to the Hessenberg form. Datta (1974) notes that a lower Hessenberg matrixx with a non-zero supper diagonal is nonderogatory.

Some of the important applications depend upon the fact that it is nonderogatory. Datta (1973) has shown that it is easier to find a symmetrizer to a given matrix, while it is in the Hessenberg form.

It should be noted that the eigenvalue problem becomes easier to handle when the matrix is in the Hessenberg form Wilkinson (1965, 1958).

Also in the solution of the Lyapunov matrix equations

$$A X + X A^* = - I \qquad (1.3.1)$$

$$X - A X A^* = I \qquad (1.3.2)$$

the use of A to Hessenberg forms makes the problem easier Datta and Datta (1975). Ordinarily the solution of (1.3.2) is derived from the solution of (1.3.1) by use of bilinear transformations which are very tedious and complicated Barnett and Storey (1970, 1971). Datta has

.../14....

been able to solve (1.3.1) and (1.3.2) when A is in the Hessenberg form without deriving the solution of (1.3.2) from the solution of (1.3.1). The set of equations as known play a significant role in stability analysis of linear control systems.

The Hessenberg matrix is also a form between a general matrix and its canonical form. It has been shown that it is easier to compute various canonical forms of a given matrix when the matrix is in the lower Hessenberg. Datta (1974, 1975) had made use of this fact in computing the Schwarz form and Routh canonical forms respectively.

Although the companion form of a given matrix, would have been more useful, and also a special form of Hessenberg matrix, unfortunately, remarked in Datta (1974), the transformation of an arbitrary matrix to its companion form is a very tedious task- also it has been found that, most methods available in the literature are numerically unstable.

.../15...

CHAPTER TWO

SYMMETRIZATION OF MATRICES

A matrix A is non-derogatory, if there is not more than one Jordan Submatrix (and therefore not more than one eigenvector) associated with λ_i for some i. In other words a matrix A is said to be non-derogatory, if its characteristic polynomial is equal to its minimal polynomial.

Example 2.1.1:

A derogatory matrix.

$$\begin{bmatrix} -5 & 4 & -6 & -3 & 8 \\ -2 & 3 & -2 & 1 & 2 \\ 4 & -3 & 4 & 1 & -6 \\ 4 & -2 & 4 & 0 & 4 \\ -1 & 0 & -2 & 1 & 2 \end{bmatrix}$$

Characteristic equation $\Phi(\lambda) = (\lambda - 2)(\lambda^2 - \lambda - 2)^2$

minimum polynomial $m(\lambda) = \lambda^2 - \lambda - 2$

$$m(\lambda) \neq \Phi(\lambda).$$

Example 2.1.2. (Non-derogatory matrix)

$$\begin{bmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ -2 & -1 & -3 \end{bmatrix} \qquad \Phi(\lambda) = m(\lambda) = \lambda^3$$

.../16...

THEOREM 2.1 (Olga Taussky and Zassenhaus 1969) For every nxn matrix $A = (a_{ij})$ with coefficients in a field f there is a non-singular matrix transforming A into its transpose A^T.

THEOREM 2.2: Every non-singular matrix X transforming A into its transpose is symmetric, if and only if, A is non-derogatory. From the above two theorems, we can therefore say that an arbitrary non-derogatory matrix A is said to be symmetrizable by the matrix **X**, if the matrix product XA is symmetric or $XA = A^TX$ ..... (2.1.1)

2.2. Applications:

Symmetrizers play an important role in the eigen-problem and in the stability problems of control theory.

Firstly, if a symmetrizer X of **A** is known then by using this we can reduce an apparently non-symmetric matrix eigenvalus problem, of the form,

$$Ay = \lambda y \quad \ldots\ldots\ldots \qquad (2.2.1)$$

to a symmetric problem of the form

$$E\,Z = \lambda Z \qquad (2.2.2)$$

The process of which is shown below

$$Ay = \lambda y$$

$$XAy = \lambda Xy$$

$$Cy = \lambda Xy \qquad (2.2.3)$$

where C = XA and X are symmetric and positive definite.

Then $X = S^TDS$ where $S^T = S^{-1}$ and D is a diagonal matrix.

.../17...

(2.2.3) becomes

$$Cy = \lambda S^{T} D Sy - \qquad (2.2.4)$$

$$SCy = \lambda DSy$$

Let $G = S C S^{T}$ and $F = Sy$

then (2.2.) becomes

$$GF = \lambda DF = \lambda D_1^{2} F \text{ where } D_1^{2} = D \qquad (2.2.5)$$

$$GF = \lambda D_1 D_1 F$$

$$D_1^{-1} G D_1^{-1} D_1 F = \lambda D_1 F \qquad (2.2.6)$$

Let $D_1^{-1} G D_1^{-1} = E$ and $D_1 F = Z$ then

becomes $EZ = \lambda Z \qquad (2.2.7)$

where E is symmetric and real.

Another useful aspect of the symmetrizer (Howland and Senez (1970) is the solution of the Lyapunov matrix equation.

$$XA + A^{T} X = -I \qquad (2.2.8)$$

It was discovered that A is greatly simplified when A is in upper Hessenberg form.

The idea of skew symmetrizer was introduced i.e. $XA = T+D$ where T is a skew-symmetric matrix and D is a diagonal matrix. It was proved that the solution of the equation (2.2.8) may be written as a linear combination of n linearly independent Lyapunov pairs, X_j D_j

For n, parameters P_j may be defined by non-singular linear system

$$\sum_{1}^{n} P_j D_j = \tfrac{1}{2} I \qquad (2.2\ 9)$$

whence $X = \Sigma P_j X_j \qquad (2.2.10)$

.../18...

is the required solution whose rank and signature solve the stability problem.

It was also shown Howland and Farell (1965) that in adopting Lanczo's method of tridiagonalization called minimized iterations, it provides the symmetrizing matrix X = D in diagonal form and the symmetric product DA = T in tridiagonal form.

Also the Rayleigh Quotient iteration, Howland (1972) as generalized and applied to the calculations of the characteristic values of arbitrary matrices M by Ostrowski, is discribed by the equations.

$$\lambda_i = \frac{y_i M X_i}{y_i X_i}$$

$$(M - \lambda_i I) X_{i+1} = X_i \qquad (2.2.11)$$

$$y_{i+1} (M - \lambda_i I) = y_i$$

which are applied consectively for i = 0,1,2, ... with arbitrarily chosen starting vectors X_o and y_o for which $y_o X_o \neq 0$. This iteration requires the solution of two nxn linear systems at each step, and has cubic convergence when M has only linear elementary divisors, and geometric convergence in the case of a non-linear elementary divisor.

When a symmetrizer matrix B for M is known the eigen problem for M becomes

$$(A - \lambda B) X = 0 \qquad (2.2.11)$$

and the generalized Rayleigh quotient iteration may be formulated

.../19...

$$\lambda_i^{\cdot} = \frac{X_i\ A\ X_i}{X_i\ B\ X_i} = \lambda_{i-1} + \frac{X_i\ B\ X_{i-1}}{X_i\ B\ X_i} \qquad (2.2.12)$$

$$(A - \lambda_i B)\ X_{i+1} = B\ X_i \qquad (2.2.12)$$

In this case, the two linear systems of the previous case are replaced with one, and, moreover, this one system is simpler than either of the preceding equations due to the simple forms of A and B. Similarly computa-tation of $X_i\ B\ X_{i-1}$ and $X_i\ B\ X_i$, together involves fewer operations than the computation of $y_i\ M\ X_i$ and $y_i\ X_i$

*Finally symmetrizers could be of use in the numeri-*cal solution of polynomials or root-finding. Howland (1961 , I.F.I.P. Con 1962) .

As an example, we illustrate the procedure by which on arbitrary polynomial with real coefficients may be written in form

$$(A - \lambda B) = 0 - \qquad (2.2.13)$$

For, consider the polynomial

$$P_4(\lambda) = \lambda^4 - a_1 \lambda^3 - a_2 \lambda^2 - a_3\lambda - a_4 = 0 \qquad (2.2.14)$$

The companion matrix of (2.2.14) is given by

$$M = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

This has the property that det $(M - \lambda I) = P_4(\lambda)$; this may be symmetrized by premultiplication with a non-singular symmetric matrix.

.../20...

$$B = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_2 & a_3 & a_4 & 0 \\ a_3 & a_4 & 0 & 0 \\ a_4 & 0 & 0 & 0 \end{bmatrix}$$

by finding

$$A = B\,M = \begin{bmatrix} a_1^2+a_2 & a_1a_2+a_3 & a_1a_3+a_4 & a_1a_4 \\ a_1a_2+a_3 & a_2^2 + a_4 & a_2a_3 & a_2a_4 \\ a_1a_3+a_4 & a_2a_3 & a_3^2 & a_3a_4 \\ a_1a_4 & a_2a_4 & a_3a_4 & a_4^2 \end{bmatrix}$$

Since A and B are symmetric the equation (2.2.14) can be written in the form (2.2.13).

The equations $\det(A - \lambda B) = 0$ where A, B are real, symmetric nxn matrices arise in the study of undamped small vibration of a mechanical system about a position of stable equilibrium .

2.3. METHODS FOR THE CONSTRUCTION OF SYMMETRIZERS.

Method 1. An algorithm for computing a symmetrizer of a companion matrix A Batta (1972). Let $X_1, X_2, \ldots\ldots, X_n$ be the successive rows of a symmetrizer X.

Then

(i) Choose X_n arbitrarily

(ii) Compute the remaining rows recursively from the relations (2.3.1)

$$X_{i-1} = X_iA - a_i X_n; \quad i = n, n-1, \ldots, 3, 2, \qquad (2.3.1)$$

.../21...

Example: (2.3.1)

Let $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_1 & a_2 & a_3 \end{bmatrix}$

be a companion matrix of order 3. Then if x_1, x_2, x_3 are the successive row s of X, choosing $x_3 = (1,1,1)$, x_1 and x_2 are given by

$$x_2 = x_3 A - a_3 x_3$$

$$= (1,1,1) \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_1 & a_2 & a_3 \end{bmatrix} - a_3(1,1,1)$$

$$X_1 = (a_1 - a_3,\ 1 + a_2 - a_3,\ a)$$

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a_1 & a_2 & a_3 \end{bmatrix} - a_2(1,1,1)$$

$$= (a_1-a_2,\ a_1-a_3,\ 1)$$

Thus $X = \begin{bmatrix} a_1 - a_2 & a_1 - a_3 & 1 \\ a_1 a_3 & 1+ a_2 - a_3 & 1 \\ 1 & 2 & 1 \end{bmatrix}$

Method 2: Algorithm for computing a symmetrizer of a Hessenberg matrix. (Datta (1974)

Case A: Matrix A is the lower Hessenberg form. The method consists of choosing the last row vector X_n of a symmetrizer arbitrárily and then computing the

.../22...

remaining row vectors X_{n-1}, X_{n-2}, ...X_2, X_1 in order from a recursive relation as follows:

Let $A = \begin{bmatrix} a_{11} & a_{12} & 0 \ldots 0 \\ a_{21} & a_{22} & a_{23} \ldots 0 \\ a_{n-11} & \cdots & a_{n-1n} \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$

be a lower Hessenberg matrix with a non-zero co-diagonal i.e. $a_{i,\ i+1} \neq 0$ and let X_i, $i = 1,2 \ldots .n.$ be the successive rows of matrix S. Then the matrix equation $XA = A^T X$ is completely equivalent to

$$X_1 A = a_{11} X_1 + a_{21} X_2 + \ldots . + a_{n1} X_n \qquad (2.3.2)$$

$$X_{i+1} A = a_{i\ i+1} X_1 + a_{i+1,\ i+1} X_{i+1} + \ldots \ldots \ldots$$

$$\ldots + a_{n,\ i+1} X_n \quad i = 1,2, \ldots .n-1 \qquad (2.3.3)$$

Eliminating X_1, X_2, ..., X_{n-1} one gets

$$X_n \phi (A) = 0 \qquad (2.3.4)$$

where $\phi(A)$ is the characteristics polynomial of A. Since by Cayley - Hamilton Theorem, $\phi (A) = 0$, it follows from (2.3.4) that X_n can be choosen arbitrarily. Since A is non-derogatory then every solution of (2.1.1) is symmetric. Then the following steps can be used for the solution. Let X_1, X_2, ..., X_n be the successive rows of a symmetrizer X.

Then (i) Choose X_n arbitrarily

(ii) Compute the rows X_{n-1}, X_{n-2}, ... X_2 and X_1 recursively from the relation

$$X_i = \frac{1}{a_{i,i+1}} \left(X_{i+1} A - a_{i+1,\ i+1} X_{i+1} \ldots . a_{n,\ i+1} X_n\right) \qquad (2.3.5)$$

.../23...

when $x_n = (\alpha , 0, \ldots\ldots 0)$ $\quad i = n-1, n-2, \ldots\ldots\ldots\ldots 1$ (2.3.6)

Such that $\alpha \neq 0$ is choosen then (2.3.6)

$$\det X = \pm \alpha^n a_{12}^{n-1} a_{23}^{n-4} \ldots a_{n-1,n}^{2-n}$$

which is clearly non zero since and $a_{i, i+1}$ (i=1, ... , n-1) are nonzero.

So with the choice of x_n in the form (2.3.6) we are sure of X that is non-singular

Example 2.3.2

Let $A = \begin{bmatrix} 3 & 2 & 0 \\ 3 & 4 & 2 \\ 1 & 2 & 2 \end{bmatrix}$

let X_1, X_2, X_3 be the successive rows of the symmetrizer

(1) choose $X_3 = (1, 0 \ 0)$

$$X_2 = \frac{1}{a_{23}} (X_3 A - a_{33} X_3)$$

$$= \tfrac{1}{2}\{(3, 2, 0) - 2(1 \ 0 \ 0)\} = (\tfrac{1}{2}, 1, 0)$$

$$X_1 = \frac{1}{a_{12}} (X_2 A - a_{22} X_2 - a_{32} X_3)$$

$$= \tfrac{1}{2}((\tfrac{1}{2}, 1, 0) A - 4(\tfrac{1}{2}, 1, 0) - 2(1, 0 \ 0))$$

$$= (\tfrac{1}{4}, \tfrac{1}{2}, 0)$$

$$\therefore \quad X = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & 1 \\ \frac{1}{2} & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

.../24...

Case 8: A is in the upper Hessenberg form (non-zero sub-diagonal)

$$\text{Let } A = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ 0 & a_{31} & \cdot & \cdot & \cdot & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix}$$

be a upper Hessenberg with $a_{i+1,\,i} \neq 0$

$$i = 1, 2, \ldots n$$

Then $XA = A^T X$ is completely equivalent to $X_1 = a_{11} X_1 + a_{21} X_2$

$$X_i A = a_{1i} X_1 + a_{2i} X_2 + a_{3i} X_3 + \ldots\ldots + a_{i+1i} X_{i+1} \tag{2.3.6}$$

$$X_n A = a_{1n} X_1 + a_{2n} X_2 + \ldots\ldots \tag{2.3.7}$$

where $X_1, X_2, \ldots\ldots X_n$ are the successive rows of the symmetrizer X.

Eliminating $X_n \ldots X_2$ in terms of X_1 we have

$$X_1 \phi(A) = 0 \tag{2.3.8}$$

Since by the Cayley - Halmilton theorem $\phi(A) = 0$, we can therefore choose X_1 arbitrarily and compute X_2, $\ldots\ldots$, X_n completely from the recursive relation

$$X_{i+1} = \frac{1}{a_{i+1,\,i}} (X_i A - a_{1i} X_1 - a_{2i} X_2 \ldots\ldots - a_{ii} X_i) \tag{2.3.9}$$

.../25...

Example 2.3.3

Let $A = \begin{bmatrix} 2 & 1 & 1 \\ 3 & 1 & 1 \\ 0 & 2 & 1 \end{bmatrix}$

Let $X_1 = (1, \ 0 \ \ 0)$

$X_2 = \frac{1}{a_{21}} (X_1 A - a_{11} X_1)$

$= \frac{1}{3} (2, 1, 0) - 2(1, 0 \ 0)$

$= (0 \ \frac{1}{3} \ \ \frac{1}{3})$

$X_3 = \frac{1}{a_{32}} (X_2 A - a_{12} X_1 - a_{22} X_2)$

$= \frac{1}{2} (1, 1, \frac{2}{3}) - 1(0, \frac{1}{3}, \ \frac{1}{3}) - 1(1, 0, 0)$

$\frac{1}{2}(0, \frac{2}{3}, \ \frac{1}{3})$

$= (0, \frac{1}{3}, \frac{1}{6})$

$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{6} \end{bmatrix}$

$XA = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & \frac{2}{3} \\ 1 & \frac{2}{3} & \frac{1}{2} \end{bmatrix} = A^T X$

Method 3: Construction of a positive definite symmetrizer

Datta (1972)

Theorem 2.3.1 Let A be a non-derogatory matrix in its rational canonical form with its eigenvalues all distinct

.../26...

then the inverse of its associated Hankel matrix H, symmetrizes A.

$$H = \begin{bmatrix} s_0 & s_1 & s_{n-1} \\ s_1 & s_2 & \ldots . s_n \\ \ldots\ldots\ldots\ldots\ldots & & \\ s_{n-1} & s_n & s_{2n} \end{bmatrix}$$

Theorem 2.3.2: Suppose that all the eigenvalues of an arbitrary square matrix are real and that r of them are distinct. The r-rowed leading principal minor submatrix, Ur of the associated Hankel matrix H is positive definite.

From the above two theorems it follows that when all eigenvalues of A are real and distinct then the associated Hankel matrix H is a positive definite symmetrizer of A (The proof is omitted)

Example 2.3.3.

Let $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -2 & 3 \end{bmatrix}$

be the given matrix of order 3, Then, the successive rows of a symmetrizer X of A are

$$X_3 \begin{bmatrix} 2 & -3 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad X_3 \begin{bmatrix} -3 & 1 & 0 \\ 0 & -3 & 1 \\ 0 & -2 & 0 \end{bmatrix} \text{ and } X_3$$

where $X_3 = (x_{11}, x_{12}, x_{13})$ is an arbitrary vector calculated from the equations below

$$3x_{11} + 3x_{12} + 5x_{13} = 0$$

$$3x_{11} + 5x_{12} + 9x_{13} = 0$$

$$5x_{11} + 9x_{12} + 17x_{13} = 0$$

The unique solution of the system is

$$X_{11} = \tfrac{1}{2},\ X_{12} = -3,\ X_{13} = \frac{3}{2}\ .$$

The successive rows of the matrix H^{-1} are therefore

$$(1, \tfrac{-3}{2}, \tfrac{1}{2}),\ (-\tfrac{3}{2}, \tfrac{13}{2}, -3) \text{ and } (1, -3, \tfrac{3}{2})$$

The matrix

$$H^{-1} = \begin{bmatrix} 1 & \frac{-3}{2} & \frac{1}{2} \\ \frac{-3}{2} & \frac{13}{2} & -3 \\ \frac{1}{2} & -3 & \frac{3}{2} \end{bmatrix}$$

is positive definite, this symmetrizes A.

.../28.

CHAPTER 3.

Residue Arithmetic applications to Exact computation of a symmetrizer and transformation to the Hessenberg form

3.1 Residue Arithmetic with a single modulus

When m divides x-a, then x is congruent to a modulus m and is usually written

$$x \equiv a(\text{mod } m) \text{ or } |x|_m \qquad (3.1.1)$$

where x, a, m, are integers.

Put in another way if $x \equiv a$ (modulus m), then a is called a residue of x to modulus m.

If $0 \leq a \leq m-1$, then a is called the least residue of x to modulus m, with m always referred to as the modulus or the bass or the radix.

Theorem 3.1.1 (Young and Gregory, 1973)

Let x,y,z be integers and m a positive integer then

(i) $|x|_m$ is unique

(ii) $|x|_m = |y|_m = x \equiv y \text{ (mod } m)$

(iii) $|km|_m = 0$ for all integers k

(iv) $|x \pm y|_m = \left| |x|_m \pm |y|_m \right|_m = |x \pm |y|_m$

$$= ||x|_m \pm y|_m$$

(v) $|xy|_m = ||x|_m . |y|_m|_m = |x . |y|_m|_m$

$$= ||x|_m * y|_m$$

.../29...

It is interesting to note that when m=p, a prime then the set of residues (0, 1, ..., p - 1) forms an abelian group under multiplication, so that the set forms a field, a finite field called Galois Fields denoted by GF(p).

For $a \neq 0$ ϵ (GF(p) a, b such that $a.b. \equiv 1 \pmod p$ a, b are called the multiplicative inverse of each other and denoted by a^{-1}, b^{-1} respectively.

Let a be a negative integer such that $|a| < \frac{(p-1)}{2}$

Then $|a|_p = p + a$.

Let A be the set of integers

$$\left\{ -\frac{p-1}{2}; \quad -\frac{p-1}{2} + 1, \ldots\ldots, 0, \ldots\ldots \frac{p-1}{2} \right\}$$

then the set Res(A) = $|i|_p$: i ϵA is the set GF(p) itself. The following function 'value' maps the elements of GF(p) back to the elements of A.

$$\begin{aligned} \text{Value}(a) &= a \text{ if } a \leq (p-1)/2 \\ &= -(p-a) \text{ otherwise} \end{aligned} \qquad (3.1.2)$$

The above thus shows that for a given prime p the integers

$a \epsilon \left[-\frac{(p-1)}{2}, \frac{p-1}{2} \right]$ have a special property viz

$$\text{Value (Residue (a))} = a \qquad (3.1.3)$$

3.1.3 thus shows that there is a unique representation for the members of A.

When performing an arithmetic process involving only addition, multiplication and subtraction, if the operands are integers the result is also an integer. The modulo p arithmetic can then be used for some prime p when the intermediate results and final

.../30...

result of an arithmetic processs lie in the interval

$$\left[-(p-1)/2\ ,\ (p-1)/2\right]$$

Exanple 3.1.1.

Solve:- (3*4 - 5) - 8, use modulus prime 17:

Solution: Residue (3) = 3

" (4) = 4

" (-5) = 12

" (-8) = 9

(3x4-5) - 8

= (3x4 + 12) + 9 mod 17 ≡ 33 mod 17

= 16 mod 17

= - 1 given by value (16)

Division is treated as multiplication by the multiplicative inverse of the divisor E.g. In GF(17) 2/3 becomes $2x\ 3^{-1}$ (mod 17), 3/5 becomes 3×5^{-1} (mod 17) The following Euclidean algorithm can be used to compute the multiplicative inverse in a finite field Knuth (1969) Rao (1975) This algorithm computes the multiplicative inverse of a number K, $0 < K < p$ with respect to a prime number p.

Step 1 Set $(U_1, U_2, U_3) = (1, 0, K)$

$(V_1, V_2, V_3) = (0, 1, P)$

Step 2 If $V_3 = 0$ stop U_1 is the multiplicative inverse of K with respect to p otherwise go to step 3.

Step 3 Set $q = (U_3/V_3)$ where q is an integer

.../31...

Set $(t_1, t_2, t_3) = (U_1, U_2, U_3) - (V_1, V_2, V_3)q$

Set $(U_1, U_2, U_3) = (V_1, V_2, V_3)$

Set $(V_1, V_2, V_3) = (t_1, t_2, t_3)$

Go to 2

Example 3.1.2:

Find the inverse of 5 in GF(7)

Solution: $P = 7 \quad k = 5$

Step 1 $(U_1, U_2, U_3) = (1, 0, 5)$

$(V_1, V_2, V_3) = (0, 1, 7)$

Since $V_3 = 7$ we go to step 3.

Step 3. $q = \frac{U_3}{V_3} = \frac{5}{7}$ and since q is an integer, q in this case is equal to zero

$\therefore (t_1, t_2, t_3) = (U_1, U_2, U_3) - (V_1, V_2, V_3)q$

$= (1, 0, 5) - (0, 1, 7) \times 0 = (1, 0. 5)$

$(U_1, U_2, U_3) = (V_1, V_2, V_3) = (0,1,7)$

$(V_1, V_2, V_3) = (t_1, t_2, t_3) = (1, 0, 5)$

Go to step 2.

Step 2 $V_3 \neq 0$ so we go to step 3.

Step 3 $q = \frac{7}{5} = 1$

$\therefore (t_1, t_2, t_3) = (0, 1, 7) - (1, 0 \ 5) \times 1$

$= (-1, 1 \ 2)$

$(U_1, U_2, U_3) = (1, 0 \ 5)$

$(V_1, V_2, V_3) = (-1, 1, 2)$

Go to step 2. $V_3 \neq 0$ then go to step 3.

.../32..

Step 3 $q = \frac{5}{2} = 2$

$(t_1, t_2, t_3) = (1, 0, 5) - (-1, 1, 2) \times 2$

$= (3, -1, 1)$

$(U_1, U_2, U_3) = (-1, 1, 2)$

$(V_1, V_2, V_3) = (3, -1, 1)$

Since $V_3 \neq 0$ we go back to step 3.

Step 3 $q = \frac{2}{1} = 2$

$(t_1, t_2, t_3) = (-1 \quad 1 \quad 2) - (3 \ -1 \ 1) \times 2$

$= (-7, \ 3, \ 0)$

$(U_1, U_2, U_3) = (3, -1, 1)$

$(V_1, V_2, V_3) = (-7, 3, 0)$

Go to step 2.

$V_3 = 0$ $\therefore$ U_1 is the inverse

$\therefore$ the inverse is 3.

If b/a is the rational number which is represented by $\left|b.a^{-1}\right|_p$, then it is possible to compute one of a or b knowing the other, Mahedava Rao (1975). For example, if a is known the quantity value $\left|\left|a\right|_p . \ b.a^{-1}\right|_p|_p$ gives the value of b. Further it is sufficient to know any non-zero multiple k_a of a for

$$\frac{1}{ka} \cdot \left(\left|\left|ba^{-1}\right|_p . ka\right|_p\right) \qquad (3.1.4)$$

gives kb/ka. But since the value should be taken in the end, it is necessary that ka and kb are integers in the set,

$$\left\{\frac{-p-1}{2}, \ldots\ldots 0, 1, \ldots\ldots \frac{p-1}{2} .\right\}$$

.../33 ..

Example 3.1.3: Let P = 17

(i) $\frac{1}{2}$ is represented by $\left|1 \times 2^{-1}\right|_{17} = \left|1 \times 9\right|_{17} = 9$

(ii) $\frac{7}{3}$ " " " $\left|7x\ 3^{-1}\right|_{17} = \left|7 \times 6\right|_{17} = 8$

(iii) $\frac{7}{2}$ " " " $\left|7x2^{-1}\right|_{17} = \left|7x9\right|_{17} = 12$

Given that the denominators are 2 and 3 we can easily compute the numerator.

(i) $\frac{1}{6} x\left|\ 9x6\right|_{17} = \frac{3}{6}$ this shows that it is enough to know a multiple of the denominator.

(ii) $\frac{1}{6} x \left|6x8\right|_{17} = \frac{14}{6}$, although it gives the right answer but it is a bad one because $14 \notin \left[\frac{-p-1}{2}, \frac{p-1}{2}\right]$

(iii) $\frac{1}{3} x \left|12 \times 8\right|_{17} = \frac{11}{8}$ $11 \notin \left[\frac{-p-1}{2}, \frac{p-1}{2}\right]$

Example 3.1.4 Let us for illustration

evaluate $\frac{1}{2} + \frac{2}{3} + \frac{1}{4} = \alpha$ (say)

Sol. Let p = 23

$\left|1/2\right|_{23} = 12$

$\left|2/3\right|_{23} = 16$

$\left|1/4\right|_{23} = 6$

$\therefore \left|\frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right|_{23} = \left|12 + 16 + 6\right|_{23} = 11$

Suppose a multiple of the denominators is known then the solution is given by

$$12\frac{1}{12} \times \left|11 \times 12\right|_{23} = \frac{17}{12} = \alpha$$

Generally, it can now be observed that the choice of the prime to be used in performing a computation using residue arithmetic is extremely important.

.../34...

The prime chosen must ensure that the results of all calculations lie within the interval of unique representation of that prime.

3.2 Residue Arithmetic with multiple Moduli:

Let $P_1, P_2, \ldots\ldots, P_s$ be distinct primes. Let a be an integer. Then the set $|a|_{P_1}, |a|_{P_2}, \ldots\ldots |a|_{P_s}$ is called the residue representation of a with respect to the moduli $P_1, P_2, P_3, \ldots\ldots, P_s$.

Let $M = P_1 \cdot P_2 \ldots\ldots P_s = \prod_{i=1}^{s} P_i$

Then we have again a one-to-one correspondence between the set of integers $(0, 1, 2, \ldots\ldots, M-1)$ and the set of their residue representations.

Theorem 3.2.1

Let $a \sim \{|a|_{p_1}, |a|_{p_2} \ldots\ldots |a|_{p_s}\}$

$b \sim \{|b|_{p_1}, |b|_{p_2} \ldots\ldots, |b|_{p_s}\}$

be the residue representations of integers a and b respectively. Then

$$a \pm b \sim \{|c|_{p_1}, |c|_{p_2}, \ldots\ldots\ldots |c|_{p_s}\} \quad (3.2.1)$$

where $|c|_{p_i} = \left| |a|_{p_i} \pm |b|_{p_i} \right|_{p_i}$ for $i = 1, 2, \ldots\ldots s$

and $a.b. \sim |d|_{p_1}, |d|_{p_2}, \ldots\ldots\ldots |d|_{p_s}$ (3.2.2)

where $|d|_{p_i} = \left| |a|_{p_i} \cdot |b|_{p_i} \right|_{p_i}$ for $i = 1, 2, \ldots\ldots$

.../35....

The following theorem known as the chinese remainder theorem Young and Gregory (1973) gives a method of computing $|a|_M$, where a is an integer and $M = \prod_{i=1}^{s} p_i$. given the residue representation

$$a \sim \{|a|_{p_1}, |a|_{p_2} \ldots\ldots |a|_{p_s}\}$$

Theorem 3.2.2

Let $p_1, p_2, \ldots, p_s$ be prime moduli, and $M = \prod_{i=1}^{s} p_i$

Let $a \sim \{|a|_{p_i}, |a|_{p_2} \ldots\ldots, |a|_{p_s}\}$

be the residue representation of a. Also let

$$\hat{p}_j = \frac{M}{p_j} \quad \text{then}$$

$$\left|a\right|_M = \left| \left\{ \sum_{j=1}^{s} \hat{p}_j \right\} \left|a\right|_{p_j} \cdot \hat{p}_j^{-1} (p_j) \right| p_j \Big| M$$

when $\hat{p}_j^{-1} (p_j)$ denotes the multiplicative inverse of $|\hat{p}_j|_{p_j}$ with respect to the base p_j.

Example 3.2.1

Let $P_1 = 3$, $p_2 = 7$ and $p_3 = 11$

Let $a \sim \{2, 4, 5\}$

Then $M = 231$

$$\hat{P}_1 = 231/3 = 77 \quad \hat{P}_2 = 231/7 = 33 \quad \hat{P}_3 = \frac{231}{11} = 21$$

$$|\hat{P}_1|_3 = |77|_3 = 2 \quad \therefore \hat{P}_2^{-1} (P_2) = 3$$

$$|\hat{P}_2|_7 = |77|_3 = 5 \quad \therefore \hat{P}_2^{-1} (P_2) = 3$$

$$|\hat{P}_3|_{11} = 21_{11} = 10 \quad \therefore \hat{P}_3^{-1} (P_3) = 10$$

$$\therefore \quad |a|_{231} = \left| 77|2.2|_3 + 33|4.3|_7 + 21|5.10|_{11} \right|_{231}$$

$$= |77x1 + 33x5 + 21x6|_{231}$$

$$= |77+165 + 126|_{231} = |368|_{231}$$

3.3 Applications to symmetrizers and Reduction of matrix to Hessenberg form

If one is interested in error free or exact solutions of matrix equations or exact transformation of matrices, it appears that the computational gain of the congruence method over the conventional multiple precision methods is larger for large matrices (Borosh and Franckel (1966). The computation time for working in t-precision is roughly proportional to t^2. In the congruence method, where most of the work is single precision computation time is roughly proportional to t. (See also Krishamurthy (1975))....

3.3.1 Algorithm for computing a symmetrizer (single prime modulus), Datta (1973)

This algorithm was earlier discussed in chapter 2. We will reinvolve the formular (see p.22) equation number 2.3.5.

.../37...

$$X_1 = \frac{1}{a_{i,\,i+1}} \left(X_{i+1}A - a_{i+1,\,i+1} X_{i+1} \ldots\ldots - a_{n\,i+1} X_n\right)$$

$i = n-1, n-2, \ldots\ldots, 3, 2, 1$, where $X_1 \ldots X_n$ are the successive rows of the symmetrizer X. Before computing in residue arithmetic we must ensure that the following conditions hold

(i) A must be in the lower Hessenberg form

(ii) If (i) is not satisfied A is brought to this form (see Chapter 2)

(iii) A must be non-derogatory

(iv) Find a bound on the magnitude of $|x_{ij}|$ max

(v) Then (iv) will enable us to choose a prime for the modular Arithmetic.

It was found out that for the algorithm

$|x_{ij}|$ max $\leq 2^{2(n-2)}\, a^{n-1}$ where $|a|$ is the largest element of A. Hence for modular Arithmetic the prime number chosen $p \geq 2^{n(n-2)}\, a^{n-1}$. Once p is chosen the symmetrizer can be computed in the following manner:

Step 1 choose the prime p for the given matrix.

Example 3.3.1 $A = \begin{bmatrix} 3 & 2 & 0 \\ 3 & 4 & 2 \\ 1 & 2 & 2 \end{bmatrix}$

Choose p = 17.

Step 2 All elements a_{ij} of A are entred as a_{ij} (mod p)

Step 3 All $a^{-1} \in$ GF(p) are calculated from the relation

$a\,x \equiv 1$ (mod p)

Step 4 Go to the algorithm

.../38...

All multiplications and additions are done in GF(p).

To ensure that our symmetric matrix is a symmetrizer we can find XA where XA ε GF(p).

Conversion of the elements of the symmetrizer X:

Step 1 (i) Find $\prod_{1=i}^{n-1} a_{i,\, i+1}$

(ii) If $\prod_{1=i}^{n-1} a_{i,\, i+1} \geq p$ then solve the equation

$$\prod_{i=1}^{n-1} a_{i,\, i+1} \equiv \alpha \bmod p$$

(iii) Let $\prod_{i=1}^{n-1} a_{i,\, i+1} = \alpha$ if $\prod_{1=i}^{n-1} a_{i,\, i+1} < p$ or/(ii) is satisfied then α is the result we need.

Step 2 Find $\alpha X GF(p)$

Step 3 Divide the new X by $\prod_{i=1}^{n-1} a_{i,\, i+1} = \alpha$.

Steps 1 to 3 is calculating $\dfrac{\text{VALUE } (\alpha\, X_{ij})}{\text{VALUE } (\alpha)}$

Example 3.3.1

P = 17

(i) Choose X_3 arbitrarily . Let X_3 = (1, 0 0)

(ii) $X_2 = \frac{1}{a_{23}} (X_2 A - a_{33}\, x_3)$

$= \frac{1}{2} (3,2,0) - 2(1\ 0\ 0)$

$= 2^{-1} ((3,2,0) + 15 (1\ 0\ 0)$

2^{-1} = 7 and -2 = 15 in GF (17)

all calculations now in GF (17)

X_2 = 9(1 2 0)

(9 1 0)

.../39..

$$X_1 = \frac{1}{a_{12}} (X_2A - a_{22}X_2 - a_{32} X_3)$$

$$= 9\ (13,\ 5,\ 2) + (15,\ 13,\ 0) + (15,\ 0\ 0)$$

$$= 9\ (9,\ 1,\ 2) = (\ 13,\ 9,\ 1)$$

$$X = \begin{bmatrix} 13 & 9 & 1 \\ 9 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

So that X is now iin modular Arithmetic

Find XA

$$= \begin{bmatrix} 13 & 9 & 1 \\ 9 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 3 & 3 & 0 \\ 3 & 4 & 2 \\ 1 & 2 & 2 \end{bmatrix} = \begin{bmatrix} 16 & 13 & 3 \\ 13 & 5 & 2 \\ 3 & 2 & 0 \end{bmatrix}$$

which is also symmetric.

Conversion of elements x_{ij} innto rational.

$$\prod_{i=1}^{n} a_{i,\ i+1} = \alpha = 2 \times 2$$

VALUE $(\alpha) = 4$

VALUE (α, x_{ij}) = VALUE $(\alpha\ x_{ij})$

$$= \begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 0 \\ 4 & 0 & 0 \end{bmatrix}$$

$$\frac{\text{VALUE } (\alpha X_{ij})}{\text{VALUE } (\alpha)} = \frac{1}{4} \begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 0 \\ 4 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & 1 \\ \frac{1}{2} & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

.../40...

3.3.2: Reduction of a given general matrix to the Hessenberg form

The Reduction can be carried out through the following steps.

It should be noted that the prime number choosen for reduction is always very high and multiple prime moduli is therefore the best way to reduce a matrix to the Hessenberg form in residue Arithmetic.

For the choice of M See page 67

Step 1 Convert the entries into residue numbers modulo P_i for each P_i

Step 2 Carry out the reduction exercise for each P_i separately

Step 3 Combine the results of each P_i using the Chinese remainder's Theorem.

Step 4 Convert the elements back into rational after the common denominator must have been computed.

In case the symmetrization of the matrix is to take place, then the steps taken in section 3.3.1 can be carried out omiting step 4 of the above.

A problem that may arise while performing steps 1 to 3 of the above, the element $a_{i,\ i+1}$ of any mod P_i may be zero, when this happens then we can either chose a different prime p instead of P_i or go through step 4 and choose different primes for the steps in section 3.3.1.

Example 3.3.2:

Reduce

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 4 \\ 1 & 3 & 2 \end{bmatrix}$$

to the Hessenberg form using multiple prime moduli

... /41...

Solution: Let M = 385 i .e. $P_1 = 5$, $P_2 = 7$, $P_3 = 17$.

For P_1 let A mod 15 = A_5

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 4 \\ 1 & 3 & 2 \end{bmatrix}$$

$$m_{2j} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -3.2^{-1} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$(m_{2j})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A_5\ m_{2j} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 0 \end{bmatrix}$$

$$(m_{2j})^{-1}\ A_5\ m_{2j} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 4 & 1 \\ 1 & 3 & 0 \end{bmatrix}$$

Factor = 4, $|4|_5 = 4$

Let A mod 7 = A_7

$$A_7 = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 4 \\ 1 & 3 & 2 \end{bmatrix}$$

.../42...

$$m_{2j} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -3.2^{-1} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$(m_{2j})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix}$$

$$A_7 m_{2j} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 4 \\ 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 3 & 2 & 1 \\ 1 & 3 & 1 \end{bmatrix}$$

$$(m_{2j})^{-1} A_7 m_{2j} =$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 3 & 2 & 1 \\ 1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 3 & 6 \\ 1 & 3 & 1 \end{bmatrix}$$

Factor = 4, $|4|_7 = 4$

Let A mod 11 = A_{11}

$$A_{11} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 4 \\ 1 & 3 & 2 \end{bmatrix}$$

$$m_{2j} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -3.2^{-1} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix} \qquad (m_{2j})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 0 & 1 \end{bmatrix}$$

.../43...

$$A_{11}m_{2j} = \begin{bmatrix}1 & 2 & 3\\ 3 & 2 & 4\\ 1 & 3 & 2\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 4\\ 0 & 0 & 1\end{bmatrix} = \begin{bmatrix}1 & 2 & 0\\ 3 & 2 & 1\\ 1 & 3 & 3\end{bmatrix}$$

$$(m_{2j})^{-1}A_{11}\,m_{2j} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 7\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 2 & 0\\ 3 & 2 & 1\\ 1 & 3 & 3\end{bmatrix} = \begin{bmatrix}1 & 2 & 0\\ 10 & 1 & 0\\ 1 & 3 & 3\end{bmatrix}$$

Factor 4, $|4|_{11} = 4$

Using the Chinese Remainder Theorem for combining the matrix to get the Hessenberg matrix in rationals.

$$M = P_1 \times P_2 \times P_3 = 5\times 7\times 11 = 385$$

$$\hat{P}_1 = 7\times 11 = 77$$

$$d_1^{-1} = (\hat{P}_1 \bmod P_1)^{-1} \bmod P_1 = |77|_5^{-1}\,|_5 = |2^{-1}|_5$$

$$= 3$$

$$\hat{P}_2 = \frac{M}{P_2} = 55$$

$$d_2^{-1} = 6$$

$$\hat{P}_3 = 35 \quad d_3^{-1} = 6$$

We compute the elements of the Hessenberg using

$$|a|_M = \left|\sum_{j=1}^{S} \hat{P}_j\,|a|_{pj}\cdot \hat{P}_j^{-1}(P_j)\,|\,P_j|_M\right| .$$

Common denominator COMDEN

$4 \bmod 5 \equiv 4 \bmod 11 \equiv 4 \bmod 7 = 4$

Let a_{ij} be an element of the Hessenberg matrix

.../44...

$a_{11} \equiv 1 \bmod 5 \equiv 1 \bmod 7 \equiv 1 \bmod 11$

$= 1$

$a_{12} \equiv 2 \bmod 5 \equiv 2 \bmod 7 \equiv 2 \bmod 11$

$= 2$

$a_{21} \equiv 2 \bmod \equiv 1 \bmod 7 \equiv 10 \bmod 11$

$|77 \times |2x3|_5 + 55x|6x1|_7 + 35x|6x10|_{11}|_{385}$

$=197$

Similarly

$a_{22} = 199$

$a_{23} = 286$

$a_{31} = 1 \quad a_{32} = 3, \; a_{33} = 190$

$$= S = \begin{bmatrix} 1 & 2 & 0 \\ 197 & 199 & 286 \\ 1 & 3 & 190 \end{bmatrix}$$

Value (4) = 4

$$\frac{\text{VALUE } (4a_{ij})}{\text{VALUE } (4)} = \begin{bmatrix} 1 & 2 & 0 \\ \frac{9}{2} & \frac{13}{2} & \frac{-11}{4} \\ 1 & 3 & \frac{-5}{2} \end{bmatrix}$$

.../45...

CHAPTER 4

INTRODUCTION:

The digital computers work with finite length numbers in a p-ary scale, where p is any positive integer (called radix or base usually a binary or decimal scale). Such a system ; known as the p-ary system; enables us to measure any numerical quantity in steps of p^{-r} where r is the length of number of digits available for calculations .

This system has a major disadvantage, in that not all rationals have exact representations, since they may not terminate within r-digits either because they may be periodic, or recurring or too small. Hence it is imposible to represent rationals of the form a/b ($b \neq 0$) exactly if b has factors which are relatively prime to the choice p. This disadvantage in the p-ary system leads to errors. When arithmetic operations are performed, the error increases with increasing operations. The consequence of this is that the algorithm being computed becomes totally unstable and results obtained from such algorithm are therefore unreliable.

Since most practical algorithms, particularly in matrix computations are ill-conditioned, the use of residue arithmetic has been suggested Young and Gregory (1973) Rao et al(1975).

The use of residue arithmetic has been discussed in the third chapter of this dissertation. The use of residue arithmetic also has certain disadvantages. For instance while choosing a prime p for a particular algorithm - Krishnamurthy (1975) the computational complexity increases with increasing p. The multiplication complexity is found out to be

.../46...

of order

$$0 \quad \frac{(k\ (K + 1))}{2}$$

where k is the precision or number of binary digits in p. The computation of multiplicative inverse (and hence division) over GF(p) requires the use of Euclidean algorithm involving on an 6 $\log_2(p = 6k)$ k-precision divisions and 12k k - precision multiplications each of complexity $0(k^2)$ (Knuth 1969).

To reduce the complexity, the use of multiple primes has also been suggested. When this is done the matrix computations are performed for each P_i separately and results are later combined using the earlier discussed Chinese Remainder Theorem. The multiple primes moduli arithmetic is not without its own failures, computation increases with increasing prime and with increasing arithmetic operations , the computation therefore becomes complex.

These difficulties have made it necessary for us to look for an alternative arithmetic system which though modular in nature has the simplicity of arithmetic operations as in the p-ary system yet free from round-off errors and is simple for hardware realization.

The segmented p-adic system is such a system. It is also equivalent to modulo p^r (where p is a prime) arithemetic, and has all these attractive features Krishnamurthy (1975)

(a) The conversion of integers and other rationals into p-adic representation is simple and realizable by existing hardware.

.../47...

(b) The evaluation of multiplicative inverse is of many orders less in complexity, infact it is simpler than binary division and has a complexity 0 $\frac{(k(k+1)}{2}$ for k bit numbers.

(c) The fast high-precision multiplication algorithm due to Toom - Cook and schonhange - strassen Knuth (1969) or other high speed switching schemes could be readily used for speeding up the p-adic arithmetic, when very large number of bits are involved.

(d) The conversion from modulo p^r to the positional representation is trivial, based on the nature of the algorithm.

(e) The choice of r can be made in regular integral steps to meet the requirement of the specific matrix problem in hand. Accordingly the hardware design becomes simple.

The p-adic was first introduced by Hensel in 1900. A detailed introduction to p-adic can be found in Bachmen (1964) Borevich and Shafarevich (1966) and Mahler (1973).

This chapter deals with the segmented p-adic arithmetic The expansion of a rational number into a p-adic form is what we call the Hensel code (in honour of Hensel) denoted by H(p,r) where p is the prime base and r is the number of digits used.

4.1. Properties of H(p, r) codes .

(1) Every rational number a/b with GCD(a,b) = 1 and $b \neq 0$ in the range $0 \leq a \leq N$, $0 < b \leq N$ where N is a positive integer such that $N < p^{r/2}$ can be uniquely coded as an r-digit ordered sequence, where each digit as sume values from 0, 1, -----, p - 1 where p is the prime chosen.

.../48...

(2) The negative rational numbers have a valid radix complement representation as in p-ary arithmetic and the weighted structure is retained.

(3) All the four basic arithmetic operations using H(p,r) codes proceed in direction (left to right); also the length of resulting code remains constant provided the resulting rational number is in the range specified in (1)

(4) The H(p,r,) code is a unified representation for all rationals and arithmetic operations are somewhat similar to the residue system , unlike the residue system, division using H(p,r) is deterministic in the sense that it does not involve the comparision of relative magnitude of the operands. The operation is exactly similar to multiplication .

(5) The essential distinction between using a truncated p-ary code and a truncated p-adic code using r-digits is that in the p-ary system the numbers are exactly representable only if they are either integers or radix fractions; hence all other rationals can only be approximately represented . In the H(p,r) code, however every rational number a/b which is an element of the Farey sequence with numerators and denominators less than N are exactly representable .

(6) Since all the basic arithmetic operations are performed exactly for a given range of the rationals, the H(p,r) codes can be used for exact numerical computations.

.../49...

4.2. (A) Algorithm for the conversion of Rational number to p-adic form.

Mahedeva Rao (1975)

Given a rational number $\alpha = \frac{a}{b}$ with GCD(a, b) = 1 and $b \neq 0$ its p-adic expansion can be obtained by the following algorithm.

Step 0 Set $\beta = \alpha$ find n such that $\alpha = \frac{c}{d}\ p^n$ if $n > 0$

(c and d are integers)

Set $a_0, a_1, \ldots\ldots, a_{n-1} = 0$

e.g. $\frac{25}{31} = \frac{1}{31} \cdot 5^2$ $\therefore$ n = 2 and n > 0 so

$a_0, = 0$ and $a_1 = 0$

Go to step 2

Step 1 Find n such that $\beta = \frac{c}{d}\ p^n$

e.g. $\frac{25}{31} = \frac{1}{31} \cdot 5^2$ for p = 5

Step 2 Solve the conqurence $dx \equiv 1 \pmod p$

If Xn is a solution then

$$a_n = c\ X_n \pmod p$$

Step 3 Set $\gamma = \beta - a_n p^n$

If $\gamma = 0$ set $a_i = 0$ for $i > n$ and stop otherwise, set $\gamma = \beta$ and go to step 1.

The algorithm is however nonterminating since the convergence to the given rational can be achieved only for infinite number of digits if the rational is neither an integer nor a radix.

(B) P-Adic Number Representation

Any rational number $\alpha = (a/b)\ p^n$ where n, a, b, are positive or negative integers, $b \neq 0$ and G.C.D(a,b) = 1 can be written in the p-adic form

$$\alpha = \sum_{j=n}^{\infty} a_j p_j \qquad 4.2.1.$$

where, the coefficient a_j are integers such that

$$0 \leq a_j \leq (p - 1) \text{ and } p \text{ is any prime number .}$$

The infinite series converges to α with respect to the p-adic norm.

e.g. $\frac{1}{3}$ has the p-adic expansion

$$\frac{1}{3} = .2313131 \ldots\ldots \text{ (for } p = 5)$$

$$= 2 + 3p\,(1 + p^2 + p^4 \ldots.) + 1$$

$$(1 + p^2 + p^4) \qquad 4.2.2$$

The infinite series $1 + p^2 + p^4 \ldots\ldots$ converges to $\frac{1}{1 - p^2}$ with respect to the p-adic norm (Rao 1975),

Note - In the conventional sense the series 4.2.2 under usual norm is divergent. However, in the sense of p-adic norm it is divergent.

Thus for p = 5

For convenience, the expansion of 4.2.1 will be denoted by the sequence

$$a_{-m}\, a_{-m+1} \ldots\ldots a_0\, a_1 \ldots\ldots \text{ for n -ve}$$

$$a_0\, a_1 \ldots\ldots \text{ for n = 0}$$

$$0 \ldots\ldots 0a_n a_{n+1} \ldots. \text{ for n + ve}$$

Integers as well as radix fractions have representations which are identical to the p-ary written here from left to right in the ascending powers of p instead of right to left with the actual rational being realised only as a sum of the infinite series.

.../51....

4.3. Arithmetic Algorithms. Krishnamurthy et al (1975)

The 4 basic arithmetic algorithms using Hensel codes are described in this section. We shall note that all operations including division proceed from lower index (power) position of the radix towards higher position.

In a finite segmented p-adic system any rational number is of the form

$$a_{-m}\ \bar{a}_{m+1} \cdots\cdots \bar{a}_1\ a_0\ a_1 \cdots\cdots a_n$$

where $r = m + n + 1$

or $\alpha = (m_\alpha , e_\alpha)$ as an ordered pair in the mantissa and exponent form.

Since we keep the length of $H(p, r, \alpha)$ constant, we denote only the negative exponents in e_α

When $e_\alpha = m$ the radix point is placed m digits to the right of the left most digit.

Addition.

Addition in a p-adic system is similar to that in positive p-ary base. In general, the algorithm for addition of any two numbers α and β (+ ve or - ve) is given by

$$\alpha = a_{-m}\ a_{-m+1} \cdots\cdots a_{-1} \cdot a\ a_1 \cdots\cdots a_n$$

$$\beta = b_{-m}\ b_{-m+1} + \cdots\cdots b_{-1} \cdot b_0\ b_1 \cdots\cdots b_n$$

For convenience, we take that α and β have identical number of digits) align the p-adic point and proceed finding the sum digit s_i and carry digit c_{i+1} from a knowledge of a_i, b_i and c_i. The carry c_{n+1} arising from the addition of $a_n + b_n + c_n$ (Mod p) is ignored

.../52...

Thus $S_i = (a_i + b_i + c_i) \pmod p$

for $i = -m, -m+1, \ldots 0, 1, \ldots n$ with $c_{-m} = 0$

and $c_{i+1} = 1$ if $a_i + b_i + c_i \geq p$

$= 0$ otherwise

Ignore C_{n+1}

e.g. $p = 5$, $r = 4$

$\alpha = {}^4/9 = .1124 \quad \beta = {}^8/9 = .2243$

and $v = \alpha + \beta = \frac{12}{9} = .3313$. See table 1

Table 1

i	0	1	2	3
a_i	1	1	2	4
b_i	2	2	4	3
c_i	0	0	0	1
s_i	3	3	1	3

B. Subtraction:

Subtraction is realized as a complimented addition.

Let $\alpha = a_{-m}\ a_{-m+1} \ldots a_{-1}\ a_0\ a_1 \ldots a_n$

The $\bar{\alpha} = \bar{a}_{-m}\ \bar{a}_{-m+1} \ldots \bar{a}_1\ \bar{a}_0\ \bar{a}_1 \ldots \bar{a}_n$

The algorithm consists the following rules

Rule 1. If $a_i \neq 0$ for $i = -m, -m+1, \ldots, n$

then $\bar{a}_i = p - a_i$ for $i = -m$

$\bar{a}_i = (p-1) - a_i$ for $-m+1 \leq i \leq n$

Rule 2. If $a_i = 0$ for $-m \leq i \leq j$, then

$\bar{a}_i = 0$ for $-m \leq i \leq j$

and $\bar{a}_{j+1} = p - a_{j+1}$

$\bar{a}_i = (p-1) - a_i$ for $i = j+2 \leq i \leq n$

.../53...

e.g. $p = 5 \quad r = 4$

Let $\alpha = {}^{5}/_{4} = .0433$ then $\bar{\alpha} = {}^{-5}/_{4} = .0111$

C. Multiplication

Let α and β be the multiplicand and multiplier respectively

$$\alpha = a_{-m}\, a_{-m+1} \ldots\ldots\ldots a_{-1}\, a_0\, a_1 \ldots\ldots a_n$$

$$\beta = m_{-m}\, b_{-m+1} \ldots\ldots\ldots b_{-1}\, b_0\, b_1 \ldots\ldots b_n$$

The algorithm consists in forming

$$P_{ij} = b_i a_j \text{ for } i = -m, -m+1, \ldots, n \text{ and}$$

for the values of $j = -m, -m+1, \ldots, n-1$ and then forming the partial products p_i and the final product p shifts and additions as specified by the following recursions.

For each i form

$$P_{ij} = b_i a_j \text{ for } j = -m \ldots\ldots n - 1$$

Then obtain

$$P_i = \sum_{j=-m}^{n} P_{ij}\, \Delta\,(m + j)$$

Where $\Delta\,(x)$ indicate right shift of the number by x digits for $x \geq 0$ and $\alpha \cdot \beta = \sum_{j=-m}^{n} P_i\, \Delta\,(m+1) = \sum_{j=-m}^{n} P_i\, \Delta\,(m+1)$,

E.g.

Take $\alpha = \frac{1}{4} = .4333$

$\beta = \frac{1}{3} = .2313$ for $p = 5$ $r = 4$

$\alpha \cdot \beta = {}^{1}/_{12} = .3424$

See table (2)

.../54...

Table (2)

i_j	0	1	2	3
a_i	4	3	3	3
b_i	2	3	1	3
p_{oo}	3	1		
p_{o1}		1	1	
p_{o2}			1	1
p_{o3}				1
p_o	3	2	2	2
p_1		2	1	1
p_2			4	3
p_3				2
p	3	4	2	4

D. Division

Let $\alpha = .a_0\ a_1 \ldots\ldots a_n$ be the dividend

$\beta = .b_0\ b_1 \ldots\ldots b_n$ " " divisor

$\gamma = .q_0\ q_1 \ldots\ldots q_n$ " " quotient

The division algorithm consists of the following steps, we assume here $b_o \neq 0$ for $b_o = 0$ see below

Set $R_o = \alpha =$ Zeroth partial remainder.

.../55...

Let R_i denote the partial remainder at the ith stage and R_{ii} denote the ith digit of R_i , Then

$$q_i = R_{ii}\ b_o^{-1} \pmod{p} \text{ for } r = 0, 1, \ldots\ldots, n$$

where b_o^{-1} is the multiplicative inverse of b_o , the leading digit of β. The next partial remainder R_{i+1} is then formed using $R_{i+1} = R_i + q_i \beta \Delta(i)$ where

$\overline{\beta}$ is the complement of β and Δ (i) denotes right shift by i digits. The algorithm terminates when q_n is obtained.

Remark: If $b_o = 0$, shift the divisor left keeping count until the first digit is non-zero and suitably adjust the quotient. It is assumed that by shifting the divisor, it does not go outside the range of the order N Farey Fraction for the given H(p,r), in such a case the division operation is invalid.

E.g. $p = 5, \quad r = 4$

$\alpha = {}^{8}/_{9} = .2243$

$\beta = {}^{1}/_{2} = .3222$

$\gamma = {}^{16}/_{9} = .4432$

$b_o^{-1} = 2 \quad \overline{\beta} = .2222$

See table 3.

.../56...

Table 3

i	0	1	2	3
R_o	2	2	4	3
$q_0\overline{\beta}\Delta(0)$	3	4	4	4
R_i	0	2	4	3
$q_1\overline{\beta}\Delta(1)$	0	3	4	4
R_2	0	0	4	3
$q_2\overline{\beta}\Delta(2)$	0	0	1	2
R_3	0	0	0	1
$q_3\overline{\beta}\Delta(3)$	0	0	0	4
R_4	0	0	0	0
γ	4	4	3	2

from the above table R_{oo} equals 2 , then $q_0 = 2x2 \pmod 5$ since $b_o^{-1} = 2$.

4.4 Conversion Algorithms (Krishnarmurthy 1975)

It can be proved Rao (19 75) that the p-adic arithmetic with r-digits can be considered equivalent to the residue arithmetic modulo p^r except for the fact that the multiplicative inverse of b(GCD(b,p) $\neq$1) is taken as the inverse of the factor c where

$$b = p^e c \text{ with } GCD(p, c) = 1$$

.../57...

Therefore as in residue arithmetic, if a multiple $\phi = kb$ (k is integral of b is known or computed during the course of the algorithm then a/b can be computed. We call ϕ as the CODE. The computation of ϕ will be discussed with.

Since the p-adic is equivalent to residue arithmetic modulo p^r then we can introduce the idea of functions and function values

(i) For each integer a, $0 \leq a \leq p^r - 1$

$$\text{value } (a) = a \text{ if } 0 \leq a \leq \frac{p^r - 1}{2}$$

$$= a - p^r \text{ otherwise} \qquad 4.4.1$$

(ii) Let m_α be the mantissa part of a p-adic number, then $I(m_\alpha)$ is the integer given by

$$I(m_\alpha) = \sum_{i=0}^{r-1} \alpha_i p_i \qquad 4.4.2$$

Let $\alpha = (m_\alpha, e_\alpha)$ be the H (p,r) code of α and $\phi = (m\phi, e\phi)$ be that of the code. Since ϕ is an integer $e\phi = 0$. Note that the value (I $(m\phi)$) gives the sign and magnitude of ϕ

Let $\gamma = \alpha \phi = (m_\gamma, e_\gamma)$ say where $m_\gamma = m_\alpha . m_\phi$ and $e_\gamma = e_\alpha$ clearly ka and equivalent of $a/b = \frac{1}{ka} \left| ab^{-1}.kb \right|_p$ in residue arithmetic is given by

$$\frac{1 \text{ VALUE } (I \ (m))}{p^{e\alpha} \text{ VALUE I } (m\phi)} \qquad 4.4.3$$

Thus, given the H(p,r) code of a rational, it can be recovered with the knowledge of denominator or that of integral multiple. One way to know the denominator is to carry it along with the H(p,r) code of the rational throughout the computation and modify it along with the arithmetic

.../58..

operations. The first procedure section 4.4.1 is based on this principle. A different way to compute a multiple of the denominator of the result is to keep track of the denominators in the process. In this regard a 'CODE register" can be maintained and computing the code as a byproduct in the computational procedure.

4.4.1 Method of self-denominator Krishnamurthy (1975)
Here every rational number a/b is represented as an ordered pair (M, E), where M is the p-adic representation or the H(p,r) code (m_α , e_α) ofα and E (an integer) is the p-adic representation of the denominator b, (usually taken as positive). Then the basic arithmetic operations involving two operands(M_1, E_1) and (M_2, E_2) are carried out according to the following rules in order to arrive at the (M_3, E_3). Here +, -, ., and/between M_1, M_2 denote the p-adic arithmetic operations and $\oplus$, $\ominus$, $\odot$, between (M_1, E_1) and (M_2, E_2) denote the generalized p-adic floating denominator operations.

(a) Addition/Subtraction

$$(M_1, E_1) \oplus (M_2, E_2) = (M_3, E_3)$$

where $M_3 = M_1 \pm M_2$ and $E_2 \equiv \text{LCM}(E_1, E_2)$

(b) Multiplication

$$(M_1, E_1) \odot (M_2, E_2) = (M_3, E_3)$$

where $M_3 = M_1.M_2$ and $E_3 = E_1.E_2$

.../59...

(c) Division

$(M_1\ E_1) \oslash (M_2\ E_2) = (M_3\ ,\ E_3)$

Where $M_3 = M_1/M_2$ and $E_3 = E_1\ |M_2.E_2|$

where $|M_2.E_2|$ is the absolute value of $M_2.E_2$ in p-adic representation.

When the above representation is used then the rational equivalent of a number (M.E) is given by

$$\frac{\text{VALUE } (I(M.E))}{\text{VALUE } (I(E))} = \frac{\text{VALUE } (I(M.E.))}{I(E)}$$

since E is assumed positive 4.3.2

Example 4.4.1

Consider the evaluation of the expression

$$(({}^{1}/2 + {}^{3}/7)/({}^{3}/5 - {}^{4}/3) - ({}^{5}/12)$$

In Table 4, we tabulate the results using the above method.

.../60...

Table 4

Comments	Operands	Results
Initial	$^{1}/_{2}$	(.322222 E0; .200000)
Conversion	$^{3}/_{7}$	(.402141 E0; .210000)
(Rational to p-adic)	$^{3}/_{5}$	(.300000 E1; .010000)
	$^{4}/_{3}$	(.331313 E0; .300000)
	$^{5}/_{12}$	(.034242 E0; .220000)
	$\frac{1}{2} + \frac{3}{7}$	(.234310 E0, .420000)
	$\frac{3}{5} - \frac{4}{3}$	(.321313 E1; .030000)
	$\frac{\frac{1}{2} + \frac{3}{7}}{\frac{3}{5} - \frac{4}{3}}$	(.041232 E0; .401100)
	$\frac{\frac{1}{2} + \frac{3}{7}}{\frac{3}{5} - \frac{4}{3}} - \frac{5}{12}$	(.012434 E0; .441210)
Result		$(-1555/_{924})$

4.4.2 Algorithmic common denominator method. This method is more economical than the method of self denominator. It is based on the principle that every algorithm, has its own common denominator that can always be computed in advance or in parallel with the algorithmic operations. The output can be converted into rationals using

$$\alpha = \frac{1}{p^{e_\alpha}} \quad \frac{\text{VALUE (I (mc. m) mod } p^{r})}{\text{VALUE (I(mc))}} \qquad 4.4.4$$

.../61...

where $H(p, r, \alpha) = (m_\alpha , e_\alpha)$ and CONDEN = $(m_c, e_c), e_c = 0$ and COMDEN is the common denominator of the algorithm. The example of this method is given in the applications of p-adic arithmetic procedure to compute symmetrizer.

4.5 Application of p-adic arithmetic to computing symmetrizer and Reduction of a matrix to Hessenberg form.

As said earlier, a given prime number p and a positive integer $r(\geq 1)$ the rational numbers a/b $|a|$, $|b| \leq \sqrt{(p^r-1)/2}$ have a unique H(p, r) code. The choice of p and r should be such that all the input and output entries are representable uniquely by H(p,r) codes.

It can be showed that for Datta (1973) Algorithm for symmetrization of a given matrix A

$$p^r > 2^{2(n-2)} \; a^{n-1} \qquad 4.5.1$$

$$\text{or } r > \frac{1}{\log p} \log 2^{2(n-1)} \; a^{n-1} \qquad 4.5.2$$

4.5.1. The Datta Algorithm

$$X_i = \frac{1}{a_{i,\, i+1}} \left[X_{i+1} A - a_{i+1,\, i+1} X_{i+1} \cdots\cdots \right.$$

$$\left. \cdots - a_{n,\, i+1} X_n \right] \qquad 4.5.3$$

$$i = n - 1, \ldots\ldots 3, 2, 1.$$

With this algorithm the COMDEN can be computed in advance and so conversion into rational will therefore use method 4.4.2.

$$\text{COMDEN} == \sum_{i=1}^{n-1} a_{i,\, i+1} \quad \text{in p-adic.}$$

Example 4.5.1.

To calculate the symmetrizer x of the given matrix

$$A = \begin{bmatrix} 3 & 2 & 0 \\ 3 & 4 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

.../62...

Solution

Let P = 5 r = 4

p^r = 625 > 128

Step I Compute COMDEN

COMDEN = 2 x 2 = 4 .4000, E0.

Step II CONVERT all entries of A into p-adic.

$$A = \begin{bmatrix} .3000\ E0 & .2000\ E0 & .0000\ E0 \\ .3000\ E0 & .4000\ E0 & .2000\ E0 \\ .1000\ E0 & .2000\ E0 & .2000\ E0 \end{bmatrix}$$

Step III Choose X_n = (1, 0 0 0 .... 0)

X_3 = (.1000 E0, .0000 E0 .0000 E0)

Step III Compute X_{n-1}, ... X_1

Using the relation . 3.5.3

$$X_2 = \frac{1}{a_{23}} (X_3A - a_{33}x_3)$$

= (.3222 E0 .1000 E0 .0000 E0)

$$X_1 = \frac{1}{a_{12}} (x_2A - a_{22}x_2 - a_{32}x_3)$$

= (.4333 E0 .3222 E0 .1000 E0)

$$X = \begin{bmatrix} .4333\ E0 & .3222\ E0 & .1000\ E0 \\ .3222\ E0 & .1000\ E0 & .0000\ E0 \\ .100\ 0\ E0 & .0000\ E0 & .0000\ E0 \end{bmatrix}$$

Then conversion into the rational can be done for each element of the symmetrizer using equation 4.4.4.

..../63...

$$\text{e.g.} \quad X_{11} = \frac{\text{VALUE I } (.4333 \times .4000) \bmod 625)}{5^{\circ} \text{ VALUE I } (.4000)}$$

$$= \frac{1}{4}$$

$$\therefore \quad = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & 1 \\ \frac{1}{2} & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

4.5.2 Reduction of a matrix to the Hessenberg form.

The reduction of a general matrix to a Hessenberg form using p-adic arithmetic can be carried out in the following steps.

Step I Select the Prime and therefore find out r the length of the Hencel code

For the choice of p^r See page 67.

Step II Convert all entries of the given matrix to p-adic codes using the algorithm B of section 4.2.

Step III Carry out the transformations as explained in chapter one but interchanging rows and columns as well as choosing the largest element are unnecessary.

Step IV Compute the COMDEN

$$\text{COMDEN} = \prod_{i=1}^{n-2} a_{i,\, i+1}^{2}$$

Where $a_{i,\, i+1}$ is the current value after post and pre multiplication of A by m_j and m_j^{-1} respectively.

a_{12} been the a_{12} element of the original matrix.

.../64...

Step V Convert back into rationals using algorithm of common denominator method as explained in section 4.4.2.

The symmetrization can be carried out using the steps in 4.5.1.

Example 4.5.1.

Reduce

$$\begin{bmatrix} 5 & 4 & 6 & 8 \\ 2 & 3 & 2 & 1 \\ 4 & 3 & 4 & 2 \\ 3 & 3 & 1 & 4 \end{bmatrix}$$

to the Hessenberg form.

$P = 5,\ r = 6.$

$$\text{Step II A} = \begin{bmatrix} .010000 & .400000 & .110000 & .310000 \\ .200000 & .300000 & .200000 & .10000 \\ .400000 & .300000 & .400000 & .20000 \\ .300000 & .300000 & .100000 & .40000 \end{bmatrix}$$

$$m_{2j} = \begin{bmatrix} .100000 & .000000 & .000000 & .000000 \\ .000000 & .100000 & -\frac{.110000}{.400000} & -\frac{.310000}{.400000} \\ .000000 & .000000 & .100000 & .000000 \\ .000000 & .000000 & .000000 & .100000 \end{bmatrix}$$

$$= \begin{bmatrix} .100000 & .00\ 0000 & .000000 & .000000 \\ .000000 & .100000 & .122222 & .344444 \\ .000000 & .000000 & .100000 & .000000 \\ .000000 & .000000 & .000000 & .100000 \end{bmatrix}$$

.../65...

$$m_{2j}^{-1} = \begin{bmatrix} .100000 & .000000 & .000000 & .000000 \\ .000000 & .100000 & .422222 & .200000 \\ .000000 & .000000 & .100000 & .000000 \\ .000000 & .000000 & .000000 & .100000 \end{bmatrix}$$

$$AM^{(2)} = \begin{bmatrix} .010000 & .400000 & .000000 & .000000 \\ .200000 & .300000 & .022222 & .044444 \\ .400000 & .300000 & .222222 & .144444 \\ .300000 & .300000 & .412222 & .344444 \end{bmatrix}$$

$$(M^{(2)})^{-1} AM^{(2)} = \begin{bmatrix} .010000 & .400000 & .000000 & .000000 \\ .420000 & .103222 & .140111 & .024444 \\ .400000 & .300000 & .222222 & .144444 \\ .300000 & .300000 & .412222 & .344444 \end{bmatrix}$$

$$m_{3j} = \begin{bmatrix} .100000 & .000000 & .000000 & .000000 \\ .000000 & .100000 & .000000 & .000000 \\ .000000 & .000000 & .100000 & -\frac{.024444}{.140111} \\ .000000 & .000000 & .000000 & .100000 \end{bmatrix}$$

$$= \begin{bmatrix} .100000 & .000000 & .000000 & .000000 \\ .000000 & .100000 & .000000 & .000000 \\ .000000 & .000000 & .100000 & .033041 \\ .000000 & .000000 & .000000 & .100000 \end{bmatrix}$$

..../66...

$$m_{2j}^{-1} = \begin{bmatrix} .100000 & .000000 & .000000 & .000000 \\ .000000 & .100000 & .000000 & .000000 \\ .000000 & .000000 & .100000 & .021403 \\ .000000 & .000000 & .000000 & .100000 \end{bmatrix}$$

$$(M^{(2)})^{-1} \; A \; M^{(2)} \; M^{(3)}$$

$$= \begin{bmatrix} .010000 & .400000 & .000000 & .000000 \\ .420000 & .103222 & .140111 & .000000 \\ .400000 & .300000 & .222222 & .103421 \\ .300000 & .300000 & .412222 & .312201 \end{bmatrix}$$

$$(M^{(3)})^{-1} \; (M^{(2)})^{-1} \; A \; M^{(2)} \; M^{(3)}$$

$$= \begin{bmatrix} .010000 & .400000 & .000000 & .000000 \\ .420000 & .103222 & .140111 & .000000 \\ .414224 & .314224 & .200021 & .103421 \\ .300000 & .300000 & .41222 & .312201 \end{bmatrix}$$

COMDEN = .112320 I(C) = Integer (comden) = 1681

Then using $a_{ij} = \dfrac{\text{VALUE } (I(C\alpha_{ij}))}{p^{e\alpha} \text{ VALUE } (I(C))}$

With the above example e_α = E0 ; $p^{e\alpha} = 1$

$$S = \begin{bmatrix} 5 & 4 & 0 & 0 \\ 14 & \frac{27}{2} & \frac{-41}{4} & 0 \\ \frac{344}{} & \frac{303}{41} & \frac{-461}{82} & \frac{2186}{1681} \\ 3 & 3 & \frac{-7}{2} & \frac{128}{4} \end{bmatrix}$$

= Rational Hessenberg form.

.../67...

4.5.3. Choice of prime

The choice of primes for multiple moduli arithmetic or p and r for p-adic arithmetic dep ends on the maximum value any intermediate element may assume in the reduction process.

The following criteria for the selection of prime is suggested in Howell (1973) for the reduction of matrices to Frobenius form- since the procedures are similar, we also suggest that same criteria can be used for our purpose when reducing the matrix to the Hessenberg form:

For multiple prime

$$M \geq 2^{\left(\left[\frac{n}{2}\right]\right)^n} (\alpha)^n$$

or $p^r \geq 2^{\left(\left[\frac{n}{2}\right]\right)^n} (\alpha)^n$ for p -adic where is a bound on the maximum eigen-value of A, and n is the size of the matrix. ([ ] = lower integral part).

..../68....

CONCLUDING REMARKS

The objective of this dissertation is to describe error-free methods for transformation of a matrix to the Hessenberg form as well as symmetrization of matrix in the Hessenberg form.

In achieving this objective, Chapter 1 of this dissertation dealt with the reduction of a matrix to the Hessenberg form by the use of simple elementary similarity operations.

Chapter 2 was concerned with some applications of the symmetrizers and various methods of symmetrization.

Chapter 3 dealt with the basic concept of residue Arithmetic, this includes both the single prime modulus and multiple moduli Arithmetic, Also the use of Chinese remainder theorem and the algorithm for finding the multiplicative inverse of any given element over a finite field were described.

In the applications of the residue arithmetic, however, one important thing is the choice of the prime for a particular algorithm.

It was also observed that if any element $a_{i,\ i+1}$ turns out to be zero during the process of symmetrization, for a particular choice of prime, one has to choose a different prime and carry out the symmetrization of the matrix.

.../69...

Chapter 4 dealt with the finite segment (of length r digits) digits) p-adic Arithmetic called the Hensel Code of a rational number a/b denoted by H(p, r, α). The algorithm for conversion into the Hensel code was also dealth with. The algorithms for performing arithmetic operations were also described. For conversion back into the rationals when all operations have been performed, a number of methods were suggested, of which the method of common denominator has been found to be the most suitable for the matrix transformation algorithms.

In conclusion, however, the p-adic arithmetic offers a better method for the reduction of a matrix to the Hessenberg form and for the symmetrization procedure, since one works with a number p^r, the problem of any one of the elements of the form $a_{i,\ i+1}$ turning out to be zero does not arise. Also the use of Chinese remainder theorem is not required as the results are directly available in base p and only need be converted to the decimal base.

.../70...

B I B L I O G R A P H Y

(1) BACHMAN, G. (1964)
Introduction to p-adic Numbers and Valuation Theory;
Academic Press, New York and London.

(2) BARNETT, S. B. (1973)
Matrices, Polynomials, and Linear Time -
Invariant systems. I.E.E.E. Trans. On Aut. Cont.
Vol AC-18 No. 1 pp 1 - 10

(3) BARNETT, S and STOREY, C (1970)
Matrix Methods in stability Theory
Nelson, London.

(4) BARNETT, S. and STOREY, C. (1971)
Matrices In Control Theory
London, Van Nostrand, Reinhold.

(5) BOREVICH, Z. I. and SHAFAREVICH, I. R. (1966)
Number Theory, Academic Press, New York.

(6) BOROSH, I. and FRAENKEL, A. S. (1966)
Exact solutions of linear Equations with Rational
Coefficients by Congruence Techniques, Mathematics
of Computation Vol. 20 pp. 107 - 112.

(7) DATTA, B. N. (1972)
Quadratic forms, matrix equations and the matrix
eigenvalue problem. Ph.D. Thesis.
University of Ottawa, Canada.

(8) DATTA, B. N. (1973)
An Algorithm for computing a symmetrizer of a
Hessenberg matrix, to appear, abstract appeared
in Notices, Amer. Mathematics Society Feb. (1973)

(9) DATTA, B. N. (1974)
A constructive method for finding the Schwarz
form of a Hessenberg matrix.
I.E.E.E. Trans. Aut. Cont. AC 19 pp 616 - 617.

.../71..

(10) DATTA, B. N. (1973)
An Algorithm for computing a symmetrizer of an arbitrary matrix (to appear)

(11) DATTA, B. N. (1975)
On the Computation of Routh Canonical form of a Hessenberg Matrix.
I.E.E.E. Aut. Cont. Vol. 20 issue 4.

(12) DATTA, B. N. and DATTA, K. (1975)
A numerical method for the solution of the stability equations (to appear)

(13) DATTA, B. N. and DATTA, K. (1975)
An algorithm for Computing powers of a Hessenberg matrix and its applications (to appear)

(14) FORSYTHE, G. and MOLER, C. B. (1967)
Computer solution of linear Algebraic Systems.
Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

(15) GANTMACHER, F. R. (1959)
The Theory of Matrices Vol. I and II, Chelsea Publishing Co. New York.

(16) HARDY, G. H. and WRIGHT, E. M. (1960)
An Introduction to Theory of Numbers; Oxford At The clarendon Press.

(17) HOWELL, J. A. and GREGORY, R. T. (1969)
An Algorithm for solving Linear Algebraic Equations Using Residue Arithmetic I. B.I.T., Bind 9 Hefle NR3, pp. 200 - 204.

(18) HOWELL, J.A. and GREGORY, R. T. (1969)
An algorithm for solving Linear Algebraic Equations Using Residue Arithmetic II. B.I.T. Bind 9 Hefte NR 4, pp. 324 - 337.

..../72...

(19) HOWELL, J. A. and GREGORY, R. T. (1970)
Solving Linear Equations using Residue Arithemetic - Algorithms
B. I. T. VOL. 10 pp. 23 - 37.

(20) HOWELL, J. A. (1971)
Exact Solution of Linear Equations Using Residue Arithmetic Communications of the ACM Vol. 14
pp. 180 - 184.

(21) HOWELL J. A.
An Algorithm for the exact reduction of a Matrix to Frobenius form. Using Modular Arithmetic.
Math. Comp. Vol.27, (1973)
pp. 887 - 919 (Parts I and II)

(22) HOWLAND J. L. (1962)
On some methods for computing the roots of polynomials Information processing 1962. Proceedings of I.F.I.P. Congress 62 pp 116 - 121.

(23) HOWLAND J. L. (1961)
A method for Computing the real roots of Determinantal Equations. American Maths. Monthly Vol. 68, No. 3 pp 235 - 239.

(24) HOWLAND, J. L. (1972)
Rayleigh Quotient Iteration for Matrix Eigenvalues.
Manuscript, University of Ottawa, Ottawa.

(25) HOWLAND, J.L. and FARRELL, F. J. (1963)
Matrix symmetrization Methods for the Algebraic Eigenproblem A paper read at the 1963 ACM Conference, Denver Colorado.

(26) HOWLAND, J. L. and SENEZ, J.A. (1970)
A constructive method for the solution of the stability problem. Num. Math. 16 pp 1 - 7.

.../73...

(27) KNUTH, D. E. (1969)
The Art of Computer Programming Vol. 2
Seminumerical Algorithms,
Addison - Wesley, Reading, Mass.

(28) KRISHNAMURTHY, E. V. (1975)
Matrix Processors using p-adic Arithmetic for Exact Linear Computations (Proc. Third Symp. on Computer Arithmetic IEEE Computer Society, Southern Methodist University Dallas, Taxas, Nov. 1975, pp. 92 - 97.

(29) KRISHNAMURTHY, E. V. MAHADEVA RAO, T. and SUBRAMANIAN, K (1975) Finite Segment p-adic Number Systems with Applications to Exact Computations
Proc. Indian Academy of Science pp 58 - 79.

(30) KRISHNAMURTHY, E. V. MAHADEVA RAO, T. and SUBRAMANIAN, K. (1975) p-adic Arithmetic procedures for Exact matrix Computations Proc. Ind. Academy of Science December (1975)

(31) KRISHNAMURTHY, E. V. and SEN, S. K. (1975)
Computer Based Numerical Algorithms, Affiliated East - West Press (D. Von Nostrand), New Delhi, India

(32) MAHADEVA RAO, T. (1975)
Finite Field Computation Techniques for Exact Solution of Numerical Problems.
Ph.D. Thesis, Indian Institute of Science, Bangolore.

(33) MAHLER, K. (1973)
Introduction to p-adic Numbers and Their Functions, Cambridge University Press.

.../74...

(34) OLGA TAUSSKY and HANS ZASSENHAUS (1959)
On the Similarity Transformation between a matrix and its transpose
Pacific J. Math Vol. 9 pp 893 - 896.

(35) RICHARDS, R. K. (1955)
Arithmetic Operations in Digit Computers
D. Van Nostrand and Co. Inc. Princeton.

(36) STEWARTS, G. W. (1973)
Introduction to Matrix Computations
Academic Press, New York.

(37) SZABO, N. S. and TANAKA, R. I. (1967)
Residue Arithmetic and its Applications to Computer Technology McGrawal Hill Book Co. New York.

(38) WILKINSON, J. H. (1958)
Stability of the reduction of a matrix to almost Triangular forms by Elementary Similarity Transformations National Physical Laboratory Ma/43.

(39) WILKINSON, J. H. (1965)
The Algebraic Eigenvalue Problem
Oxford University Press, London.

(40) WILKINSON, J. H., MARTIN, K. S. and PETERS G. (1970)
The QR Algorithm for Real Hessenberg Matrices .
Numer. Math. 14 pp 219 - 231.

(41) YOUNG, D. M. and GREGORY, R. T. (1972)
A Survey of Numerical Mathematics
Vol. I and Vol. II, Addison Wesley, Reading, Mass. .