A ROUND-OFF THEORY FOR SCALAR PRODUCTS

by

R. Deane Branstetter

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subject: Mathematics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Head of/Major Department

Signature was redacted for privacy.

Dean of Graduate College

Iowa State College
1953

TABLE OF CONTENTS

Page

I. INTRODUCTION 1

A. Types of Error 1
B. Digital Numbers and Round-off 3

II. PSEUDO-OPERATIONS FOR SCALARS 6

III. A MODIFICATION OF THE BINGHAM METHOD FOR INVERTING MATRICES 12

A. Description of the Bingham Method 12
B. Description of the Modified Bingham Method . 15

IV. PSEUDO-OPERATIONS FOR MATRICES 19

A. Definitions and Some Properties of Pseudo-Operations for Matrices 19
B. Left and Right Pseudo-Multiplication 21

V. SYMMETRIC PROPERTIES 30

VI. SCALING OPERATIONS 38

A. Exact Considerations 38
B. Statistical Considerations 41

VII. COMPARISON OF EXACT SOLUTIONS AND PSEUDO-SOLUTIONS OF A LINEAR EQUATION 51

A. Discussion of Some of the Problems 51
B. Definitions and Some Properties of Pseudo-Solutions 52
C. Criterion for an Exact Solution to Be a Pseudo-Solution 59
D. Discussion of the Two, Three and Four Dimensional Cases 62
E. The Limit of the Ratio of Certain Volumes as the Dimension Increases Without Bound 69

TABLE OF CONTENTS (continued)

Page

VIII. SUMMARY 77

IX. SUGGESTIONS FOR FURTHER STUDY 80

X. BIBLIOGRAPHY 81

A. Literature Cited 81
B. Other Bibliography 82

XI. ACKNOWLEDGMENT 83

I. INTRODUCTION

A. Types of Error

Numerical calculations of problems in pure or in applied mathematics may cause errors to enter into the solution. The term error means a deviation, of the numerical solution obtained, from the exact solution. Unless a measure of such errors is possible, it is meaningless to talk about such a solution. There is more than one source for such errors. In order to distinguish between the different types of errors, consider the system of linear equations

$$(1.1) \qquad \sum_{j=1}^{n} a_{ij}x_j = b_i \qquad\qquad i = 1,2,3,\cdots,n \; .$$

First, such a system of equations may only idealize the true relationship that exists between the x_i. The exact solution of such an idealized relationship may deviate from the true values of the x_i. This then may be a source of error. Second, the a_{ij} may be parameters measured by empirical means or computed directly from theoretical considerations of measured observations. If the a_{ij} are not known exactly or cannot be represented exactly, this may be a source of error of the solution.

These two sources of error are important in the ultimate analysis of the total accumulative error, but, in this thesis, it will be assumed that the expressed relationships between the variables are correct and that the a_{ij} have been expressed exactly and within the digital capacity of the computing devices that will be used. This thesis will be a study of errors which are due to computing procedures of elementary operations. Due to the limited capacity of computing machines it is usually necessary to round-off products and quotients as they are performed. For example, if the capacity of a standard computing machine is eight digits, then (.12345678)(.64196966) is exactly equal to .0792555070812948, but it is rounded-off to .07925551 in order that subsequent operations with this number may be performed on the machine. The amount of error due to round-off is a function of the number of digits retained. If it were possible to keep all digits, there would be no error due to round-off; but since this is generally not possible, round-off errors occurring at different stages of the computing process must be duly considered since they may accumulate to a sizable error in the final solution.

Using most of the definitions, symbols and basic inequalities concerning pseudo-operations as given by J. Von Neumann and H. H. Goldstine (1) this thesis first gives some generalizations of their theory and then applies it to the process of inverting matrices. In particular, a modification of the Bingham Method for inverting matrices is introduced. Next a strict approach and a probabilistic approach to the operation of scaling is presented. The last chapter is devoted to the analysis of the effect of pseudo-checking the exact solution of a linear equation.

B. Digital Numbers and Round-off

Since the computational work will be upon the a_{ij} of equation (1.1) and also upon the matrix with elements a_{ij}, the nature of such numbers will now be given. The a_{ij} are digital numbers. A digital number $\bar{x}$ is an s-place, base β, digital aggregate with sign

$$\bar{x} = \epsilon(\alpha_1,\cdots,\alpha_s);$$

$$\epsilon = \pm 1; \ \alpha_1,\cdots,\alpha_s = 0,1,\cdots,\beta-1.$$

The sum and difference of two digital numbers will be denoted by $\bar{x} \pm \bar{y}$ and will have their ordinary meaning. It is true that the sum of two digital numbers may be an (s+1)-place digital aggregate and therefore not a digital number; but it will be assumed that such a number does not exceed the capacity of the computing machine in performing subsequent operations.

The product and quotient of two digital numbers will be rounded-off to s-place aggregates and such quantities will be called pseudo-products and pseudo-quotients. A pseudo-product will be denoted by $\bar{x} \times \bar{y}$ and a pseudo-quotient by $\bar{x} \div \bar{y}$.

If a digital number $\bar{x}$ is multiplied by an integer $m(0,\pm 1, \pm 2,\cdots)$ a number $m\bar{x}$ is obtained which is not subject to round-off since such an operation is thought of as repeated additions or subtractions. It is true that if m is large, then $m\bar{x}$ may exceed the capacity of the machine but by a slight revision of the addition or subtraction operation $m\bar{x}$ can be computed with no round-off involved.

If two digital numbers are multiplied together and rounded-off to an s-place aggregate, it is obvious that the magnitude of the round-off error depends on the location of the decimal points or more generally the β-adic points of the two digital numbers. In all operations that follow, therefore, it will generally be assumed that the decimal point of β-adic point is located at the extreme left of all s-place digital numbers. It is always possible to force this condition upon digital numbers by the introduction of proper scale factors. Scale factors will be introduced later. If the β-adic point is at the extreme left, this means that all digital numbers lie in the open interval (-1,1).

If two digital numbers are multiplied, a 2s-place number is obtained. If this is rounded-off to an s-place number and if both digital multipliers had their β-adic point at their extreme lefts, then the round-off error is numerically less than $\frac{\beta^{-s}}{2}$.

Since rounding-off numbers is not uniformly performed, the rules that follow will govern the round-off referred to in this thesis. It is true that some automatic high-speed calculating machines do not have a rounding-off operation, but it will also be assumed for this thesis that all machines that are used will be able to perform the following rules for round-off.

1. If the discarded digits amount to 5 or more than 5 in the position of the first discarded digit, then 1 is added to the last digit retained.
2. If the discarded digits amount to less than 5 in the position of the first discarded digit, no change is made in the digits retained.

II. PSEUDO-OPERATIONS FOR SCALARS

Since it is necessary to compute powers of the matrix A in order to determine its inverse by the Bingham method, a study of the properties of pseudo-multiplication and pseudo-divivion will be made. The laws of multiplication and division of digital numbers will be studied first. Later, pseudo-multiplication of matrices will be discussed. The usual associative law of multiplication, distributive law and inverse relationship between multiplication and division are affected by the pseudo-operations. The commutative law of multiplication using pseudo-operations remains valid, that is $\bar{x} \times \bar{y} = \bar{y} \times \bar{x}$. This is true since $\bar{x} \times \bar{y}$ means to compute $\bar{x}\ \bar{y}$ and then round-off. Now since $\bar{x}\ \bar{y}$ and $\bar{y}\ \bar{x}$ are equal, $\bar{x} \times \bar{y}$ and $\bar{y} \times \bar{x}$ would be rounded-off to the same value.

The basic inequalities involved are

$$(2.1) \qquad |\bar{x} \times \bar{y} - \bar{x}\ \bar{y}| \le \beta^{-s}/2$$

$$(2.2) \qquad |\bar{x} \div \bar{y} - \bar{x}/\bar{y}| \le \beta^{-s}/2.$$

Using these inequalities, the distributive law of multiplication can be analyzed. It follows, for instance, that

$$(\bar{x} + \bar{y}) \times \bar{z} - (\bar{x} \times \bar{z} + \bar{y} \times \bar{z}) = (\bar{x} + \bar{y}) \times \bar{z} - (\bar{x} + \bar{y})\bar{z} + \bar{x}\ \bar{z}$$

$$-\bar{x} \times \bar{z} + \bar{y}\ \bar{z} - \bar{y} \times \bar{z},$$

since the exact product $(\bar{x} + \bar{y})\bar{z}$ has been added and subtracted and the distributive law for exact multiplication is true. Therefore,

$$|(\bar{x} + \bar{y})\times\bar{z} - (\bar{x}\times\bar{z} + \bar{y}\times\bar{z})| \le 3\beta^{-s}/2 .$$

Since the left hand member is the difference of two s-place numbers, it is apparent that such a difference is an integer multiple of β^{-s}, therefore,

$$|(\bar{x} + \bar{y})\times\bar{z} - (\bar{x}\times\bar{z} + \bar{y}\times\bar{z})| \le \beta^{-s} .$$

These results can be found in the paper by J. Von Neumann and H. H. Goldstine (1). As an extension of these results, one obtains

$$\begin{aligned} &|(\bar{x}_1 + \bar{x}_2 + \cdots + \bar{x}_n)\times\bar{y} - (\bar{x}_1\times\bar{y} + \bar{x}_2\times\bar{y} + \cdots + \bar{x}_n\times y)| \\ &\qquad \le [(n+1)/2]\ \beta^{-s} \end{aligned} \tag{2.3}$$

where $[(n + 1)/2]$ means the largest integer less than or equal to $(n + 1)/2$. The distributive law may not hold then for pseudo-multiplication. Equation (2.3) gives the maximum error that could occur.

Although the distributive law for pseudo-multiplication does not always hold there is a particular case in which it does. Since this case will be utilized in the last chapter, the following theorem will be proved.

THEOREM 1. If $\bar{a}\ \bar{b}$ (mod an integer) is a digital number then $\bar{a}\times(\bar{b} + \bar{c}) = \bar{a}\times\bar{b} + \bar{a}\times\bar{c}$. In other words, the distributive law for pseudo-multiplication holds in this case.

Proof. Since $\bar{a} \times (\bar{b} + \bar{c})$ is to be computed by exactly multiplying $\bar{a}$ times $(\bar{b} + \bar{c})$ and then rounding-off the product, one obtains

$$\bar{a} \times (\bar{b} + \bar{c}) = \bar{a}(\bar{b} + \bar{c}) + \varepsilon_1$$

where ε_1 is the round-off error contributed by rounding-off $\bar{a}\ \bar{c}$. The right member of the above equation is equal to $\bar{a}\ \bar{b} + \bar{a}\ \bar{c} + \varepsilon_1$ since the distributive law holds for exact multiplication. But

$$\bar{a}\ \bar{b} + \bar{a}\ \bar{c} + \varepsilon_1 = \bar{a} \times \bar{b} + \bar{a} \times \bar{c}$$

since $\bar{a}\ \bar{b} = \bar{a} \times \bar{b}$ and $\bar{a}\ \bar{c} + \varepsilon_1 = \bar{a} \times \bar{c}$ by definition. This completes the proof.

The associative law of pseudo-multiplication can be investigated as follows. Since

$$\bar{x} \times (\bar{y} \times \bar{z}) - \bar{x}\ \bar{y}\ \bar{z} = \bar{x} \times (\bar{y} \times \bar{z}) - \bar{x}(\bar{y} \times \bar{z}) + \bar{x}\ \ (\bar{y} \times \bar{z}) - \bar{x}\ \bar{y}\ \bar{z},$$

then

$$|\bar{x} \times (\bar{y} \times \bar{z}) - \bar{x}\ \bar{y}\ \bar{z}| \leq \beta^{-s}/2 + |\bar{x}|\ \beta^{-s}/2 = \beta^{-s}/2\ (1 + |\bar{x}|).$$

Also

$$(\bar{x} \times \bar{y}) \times \bar{z} - \bar{x}\ \bar{y}\ \bar{z} = (\bar{x} \times \bar{y}) \times \bar{z} - (\bar{x} \times \bar{y})\bar{z} + (\bar{x} \times \bar{y})\bar{z} - \bar{x}\ \bar{y}\ \bar{z},$$

so

$$|(\bar{x} \times \bar{y}) \times \bar{z} - \bar{x}\ \bar{y}\ \bar{z}| \leq \beta^{-s}/2 + |\bar{z}|\ \beta^{-s}/2 = \beta^{-s}/2\ (1 + |\bar{z}|).$$

This means that

$$(2.4) \qquad |\bar{x} \times (\bar{y} \times \bar{z}) - (\bar{x} \times \bar{y}) \times \bar{z}| \leq \beta^{-s}/2\ \left[2 + |\bar{x}| + |\bar{z}|\right].$$

If $|\bar{x}|$, $|\bar{z}| < 1$, then the left member of equation (2.4) will be $< 2\beta^{-s}$. Now since the difference between two digital numbers is an integer multiple of β^{-s}, the left member of equation (2.4) will be $\leq \beta^{-s}$. If $|\bar{x}| = |\bar{z}| = 1$, the left member of equation (2.4) will be zero. Thus, for all $|\bar{x}|$, $|\bar{z}| \leq 1$,

$$|(\bar{x} \times \bar{y}) \times \bar{z} - \bar{x} \times (\bar{y} \times \bar{z})| \leq \beta^{-s} .$$

This result is given by J. Von Neumann and H. H. Goldstine (1).

To show that the difference between these associated pseudo-products can be as much as β^{-s}, the following example is cited using $\beta = 10$ and $s = 3$.

$$(.986 \times .749) \times .837 = .619,$$

$$.986 \times (.749 \times .837) = .618 .$$

The associative law of multiplication with n factors will now be studied using right pseudo-multiplication. One obtains the identity

$$\begin{aligned} &((\cdots((\bar{x}_1 \times \bar{x}_2) \times \bar{x}_3) \cdots) \times \bar{x}_n - \bar{x}_1 \, \bar{x}_2 \, \bar{x}_3 \cdots \bar{x}_n \\ &= ((\cdots((\bar{x}_1 \times \bar{x}_2) \times \bar{x}_3) \cdots) \times \bar{x}_{n-1}) \times \bar{x}_n \\ &- ((\cdots((\bar{x}_1 \times \bar{x}_2) \times \bar{x}_3) \cdots) \times \bar{x}_{n-1}) \bar{x}_n \\ &+ ((\cdots((\bar{x}_1 \times \bar{x}_2) \times \bar{x}_3) \cdots) \times \bar{x}_{n-1}) \bar{x}_n \\ &- ((\cdots((\bar{x}_1 \times \bar{x}_2) \times \bar{x}_3) \cdots) \, \bar{x}_{n-1}) \bar{x}_n \\ &+ (\cdots((\bar{x}_1 \times \bar{x}_2) \times \bar{x}_3) \cdots) \bar{x}_{n-1} \bar{x}_n \\ &\pm \cdots + (\bar{x}_1 \times \bar{x}_2) \, \bar{x}_3 \, \bar{x}_4 \cdots \bar{x}_n - \bar{x}_1 \, \bar{x}_2 \cdots \bar{x}_n . \end{aligned}$$

Taking the absolute value of both sides of this identity, one obtains

$$\begin{aligned} &|((\cdots((\bar{x}_1 \times \bar{x}_2) \times \bar{x}_3) \cdots) \times \bar{x}_{n-1}) \times \bar{x}_n - \bar{x}_1 \, \bar{x}_2 \, \bar{x}_3 \cdots \bar{x}_n| \\ (2.5)\quad & \\ &\leq \beta^{-s/2} \left[1 + |\bar{x}_n| + |\bar{x}_n \, \bar{x}_{n-1}| + \cdots + \Big| \prod_{i=n}^{3} \bar{x}_i \Big| \right] . \end{aligned}$$

In a similar manner the upper bound to the numerical difference between the pseudo-product multiplied in the reverse order and the exact product can be obtained. Combining these two results, it follows that

$$|((\cdots((\bar{x}_1\times\bar{x}_2)\times\bar{x}_3)\cdots)\times\bar{x}_n)-(\bar{x}_1\times(\bar{x}_2\times(\cdots(\bar{x}_{n-1}\times\bar{x}_n)\cdots))|$$

(2.6)

$$\leq \beta^{-s/2}\left[2+\sum_{i=3}^{n}|\bar{x}_i\cdots\bar{x}_n|+\sum_{k=1}^{n-2}|\bar{x}_1\cdots\bar{x}_k|\right].$$

Of particular interest is the case where $\bar{x}_i = \bar{a}$ for all i. The left member of equation (2.6) obviously reduces to zero for this case. The left member of equation (2.5) is numerically less than $\beta^{-s/2}\left[1+|\bar{a}|+\cdots+|\bar{a}^{n-2}|\right]$, which is finite even if n becomes infinite provided $|\bar{a}| \leq 1$. For example, if $|\bar{a}| \leq 1/2$, regardless of how many times $\bar{a}$ is pseudo-multiplied by itself, the difference between this result and the exact product is less than or equal to β^{-s}. It is obvious that actually the difference approaches zero as $n \longrightarrow \infty$.

Next, if one has to pseudo-multiply $\bar{x}$ by $\bar{y}$ and pseudo-divide by $\bar{z}$, is there a preferred order? In other words, is it better to calculate $(\bar{x}\times\bar{y})\div\bar{z}$ or $(\bar{x}\div\bar{z})\times\bar{y}$ or is the order immaterial? Now since

$$(\bar{x}\times\bar{y})\div\bar{z}-\frac{\bar{x}\,\bar{y}}{\bar{z}}=(\bar{x}\times\bar{y})\div\bar{z}-\frac{\bar{x}\times\bar{y}}{\bar{z}}+\frac{\bar{x}\times\bar{y}}{\bar{z}}-\frac{\bar{x}\,\bar{y}}{\bar{z}},$$

then

$$(2.7)\qquad |(\bar{x}\times\bar{y})\div\bar{z}-\frac{\bar{x}\,\bar{y}}{\bar{z}}|\leq \beta^{-s/2}\ (1+|\bar{z}|^{-1}).$$

Next

$$(\bar{x} \div \bar{z})\times\bar{y} - \frac{\bar{x}\,\bar{y}}{\bar{z}} = (\bar{x} \div \bar{z})\times\bar{y} - (\bar{x} \div \bar{z})\,\bar{y} + (\bar{x} \div \bar{z})\bar{y} - \frac{\bar{x}\,\bar{y}}{\bar{z}}$$

so

$$(2.8) \qquad \left|(\bar{x} \div \bar{z})\times\bar{y} - \frac{\bar{x}\,\bar{y}}{\bar{z}}\right| \leq \beta^{-s}/2 \ (1 + |\bar{y}|) .$$

Now, since $|\bar{z}|$ and $|\bar{y}| \leq 1$, equations (2.7) and (2.8) show that for all $|\bar{z}|$ and $|\bar{y}| \leq 1$, generally speaking, it is better to first divide and then multiply. This result is also given by J. Von Neumann and H. H. Goldstine (1).

III. A MODIFICATION OF THE BINGHAM METHOD FOR INVERTING MATRICES

A. Description of the Bingham Method

If $A = (a_{ij})$ is a given square matrix of order n, one can form the matrix $\lambda I - A$, called the characteristic matrix of A. The determinant of this matrix is called the characteristic function of A and is a polynomial of degree n in λ.

Setting $f(\lambda) = |\lambda I - A|$, then $f(\lambda) = \lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1} \lambda + a_n$. From this, one sees that $a_n = f(0) = |-A|$; or that, $a_n = (-1)^n |A|$. The algebraic equation, $f(\lambda) = 0$, is called the characteristic equation of the matrix A, and the roots of this equation are called the characteristic roots of A. The Cayley-Hamilton Theorem, upon which the Bingham Method of inverting matrices depends, will now be stated but not proven.

The Cayley-Hamilton Theorem: Let

$$f(\lambda) = \lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n$$

be the characteristic function of a matrix A, and let I and 0 be the unit matrix and zero matrix respectively with an order equal to that of the order of A. Then the matric polynomial equation

$$X^n + a_1X^{n-1} + \cdots + a_{n-1} X + a_nI = 0$$

is satisfied by X = A.

If A has an inverse A^{-1}, then the determinant of A is not equal to zero. Therefore, $a_n \neq 0$, so it follows that

$$I = (-1/a_n)\ (A^n + a_1A^{n-1} + \cdots + a_{n-1}\ A)\ .$$

The right hand member can be written

$$A\left[(-1/a_n)\ (A^{n-1} + a_1A^{n-2} + \cdots + a_{n-1}\ \ I)\right]$$

which means that

$$(3.1)\qquad A^{-1} = (-1/a_n)\ (A^{n-1} + a_1A^{n-2} + \cdots + a_{n-1}\ \ I).$$

From this equation A^{-1} can be computed if the values of the a_i and A^i are known. To compute the a_i, one must first compute the trace of the matrix A and the traces of the powers of A. The traces of the matrix A is defined by $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$. Next define the numbers $s_1, s_2, \cdots, s_n$ by

$$(3.2)\qquad s_1 = \mathrm{tr}(A),\qquad s_2 = \mathrm{tr}(A^2),\cdots,\qquad s_n = \mathrm{tr}(A^n).$$

The following recursion formulas, known as Newton's formulas, can be used to compute the a_i :

$$(3.3)\qquad \begin{aligned} a_1 &= -s_1\ , \\ a_2 &= (-1/2)(a_1\ s_1 + s_2)\ , \\ a_3 &= (-1/3)(a_2s_1 + a_1s_2 + s_3)\ , \\ &\ \ \vdots \\ a_n &= (-1/n)(a_{n-1}s_1 + a_{n-2}s_2 + \cdots + a_1s_{n-1}+s_n). \end{aligned}$$

One can summarize the rule for inverting matrices by the Bingham Method as follows:

1. Compute A^k, $k= 1,2,3,\cdots,n-1$.
2. Compute the diagonal elements of A^n.
3. Compute $s_1,s_2,\cdots,s_n$ by equation (3.2).

4. Compute a_i by equation (3.3).

5. Compute A^{-1} by equation (3.1).

The following example will illustrate the use of the Bingham Method for inverting matrices.

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

$$A^2 = \begin{pmatrix} 11 & 3 & 1 \\ 6 & 1 & 1 \\ -2 & -1 & 0 \end{pmatrix}$$

$$A^3 = \begin{pmatrix} 39 & & \\ & 5 & \\ & & -1 \end{pmatrix}$$

$s_1 = 4 \qquad s_2 = 12 \qquad s_3 = 43$

$a_1 = -4$

$a_2 = (-1/2)(-16 + 12) = 2$

$a_3 = (-1/3)(2(4) + (-4)(12) + 43) = -1$

$$A^{-1} = -(1/-1)\left[\begin{pmatrix} 11 & 3 & 1 \\ 6 & 1 & 1 \\ -2 & -1 & 0 \end{pmatrix} -4\begin{pmatrix} 3 & 1 & 0 \\ 2 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix} +2\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right]$$

$$= \begin{pmatrix} 1 & -1 & 1 \\ -2 & 3 & -3 \\ -2 & 3 & -2 \end{pmatrix}$$

B. Description of the Modified Bingham Method

One of the principal differences between standard elimination procedures used for computing the inverse of a matrix and the Bingham Method is that the powers of the matrix are computed in the Bingham Method. This poses quite a storage problem if the punch card method is used. It poses quite a memory storage problem if electronic methods are used. This problem of storage becomes increasingly more acute as n increases in magnitude. For this reason, it seems necessary to modify the Bingham Method in order to conserve the space available as a memory. A modification which reduces the storage space required during any particular interval of time is therefore presented here.

If the inverse of the matrix A exists, it is given by the following:

$$A^{-1} = (-1/a_n)(A^{n-1} + a_1A^{n-2} + a_2A^{n-3} + \cdots + a_{n-1}\, I).$$

Multiplying by a_n, one obtains

$$a_nA^{-1} = -\,(A^{n-1} + a_1A^{n-2} + a_2A^{n-3} + \cdots + a_{n-1}\, I).$$

Also

$$a_nA^{-1} + a_{n-1}\, I = -A\,(A^{n-2} + a_1A^{n-3} + \cdots + a_{n-2}\, I)\;.$$

Now let

$$B_1 = A^{n-2} + a_1A^{n-3} + \cdots + a_{n-2}\, I,$$

or

$$B_1 - a_{n-2}\, I = A\ (A^{n-3} + a_1A^{n-4} + \cdots + a_{n-3}\, I).$$

Let

$$B_2 = A^{n-3} + a_1A^{n-4} + \cdots + a_{n-3}\, I,$$

or in general,

(3.4) $B_i = A^{n-i-1} + a_1A^{n-i-2} + \cdots + a_{n-i-1}\, I, \quad i=1,2,\cdots,n-1$

if $a_o = 1$ and $A^{\circ} = I$.

The recurrence formula which enables one to compute the B_i is

(3.5) $$B_{i-1} = AB_i + a_{n-i}\, I.$$

The inverse of the matrix A is now computed by the formula

(3.6) $$A^{-1} = (-1/a_n)(AB_1 + a_{n-1}\, I)\ .$$

The rule for inverting matrices by the modified Bingham Method is summarized as follows:

1. Compute s_1 by formula (3.2).
2. Compute a_1 by formula (3.3).
3. Compute $B_{n-2} = A + a_1I$.
4. Compute A^2.
5. Compute s_2 by formula (3.2) and a_2 by formula (3.3).
6. Compute B_{n-3} by recurrence formula (3.5).
7. Compute A^3, s_3, a_3 and B_{n-4} in this order

> and continue this procedure until B_1 and a_n have been computed. Then compute A^{-1} by formula (3.6).

This means that in the memory storage, one must keep all the computed a_i and s_i, A, the current power of A and the current B_{n-1}. This cuts the maximum amount of storage space required at any particular time by about the factor n/3. For example, if n = 100, this would mean that only about 1/33 of the storage space required by the Bingham Method is required for this modified method.

The following example will illustrate the use of the Modified Bingham Method for inverting matrices.

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

$$s_1 = 4 \qquad a_1 = -4$$

$$B_2 = I \qquad B_1 = AI + a_1I = \begin{pmatrix} -1 & 1 & 0 \\ 2 & -4 & 1 \\ 0 & -1 & -3 \end{pmatrix}$$

$$A^2 = \begin{pmatrix} 11 & 3 & 1 \\ 6 & 1 & 1 \\ -2 & -1 & 0 \end{pmatrix}$$

$$s_2 = 12 \qquad a_2 = 2$$

$$A^3 = \begin{pmatrix} 39 & & \\ & 5 & \\ & & -1 \end{pmatrix}$$

$$s_3 = 43 \qquad\qquad a_3 = -1$$

$$A^{-1} = +1 \left[\begin{pmatrix} 3 & 1 & 0 \\ 2 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 2 & -4 & 1 \\ 0 & -1 & -3 \end{pmatrix} + \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \right]$$

$$= \begin{pmatrix} -1 & -1 & 1 \\ -2 & 1 & -3 \\ -2 & 3 & -4 \end{pmatrix} + \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & -1 & 1 \\ -2 & 3 & -3 \\ -2 & 3 & -2 \end{pmatrix}$$

IV. PSEUDO-OPERATIONS FOR MATRICES

A. Definitions and Some Properties of Pseudo-Operations for Matrices.

Since powers of the matrix A are needed to compute the inverse of A, the errors involved in obtaining these powers are studied. Unless otherwise stated, digital matrices are the type referred to in the discussion. A matrix is said to be a digital matrix if its coefficients are digital numbers. Digital matrices are designated by $\mathbb{A}$, $\mathbb{B}$, $\mathbb{C}$, $\mathbb{D}$, and so forth.

To determine the calculated powers of a matrix, pseudo-multiplication of matrices must be defined. The pseudo-product of a digital matrix A by a digital matrix B has coefficients c_{ij} obtained by the relationship

$$c_{ij} = \sum_{k=1}^{n} \bar{a}_{ik} \times \bar{b}_{kj} \quad .$$

The right member of this equation can be obtained in several ways. If two digital numbers are exactly multiplied, the result is a 2s-place number. Since subsequent operations may use these numbers, the last s-places are usually discarded and the remaining s-places rounded-off as described previously. On some automatic machines, these digits "spill off" to the right as the addends, making up the product, are computed. In other words, each of the pseudo-products $\bar{a}_{ik} \times \bar{b}_{kj}$ is obtained as an s-place number and the sum of these n digital numbers is then computed. If the round-off errors for each of these products were a maximum value and each of the same sign, then excessive errors due to the round-off would occur. In order to improve the estimate of the round-off error, it is assumed throughout this thesis, unless otherwise stated, that the c_{ij} are computed by a method called double precision multiplication. A description of this procedure follows.

In determining c_{ij}, first form the true 2s-place products $\bar{a}_{ik} \times \bar{b}_{kj}$ $(k=1,\cdots,n)$, then form their sum correctly to 2s-places and finally round-off this sum to s places. In general

$$\left|\sum_{k=1}^{n} \bar{a}_{ik} \times \bar{b}_{kj} - \sum_{k=1}^{n} \bar{a}_{ik} \bar{b}_{kj}\right| \leq \beta^{-s}/2$$

if double precision is used and

$$\left|\sum_{k=1}^{n} \bar{a}_{ik} \times \bar{b}_{kj} - \sum_{k=1}^{n} \bar{a}_{ik} \bar{b}_{kj}\right| \leq n\beta^{-s}/2$$

if the pseudo inner product were obtained by ordinary precision multiplication. This result is given by J. Von Neumann and H. H. Goldstine (1). This shows a definite advantage for the double precision method. However, one should keep in mind that it is sometimes much more difficult to perform double precision multiplication.

If two digital matrices A and B are multiplied together according to the laws of pseudo-multiplication and double precision multiplication is utilized, the pseudo-product is denoted by $A \times B$. The resulting matrix may not be a digital matrix. If $B = A$, then the pseudo-product is denoted by $A \times A$ or $\bar{A}^2$, the latter being read the pseudo A squared, the computed square of the digital matrix A, the computed A squared or generally in this thesis as A squared bar.

It is worthwhile to note that $A \times (B \times C)$ and $(A \times B) \times C$ are generally not equal. Neither are $A \times (B + C)$ and $A \times B + A \times C$ generally equal but a very important case where the distributive law holds for pseudo-multiplication of matrices is given in the following theorem.

THEOREM 2. Regardless of whether the symbol x refers to ordinary or double precision multiplication,

$$A \times (I + B) = A + A \times B.$$

Proof. $A \times (I + B) = A(I + B) + G$ where G is the error matrix contributed to only by the round-off of AB. Therefore the right hand member can be written as $A + AB + G$ $= A + A \times B$. This completes the proof.

B. Left and Right Pseudo-Multiplication

If the multiplication of two matrices could be performed exactly then whether one computed A^1 by left multiplication, right multiplication or by squaring $A^{1/2}$, if i is even, the results would be the same since matrices are associative with respect to multiplication if exact multiplication is performed. For example, if A^4 is computed by multiplying A times A^3, A^3 times A or by squaring A^2 the results are the same. Since, however, the powers of the matrices have their coefficients rounded-off, it is necessary to study the possible ways of computing A^i. In this thesis only the possibilities of left and right multiplication are considered.

DEFINITION 1. The computed i^{th} power of the matrix A obtained by left multiplication, $\overline{A}_L^i$, is computed by the formula

$$\overline{A}_L^i = A \times (A\times(A\times(A\times\cdots(A\times(A\times A))\cdots)))$$

the right member containing i factors.

DEFINITION 2. The computed i^{th} power of the matrix A obtained by right multiplication, $\overline{A}_R^i$, is computed by the formula

$$\overline{A}_R^i = (\cdots(((A\times A)\times A)\times A)\times\cdots\times A)\times A$$

the right member containing i factors.

DEFINITION 3. The symbol $((a,b,c))$, called pseudo-associator, is equal to $a\times(b\times c) - (a\times b)\times c$.

DEFINITION 4. The symbol $[[a,b]]$,called pseudo-commutator, is equal to $a\times b - b\times a$.

DEFINITION 5. The symbol (a,b), called commutator, is equal to $ab - ba$.

The commutator is zero in a commutative algebra.

Since $A*A$ is obtained by exactly squaring and then rounding-off the result,

$$A*A = A^2 + E_1$$

where E_1 is the error matrix due to rounding-off the exact product. In this thesis, E_i represents the error matrix due to round-off occurring when left multiplication is used, and F_i refers to the error matrix obtained when right multiplication is used. Using this notation, one obtains

$$A*(A*A) = A(A*A) + E_2 = A^3 + AE_1 + E_2$$

and

$$(A*A)*A = (A*A)A + F_2 = A^3 + E_1A + F_2.$$

In order to study some of the properties of left and right pseudo-multiplication, some of the properties of pseudo-associators and pseudo-commutators are developed. The pseudo-associator ((a,b,c)) is studied first. It is assumed that a,b, and c are from the field of real numbers. If all possible permutations of the letters are taken and these are added together, the result is zero.

That is,

$$((a,b,c)) + ((a,c,b)) + ((b,a,c)) + ((b,c,a)) + ((c,a,b)) + ((c,b,a)) = 0.$$

If $c = a$,

$$((a,b,c)) = a*(b*a) - (a*b)*a = 0.$$

Also

$$((a,a,b)) = a\times(a \times b) - (a \times a)\times b$$

which is generally not zero. Neither is $((b,a,a))$ generally equal to zero but it is of interest to note that

$$((a,a,b)) = -((b,a,a)).$$

If $b = c = a$, then the associator becomes $((a,a,a))$ which is obviously zero. Since $((a,a,a)) = \overline{a}_L^3 - \overline{a}_R^3$ and this is zero, one can show by mathematical induction that $\overline{a}_L^k - \overline{a}_R^k = 0$, for all positive integral values of k.

Matrices are considered now. It is shown here how $\overline{A}_L^i - \overline{A}_R^i$ can be written in terms of the associators. For instance

$$\overline{A}_L^4 - \overline{A}_R^4 = ((A,A,\overline{A}^2)) + ((\overline{A}^2,A,A)),$$

and

$$\overline{A}_L^5 - \overline{A}_R^5 = ((A,A,\overline{A}_L^3)) + ((\overline{A}_R^3,A,A)) + ((\overline{A}^2,A,\overline{A}^2)).$$

In general, if i is odd

$$\overline{A}_L^i - \overline{A}_R^i = ((A,A,\overline{A}_L^{i-2})) + ((\overline{A}_R^{i-2},A,A)) + ((\overline{A}^2,A,\overline{A}_L^{i-3}))$$

$$+ ((\overline{A}_R^{i-3},A,\overline{A}^2)) + ((\overline{A}_R^3,A,\overline{A}_L^{i-4})) + ((\overline{A}_R^{i-4},A,\overline{A}^3))$$

$$+ \cdot \cdot \cdot + ((\overline{A}_R^{(i-1)/2},A,\overline{A}_L^{(i-1)/2})) .$$

If i is even

$$\overline{A}_L^i - \overline{A}_R^i = ((A, A, \overline{A}_L^{i-2})) + ((\overline{A}_R^{i-2}, A, A)) + ((\overline{A}^2, A, \overline{A}_L^{i-3}))$$

$$+ ((\overline{A}_R^{i-3}, A, \overline{A}^2)) + ((\overline{A}_R^3, A, \overline{A}_L^{i-4})) + ((\overline{A}_R^{i-4}, A, \overline{A}_L^3))$$

$$+ \cdot \cdot \cdot + ((\overline{A}_R^{i/2-1}, A, \overline{A}_L^{i/2})) + ((\overline{A}_R^{i/2}, A, \overline{A}_L^{i/2-1})) .$$

Thus one sees that the difference between left and right multiplication can be expressed as the sum of appropriate associators. However, since this does not give a measure of the size of the difference between the two methods of calculating powers of A, a different approach is tried to ascertain if there is any preference of one method over the other.

Since $\overline{A}_R^2 = \overline{A}_L^2$ there is no preference of left multiplication over right or vice-versa in determining the pseudo-square of the matrix A. Now higher computed powers of the matrix A are considered. One can write

$$A \times \overline{A}^2 - A^3 = A \times \overline{A}^2 - A(\overline{A}^2 - E_i)$$

where all E_i have coefficients which are numerically less than $\beta^{-s}/2$. Using the maximum coefficient norm as a measure of the magnitude of $\overline{A}_L^3 - A^3$, the following result is obtained:

$$M(\overline{A}_L^3 - A^3) \leq \beta^{-s}/2 + n\beta^{-s}/2 = \beta^{-s}/2 \quad (1 + n)$$

where $M(\overline{A}_L^3 - A^3)$ means the maximum coefficient norm. If right multiplication is used, the result

$$\overline{A}^s \times A - A^s = \overline{A}^s \times A - (\overline{A}^s - F_s)A$$

is obtained. All F_i have coefficients which are numerically less than $\beta^{-s}/2$. Taking the maximum coefficient norm, one obtains

$$M(\overline{A}^s_R - A^s) \leq \beta^{-s}/2 + n\,\beta^{-s}/2 = \beta^{-s}/2\,(1 + n)\ .$$

Thus it is seen that there is no preference for one method over another if the norm M(A) is used as the measure of error. This does not mean that the results are identical in both cases, but that, in general, the norm M(A) cannot measure such a difference. To illustrate this, the following examples are cited:

Example 1.

$$A = \begin{pmatrix} .3 & .7 & .4 \\ -.2 & .4 & .8 \\ .9 & .8 & .1 \end{pmatrix}$$

$$A^s = \begin{pmatrix} .31 & .81 & .72 \\ .58 & .66 & .32 \\ .20 & 1.03 & 1.01 \end{pmatrix}$$

$$\overline{A}^s = \begin{pmatrix} .3 & .8 & .7 \\ .6 & .7 & .3 \\ .2 & 1.0 & 1.0 \end{pmatrix}$$

$$A^s = \begin{pmatrix} .579 & 1.117 & .844 \\ .330 & .926 & .792 \\ .763 & 1.360 & 1.005 \end{pmatrix}$$

$$\overline{A}_L^s = \begin{pmatrix} .6 & 1.1 & .8 \\ .3 & .9 & .8 \\ .8 & 1.4 & 1.0 \end{pmatrix}$$

$$\overline{A}_R^s = \begin{pmatrix} .6 & 1.1 & .8 \\ .3 & .9 & .8 \\ .8 & 1.3 & 1.0 \end{pmatrix}$$

$$\overline{A}_L^s - A^s = \begin{pmatrix} .021 & -.017 & -.044 \\ -.030 & -.026 & +.008 \\ .037 & +.040 & -.005 \end{pmatrix}$$

$$\overline{A}_R^s - A = \begin{pmatrix} .021 & -.017 & -.044 \\ -.030 & -.026 & +.008 \\ .037 & -.060 & -.005 \end{pmatrix}$$

In this example,

$$M(\overline{A}_L^s - A^s) = .044$$

and

$$M(\overline{A}_R^s - A^s) = .060 .$$

This illustrates the case where left multiplication is preferred to right multiplication if the norm M(A) is the measure used.

Example 2.

$$A = \begin{pmatrix} .8 & 0 & .9 \\ .4 & 1.0 & -.7 \\ .3 & .8 & .6 \end{pmatrix}$$

$$A^2 = \begin{pmatrix} .91 & .72 & 1.26 \\ .51 & .44 & -.76 \\ .74 & 1.28 & .07 \end{pmatrix}$$

$$\overline{A}^2 = \begin{pmatrix} .9 & .7 & 1.3 \\ .5 & .4 & -.8 \\ .7 & 1.3 & .1 \end{pmatrix}$$

$$A^3 = \begin{pmatrix} 1.394 & 1.728 & 1.071 \\ .356 & -.168 & -.305 \\ 1.125 & 1.336 & -.188 \end{pmatrix}$$

$$\overline{A}_L^3 = \begin{pmatrix} 1.4 & 1.7 & 1.1 \\ .4 & -.2 & -.4 \\ 1.1 & 1.3 & -.2 \end{pmatrix}$$

$$\overline{A}_R^3 = \begin{pmatrix} 1.4 & 1.7 & 1.1 \\ .3 & -.2 & -.3 \\ 1.1 & 1.4 & -.2 \end{pmatrix}$$

In this example,

$$M(\overline{A}_L^3 - A^3) = .095$$

and

$$M(\overline{A}_R^3 - A^3) = .064,$$

so, for this particular case, right multiplication is preferred over left multiplication if the maximum coefficient norm is used as a measure of the error. In general, it can be shown that

$$M(\overline{A}_L^k - A^k) = \beta^{-s}/2 \left[1 + n + n^2 + \cdots + n^{k-2}\right].$$

Also it can be shown that

$$M(\overline{A}_R^k - A^k) = \beta^{-s}/2 \left[1 + n + n^2 + \cdots + n^{k-2}\right].$$

Thus, if the maximum coefficient norm is used as a measure of the size of the error matrices, one cannot formulate a preference for either right or left multiplication. In any particular case, however, one method of multiplication might be preferred over the other method. This was illustrated above for the three by three case. Since one cannot express a preference for either method, left multiplication is used for most cases that follows.

V. SYMMETRIC PROPERTIES

Since there seems to be no particular preference for the order of multiplication, the number of operations required to obtain the inverse of a matrix is discussed now.

Table 1

Number of Operations Necessary to Compute A^{-1} by the Bingham Method

Operation	Types of Calculations		
	Additions	Multiplications	Divisions
Calculate A^k $k=1,2,3,\cdots,n-1$	$n^2(n-1)^2$	$n^3(n-2)$	
Calculate A^n	$n(n-1)$	n^2	
Calculate s_i $i=1,2,3,\cdots,n$	$n(n-1)$		
Calculate a_i $i=1,2,\cdots,n$	$(n/2)(n-1)$	$(n/2)(n-1)$	$n-1$
Calculate A^{-1}	$n(n-1)$ $+(n^2-n)(n-2)$	$n^2(n-2)$	n^2
Total	$n^4-n^3+\frac{3n^2}{2}-\frac{3n}{2}$	$n^4-n^3-\frac{n^2}{2}\quad\frac{n}{2}$	n^2+n-1

Table 2

Number of Operations Necessary to Compute A^{-1} by the Modified Bingham Method.

Operations	Types of Calculations		
	Additions	Multiplications	Divisions
Calculate A^k $k=1,2,\cdots,n-1$	$n^2(n-1)^2$	$n^3(n-2)$	
Calculate A^n	$n(n-1)$	n^2	
Calculate s_i $i=1,2,\cdots,n$	$n(n-1)$		
Calculate a_i $i=1,2,\cdots,n$	$n(n-1)/2$	$n(n-1)/2$	$n-1$
Calculate B_i $i=1,2,\cdots,n-2$	$(n-3)\left[n+n^2(n-1\right]+n$	$(n-3)n^3$	
Calculate A^{-1}	$n+n^2(n-1)$	n^3	n^2
Total	$2n^4-5n^3+\frac{13}{2}n^2-\frac{7}{2}n$	$2n^4-4n^3+\frac{3n^2}{2}-\frac{n}{2}$	n^2+n-1

Since the number of operations in determining the inverse of a matrix is more using the Bingham or modified Bingham Method than that for most of the other elimination methods, techniques to eliminate some of these operations for certain classes of matrices will now be investigated. Tables 1 and 2 respectively give the numbers of operations that are required to compute A^{-1} by the Bingham and modified Bingham Methods. The numbers given in these two tables do not take into consideration the types of matrices that one might be inverting. For example, if one is required to invert a triangular matrix, the number of operations required is considerably reduced over that given in the tables as is also the case for symmetric and positive definite matrices. The case for symmetric matrices is considered at this time.

A few theorems concerning symmetric matrices are given.

THEOREM 3. If A is a symmetric matrix, then A^k, for all k, is symmetric if A^k denotes the exact k^{th} power of the matrix A. This of course is well-known.

Proof. If $\overset{k}{a}_{ij}$ is the coefficient of A^k in the i^{th} row and j^{th} column, then

$$\overset{k}{a}_{ij} = \sum \cdots \sum a_{i\sigma_1} a_{\sigma_1\sigma_2} a_{\sigma_2\sigma_3} \cdots a_{\sigma_k j}$$

which is equal to

$$\sum \cdots \sum a_{j\sigma_k} a_{\sigma_k\sigma_{k-1}} \cdots a_{\sigma_1 i}$$

since A is symmetric and its coefficients satisfy the commutative law of multiplication. The last expression however, is equal to $\overset{k}{a}_{ji}$ by definition. This completes the proof that A^k is symmetric for all k. In the actual calculation of A^k, however, it is generally not practical to compute the coefficients exactly.

The question arises, then, whether $\bar{A}^k$ is a symmetric matrix for all values of k. In general, $\bar{A}^k$ is not symmetric. The following theorems are noted, however.

THEOREM 4. $A \times A$ is a symmetric matrix if A is a symmetric matrix.

Proof. Let $A \times A = (\bar{a}^*_{ij})$. Then

$\bar{a}^*_{ij} = \sum_{k=1}^{n} a_{ik} \times a_{kj} = \sum_{k=1}^{n} a_{jk} \times a_{ki} = \bar{a}^*_{ji}$. The second term is equal to the third term due to the commutivity of pseudo-products of numbers and the symmetric properties of the matrix A. The last equality is true by definition.

THEOREM 5. $A \times B$ is a symmetric matrix if AB is a symmetric matrix and if double precision multiplication is used to determine $A \times B$.

Proof. Since AB is a symmetric matrix,

$$\sum_{k=1}^{n} a_{ik}b_{kj} = \sum_{k=1}^{n} a_{jk}b_{ki} .$$

Now, remembering that $\sum_{k=1}^{n} a_{ik} \times b_{kj}$ is the same, before round-off is performed, as $\sum_{k=1}^{n} a_{ik}\ b_{kj}$ since double precision multiplication is used and $\sum_{k=1}^{n} a_{jk} \times b_{ki}$ is the same before round-off as $\sum_{k=1}^{n} a_{jk}b_{ki}$, one readily sees that $\sum_{k=1}^{n} a_{ik} \times b_{kj}$ and $\sum_{k=1}^{n} a_{jk} \times b_{ki}$ have been rounded-off to the same value. This proves the theorem.

Although it can be stated that generally $\overline{A}^k$ is not symmetric even though A is symmetric, it would be advantageous if the symmetric properties of A^k were also true in $\overline{A}^k$. In the first place, the amount of computation would be greatly reduced if only the upper triangular matrix coefficients and the diagonal coefficients had to be computed and the lower triangular coefficients obtained by a flip-over technique. If such a technique is used, does it affect the accuracy of A^{-1}? As a partial answer to this question, the errors due to the flip-over technique in obtaining the B_i are studied.

Equation (3.4) states that

$$B_i = A^{n-i-1} + a_1 A^{n-i-2} + \cdots + a_{n-i-1}\, I, \quad i=1,2,\cdots,n-1$$

and equation (3.5) is the recurrence formula

$$B_{i-1} = AB_i + a_{n-i}\, I.$$

DEFINITION 6. The computed B_i, written $\overline{B}_i$, are determined by means of the recurrence formula

$$\overline{B}_{i-1} = A \times \overline{B}_i + \overline{a}_{n-1}\ \ I$$

where

$$B_{n-1} = \overline{B}_{n-1} = I.$$

DEFINITION 7. If the B_i are computed by means of the flip-over technique, the resulting matrix is referred to as $\overline{B}_i^* = (\overline{b}^{i\,*}_{i'j})$ where

$$\overline{b}^{i\,*}_{i'j} = \begin{cases} \sum_{k=1}^{n} a_{i'k} \times \overline{b}^{i+1\,*}_{kj} + a_{n-i-1}\delta_{i'j} & i' \le j. \\ \overline{b}^{i\,*}_{ji'} & i > j . \end{cases}$$

The recurrence formula

$$\overline{B}^*_{i-1} = A \times \overline{B}^*_i + \overline{a}_{n-1} I$$

is used to compute the $\overline{B}^*_i$. Using equation (3.5) and the definitions above, one obtains

$$B_{n-1} = \overline{B}^*_{n-1} = I,$$

$$B_{n-2} = \overline{B}^*_{n-2}$$

and in general

$$\overline{B}^*_{n-3} = A \times \overline{B}^*_{n-2} + \overline{a}_2 I .$$

The $\overline{a}_i$ is assumed to be equal to the exact a_i. This means that

$$\overline{B}^*_{n-3} - B_{n-3} = A \times \overline{B}^*_{n-2} - AB_{n-2} = S_2^*$$

where S_2^* is a matrix whose elements are numerically less than the round-off error $\beta^{-s/2}$. In general

$$S_i^* = A \times \overline{B}^*_{n-i} - A\overline{B}^*_{n-i} .$$

Furthermore, if the $\bar{a}_i$ are assumed to be exact, then

$$\bar{B}_i^* - B_i = A \times \bar{B}_{i+1}^* - AB_{i+1}$$

$$= A \times \bar{B}_{i+1}^* - A \Big[\bar{B}_{i+1}^* - S_{n-i-2}^* - AS_{n-i-3}^*$$

$$- \cdots - A^{n-i-4} S_2^* \Big]$$

$$= S_{n-i-1}^* + AS^*_{n-i-2} + \cdots + A^{n-i-3} S_2^* .$$

If i = 1, this reduces to

$$\bar{B}_1^* - B_1 = S_{n-2}^* + AS_{n-3}^* + \cdots + A^{n-4} S_2^* .$$

Since

$$\bar{A}^{-1^*} = (-1/\bar{a}_n)(A \times \bar{B}_1^* + a_{n-1} \quad I) ,$$

$$\bar{A}^{-1^*} - A^{-1} = (-1/\bar{a}_n)(A \times \bar{B}_1^* - AB_1)$$

$$= (-1/\bar{a}_n)(S_{n-1}^* + AS_{n-2}^* + \cdots + A^{n-3} S_2^*) .$$

Using the same technique, it can be shown that

$$\bar{A}^{-1} - A^{-1} = (-1/\bar{a}_n)(S_{n-1} + AS_{n-2} + \cdots + A^{n-3} S_2)$$

where the S_i have coefficients whose numerical values are less than or equal to $\beta^{-s}/2$. In both cases

$$M(\bar{A}^{-1^*} - A^{-1}) \le \frac{n\beta^{-s}}{2|\bar{a}_n|} \Big[1+M(A)+M(A^2) + \cdots + M(A^{n-3})\Big]$$

and

$$M(\bar{A}^{-1} - A^{-1}) \le \frac{n\beta^{-s}}{2|\bar{a}_n|} \Big[1+M(A)+M(A^2) + \cdots + M(A^{n-3})\Big] .$$

Therefore, although the error matrices $(\bar{A}^{-1^*} - A^{-1})$ and $(\bar{A}^{-1} - A^{-1})$ may not have identical entries and the build up of such entries may be entirely different; nevertheless, if one uses the maximum coefficient norm as a measure of the error matrices then no difference can be detected in the two techniques. This does not mean that in all cases that one method is preferred over another but from the above considerations and also considering that the amount of multiplications required is almost cut in half, it seems desirable to use the flip-over technique to compute B_i. A similar argument holds for computing the A^k. Therefore, in computing the powers of A and the B_i the flip-over technique is used in case A is a symmetric matrix.

VI. SCALING OPERATIONS

A. Exact Considerations

Due to the limited capacity of computing machines it is necessary to retain only a fixed number of digits to make subsequent calculations a practical operation. For example, if the machine capacity is 6 digits and if 632,145 is multiplied times 732,148, the exact result is 462,823,697,460. This number is rounded-off to 462,824,000,000. Thus the magnitude of the error due to round-off is 302,540. If 6.32145 is multiplied times 7.32148, the magnitude of the error due to round-off is only .000030254. One sees then that the magnitude of the error is not a good measure of the accuracy. It is proposed, therefore, that one round-off in the s^th^ decimal place according to the method as prescribed in the introduction. If it were required that a number be rounded-off in the 6th decimal place and the machine capacity were 8 digits, then one could allow numbers whose magnitude did not exceed 99.999999. One should strive to use the full capacity of the machine and still have a consistent measure of the error due to round-off. If one rounds-off in say the 8th decimal place and numbers are written to the base 10, then the maximum magnitude of the error is $10^{-8}/2$. In general, if numbers are rounded-off in the s^{th} decimal place and the base of the number system is β, then the maximum error due to round-off is $\beta^{-s}/2$. It is important, therefore, that one restrict the number of digits to the left of the decimal point if numbers are rounded-off in a certain decimal position since the machine capacity is limited as to the number of digits it can manage. All numbers, therefore, are scaled so that their absolute value is less than one and the value chosen for s is governed by the number of digits the machine can handle.

In the case of matrices, it is not only important that the coefficients are scaled such that their absolute value is less than or equal to one; but also that the coefficients of certain powers of the matrix are numerically less than or equal to one in absolute value. This can be done by introducing appropriate scale factors as they are needed; but in this thesis scale factors will be introduced at the beginning so that all $\bar{A}^k$ $k = 1,2,\cdots n$ will have coefficients whose absolute values are less than or equal to one.

First, it is desired that all $\bar{a}_{ij}$ of the matrix A satisfy the inequality

$$|\bar{a}_{ij}| \leq 10^{-k} \qquad \text{all } i,j,$$

where the k is selected in such a way that for all relevant powers of A the coefficients are numerically less than or equal to one. One easily obtains the following inequalities:

$$|a^2_{ij}| \leq n\ (10^{-2k})\ , \qquad\qquad |a^3_{ij}| \leq n^2\ (10^{-3k})\ .$$

In general

$$|\bar{a}^h_{ij}| \leq n^{h-1} (10^{-hk}) .$$

Now let k be selected so that $(n^{h-1}) 10^{-hk}$ is numerically less than or equal to one for $h = 1,\cdots,n$. After some simplification, one obtains that k must satisfy the inequality

(6.1) $$k \geq (1 - 1/h) \log_{10} n.$$

Second, let k be selected in such a way that all coefficients of A^h for h greater than some prescribed m are zero after round-off. This is equivalent to saying that the absolute values of the coefficients are less than $\beta^{-s}/2$. If k is selected in this manner, its value is governed by the inequality

(6.2) $$k > (1 - 1/h)\log_{10} n + (s \log_{10}\beta)/h + \log_{10} 2/h$$

$$h > m.$$

Example. If one uses equation (6.1) with $n = 10$, then $k \geq 1$. This means that if a ten by ten matrix is considered then each of the coefficients must be numerically less than or equal to .1 in order that higher powers have coefficients less than or equal to unity in absolute value. This is very restrictive, however, and in the long run the k given by equation (6.1) could be selected as a smaller number and yet the coefficients of the powers of A would not exceed unity. This leads one to try a statistical approach to the problem.

Such a discussion is given in section B of this chapter.

B. Statistical Considerations

In the previous section on scaling operations, it was shown that if appropriate scale factors are introduced, one can be certain that pseudo-powers of the matrix A are digital matrices with coefficients numerically less than or equal to one. These scale factors however are too restrictive in general and it is possible to choose smaller scale factors for a large number of cases. The reason for desiring to select the optimal scale factors is in order that a maximum number of digits in each coefficient be retained. For example, if the digital number .333 is multiplied by the scale factor of $\frac{1}{10}$ the result is .033, by $\frac{1}{100}$ the result is .003, and by $\frac{1}{1000}$ the result is .000. Thus, as the scale factors change the number of non-zero retained digits may vary. Since one should retain as many digits as possible, it is exceedingly important that scale factors be selected appropriately.

In any particular case, the probability is of course, either zero or one that the selected scale factor is adequate to produce a matrix whose pseudo-powers are digital matrices. Some calculating machines are automatically shut off when the machine capacity is exceeded. A machine could be wired to shut off as soon as the entry is numerically greater than or equal to one. Thus a machine could be wired to shut off as soon as the coefficients of the pseudo-powers of the matrix exceed one in absolute value. If this occurs, subsequent calculations would have to be programmed and some of the previous calculations recomputed. This is costly since machine time is valuable. One is confronted, therefore, with two alternatives in what to do with scale factors.

First, one can select a scale factor according to the discussion in the preceding section and say with certainty that all pseudo-powers are digital matrices. An argument against using this scale factor is that valuable digits may be discarded. Second, if one selects the scale factor by probabilistic methods, one may select a smaller scale factor but then not be certain that pseudo-powers are digital matrices. If one makes certain assumptions about 1) the distribution from which the coefficients are taken and 2) the order of the matrix, then a statement can be made about the probability that the pseudo-powers are digital matrices. Since this probability is a measure of whether the machine will stop during the calculation process, it can be used to determine whether the scale factor is acceptable. For example, it may be that one is willing to accept a given scale factor if in the long run the machine will perform all of the calculations 95 per cent of the time without shut off being necessary. To decide between the two alternatives, it is necessary to study the characteristics of the sum of n products subject to certain restrictions on the variables as well as assumptions on the relevant distributions.

In order that a relevant probabilistic statement can be made about the sum of n products, some statistical properties are derived. It is assumed that each of the factors contained in the products is an independent random variable from a uniform distribution.

Briefly, if x_i and y_i are distributed uniformly from $-k$ to k, what is $P(|\sum_{i=1}^{n} x_i y_i| \leq 1)$? It should be noted that if k is less than $1/\sqrt{n}$, the probability is unity. If n is equal to one and k is greater than one, then the $P(|x_i y_i| \leq 1)$ is equal to $(1 + \ln k^2)/k^2$. For example, if k is equal to e, the $P(|x_i y_i| \leq 1)$ is equal to 0.406. If k is equal to e^2, the $P(|x_i y_i| \leq 1)$ is equal to 0.023.

Let $Y = \sum_{k=1}^{n} X_i$, where X_i is the product of two independent random variables taken from the uniform distribution lying in the interval $(-k,k)$. Then one is interested in finding the cumulative distribution of Y. The expected values of X^i, $E(X^i)$, are computed first. One has

$$(6.3) \qquad E(X^i) = \int_{-k}^{k}\int_{-k}^{k} x_1 x_2 \quad f(x_1,x_2)\ dx_1 dx_2 .$$

The distribution function, $f(x_1)$, is equal to $\frac{1}{2k}$ and the joint density function for x_1 and x_2, $f(x_1x_2)$, is equal to $(\frac{1}{2k})^2$. Equation (6.3) can be written as

$$E(X^i) = \left[\frac{1}{2k}\int_{-k}^{k} x^i \, dx \right]^2$$

$$= \left[\frac{k^i}{i+1}\right]^2 \qquad \text{if } i \text{ is even,}$$

$$= 0 \qquad \text{if } i \text{ is odd.}$$

Next, the $E(Y^i)$ are computed. Since the X_i are independent,

$$E(Y) = nE(X) = 0.$$

Also

$$E(Y^2) = nE(X^2) + n(n-1) \quad E(X)^2$$

$$= nk^4/9 \text{ .}$$

This value is important since it is the variance of Y. If i is an odd integer, $E(Y^i)$ is zero since all terms involve the expected values of odd powers of X and each of these are zero. In general,

$$(6.4) \quad E(Y^{2i}) = \frac{(2_i)!}{\prod_{m=1}^{k} (\alpha_m!)^{\beta_m}} \qquad \frac{n(n-1)\cdots(n-\sum\beta_m+1)}{\prod_{m=1}^{k} \quad \beta_m!} \prod_{m=1}^{k} \left[E(X^{\alpha_m}\right]^{\beta_m}$$

where $\sum_{m=1}^{k} \alpha_m\beta_m$ is equal to 2i and all possible selections of α_m and β_m are made. The value of k is determined by the number of different selections one makes for α_k and β_k .

Using equation (6.4) the first three non-zero values for $E(Y^i)$ are

$$E(Y^2) = nk^4/9,$$

$$E(Y^4) = nk^8/25 + 3n(n-1)\ k^8/81$$

and

$$E(Y^6) = nk^{12}/49 + n(n-1)k^{12}/15 + 5n(n-1)(n-2)k^{12}/243.$$

Now it is assumed that the cumulative density function $F(Y)$ can be represented by a Gram-Charlier series and a justification is given later. Suppose $f(Y)$ is the density function and suppose its mean and variance are μ and σ^2. If one lets $\mathring{Y} = (Y - \mu)/\sigma$, then $\mathring{Y}$ has zero mean and unit variance. The Gram-Charlier series is a series in the derivatives of the normal distribution of $\mathring{Y}$. If $n_i(\mathring{Y})$ represents the i^{th} derivative of the standard normal density $n(\mathring{Y},0,1)$ then in general

$$n_i(\mathring{Y}) = H_i(\mathring{Y})n_o(\mathring{Y}),$$

where $H_i(\mathring{Y})$ is the i^{th} Hermite polynomial and

$$n_o(\mathring{Y}) = \frac{1}{\sqrt{2\pi}}\ e^{-1/2\ \mathring{Y}^2} .$$

The first seven Hermite polynomials are:

$$H_o(y) = 1,$$

$$H_1(y) = -\ y,$$

$$H_2(y) = y^2 - 1,$$

$$H_3(y) = -\ y^3 + 3y,$$

$$H_4(y) = y^4 - 6y^2 + 3,$$

$$H_5(y) = -y^5 + 10y^3 - 15\,y,$$

$$H_6(y) = y^6 - 15y^4 + 45y^2 - 15\;.$$

The Gram-Charlier hypothesis (2) assumes that under rather general conditions f(Y) may be put in the form

$$(6.5)\quad f(Y) = n_o(\mathring{Y}) + \frac{c_3}{3!}\,n_o{}^{(3)}(\mathring{Y}) + \frac{c_4}{4!}\,n_o{}^{(4)}(\mathring{Y}) + \cdots$$

where

$$c_v = (-1)^v \int_{-\infty}^{\infty} H_v(\mathring{Y})\; f(\mathring{Y})\; d\,\mathring{Y}\;.$$

It can be shown that

$$c_o = 1,$$

$$c_1 = c_2 = 0,$$

$$c_3 = -\mu_3/\sigma^3,$$

$$c_4 = \frac{\mu_4}{\sigma^4} - 3,$$

$$c_5 = \frac{-\mu_5}{\sigma^5} + \frac{10\,\mu_3}{\sigma^3}$$

and

$$c_6 = \mu_6/\sigma^6 \;\; -15\;\mu_4/\sigma^4 + 30\;.$$

In addition, the cumulative distribution function F(Y) can be written

$$(6.6)\qquad F(Y) = \phi(\mathring{Y}) + \frac{c_3}{3!}\,\phi^{(3)}(\mathring{Y}) + \frac{c_4}{4!}\,\phi^{(4)}(\mathring{Y}) + \cdots$$

where

$$\phi(\mathring{Y}) = \int_{-\infty}^{\mathring{Y}} \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\; dx$$

Next, the c_i are computed and are expressed as follows in terms of the number of products, n:

$$c_0 = 1,$$

$$c_1 = c_2 = c_{2m+1} = 0 \qquad (m = 1,2,\cdots),$$

$$c_4 = \frac{6}{25n},$$

$$c_6 = -\frac{912}{245n^2} \quad .$$

Therefore, using equation (6.6), the Gram-Charlier series for the cumulative distribution is

$$F_{\overset{\circ}{Y}}(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \; e^{-t^2/2} \, dt + \frac{1}{100n}\int_{-\infty}^{u} (H_4) \frac{1}{\sqrt{2\pi}} \; e^{-t^2/2} \, dt$$

(6.7)

$$- \frac{19}{15(245n^2)} \int_{-\infty}^{u} \; (H_6) \frac{1}{\sqrt{2\pi}} \, e^{-t^2/2} \, dt + \cdots \; .$$

Since $\overset{\circ}{Y} = (Y - \mu)/\sigma$, $\mu = 0$ and $\sigma^2 = nk^4/9$, then $\overset{\circ}{Y} = 3Y/k^2\sqrt{n}$. Therefore,

$$P(\overset{\circ}{Y} \leq u) = P(\frac{3Y}{k^2\sqrt{n}} \leq u) = P(Y \leq \frac{uk^2\sqrt{n}}{3}) \; .$$

Now since one is to compute the probability that the numerical sum of the n products is less than or equal to one, the quantity $P(|Y| \leq 1)$ is to be computed. For example, if $k = 1$ and $n = 1$, it is certain that Y is numerically less than or equal to one. If one computes $P(|Y| \leq 1)$ by equation (6.7), the result is 0.996. This value compares favorably with the exact probability.

If $n = 100$, then in order to be assured that the square of a matrix have coefficients numerically less than one, each coefficient of the original matrix must be less than 0.1 in absolute value. If one insists that the coefficients be this small, then this is a very rigid restriction indeed. In general, the coefficients can be much larger and still the square of the matrix be a digital matrix. It is assumed for the moment that the three terms of the right member of equation (6.7) actually represent the cumulative distribution of Y. If this is true, then one finds that if the coefficients of the original matrix are numerically less than 0.4, then the probability that the square of the matrix be a digital matrix is 0.95. Thus it is seen that if the first three terms of equation (6.7) actually represent the cumulative distribution of Y, that the use of probabilistic considerations for determining scale factors is superior to using strict considerations.

Since the first three terms of the Gram-Charlier series are the same as the corresponding terms of the asymptotic Edgeworth series, the theory of the Edgeworth series is appealed to, since it is apparently of wider scope.

If one uses the same notation as Cramer [(2) p.59] , then for the cumulative distribution under consideration

$$
\begin{aligned}
V'(x) &= \frac{\ln k^2 - \ln x}{2k^2} & k^2 \geq x > 0, \\
&= \frac{\ln k^2 - \ln(-x)}{2k^2} & -k^2 \leq x < 0, \\
&= 0 & -k^2 \geq x \geq k^2 .
\end{aligned}
$$

Since the cumulative distribution satisfies the conditions of Theorem 2 given in Cramer [(2) p.59] then by Cramer (3) it can be written as an asymptotic expansion in powers of $n^{-1/2}$ with a remainder term of the same order of magnitude as the first term neglected. Remember that n is the order of the matrix. Now if the order of the matrix is allowed to increase without limit, then the difference between the distribution function and the partial sum of a fixed number of terms of the series can be made arbitrarily small. But no means of evaluating the remainder is available.

Now in practice, one wants to use a fixed number of terms in the series--usually not more than three or four--and also the order of the matrix is fixed. The two facts are not compatible with the asymptotic nature of the Edgeworth series and in general then the accuracy of the probabilistic statement is not known.

In defense, however, of the use of the first three terms of the series, the writer wishes to remind the reader that for any particular matrix, the probability is either one or zero that the scale factor accomplishes its purpose. This discussion is not to be used to ascertain the exact probability; but it is to be used only as a means of giving the programmer some idea as to what percent of the time the machines shut off automatically. Automatic shut off occurs when the capacity of the machine is exceeded.

It should be noted that the discussion in this section pertains to the square of a matrix. Higher powers are more difficult and the writer professes ignorance of the results for higher powers, unless rash (or rasher!) assumptions are made on the coefficients of powers of A.

VII. COMPARISON OF EXACT SOLUTIONS AND PSEUDO-SOLUTIONS OF A LINEAR EQUATION

A. Discussion of Some of the Problems

If one computes the inverse of a matrix, the question arises as to the accuracy of the result. In order to test the accuracy of the computed inverse, one can pseudo-multiply the original matrix times the computed inverse using ordinary or double precision multiplication. If this matrix pseudo-product is the identity matrix, then the computed inverse is usually considered to be satisfactory. If the pseudo-product is not the identity matrix, then the computed inverse may or may not be of sufficient accuracy.

By the very nature of the computational processes, the coefficients of the computed inverse are calculated to s-places. This means that unless the exact inverse is a digital matrix that one cannot obtain it by the usual machine calculations. Therefore, since it is generally impossible to obtain the exact inverse, the following questions arise:

1. Does the computed inverse satisfy the equation $\overline{A}^{-1} \times A = I$?
2. Does the exact inverse satisfy the equation $A^{-1} \times A = I$?
3. If $\overline{A}^{-1} \times A = I$, is $\overline{A}^{-1} = A^{-1}$?

4. If $\overline{A}^{-1} \times A = I$, are the rounded-off coefficients of A^{-1} equal to the coefficients of $\overline{A}^{-1}$?

5. If double precision multiplication is used, what is the effect upon pseudo-checking the computed inverse?

6. What is the effect of rounding-off the exact solution and then pseudo-checking this rounded-off solution?

The answers to these questions are not complete. In order to obtain a partial answer, the individual linear equations of the linear system

$$(7.1) \qquad \sum_{j=1}^{n} \overline{a}_{ij}\, x_j = b_i \qquad i = 1,2,\cdots,n$$

are analyzed. The $\overline{a}_{ij}$, and b_i (mod an integer) are digital numbers. Throughout this chapter, all $\overline{a}_{ij}$ are assumed to be different from zero.

B. Definitions and Some Properties of Pseudo-Solutions

For the first part of this discussion one of the n equations, say the i^{th} one, is considered.

DEFINITION 8. A set $\left\{ x_j \right\}$ which exactly satisfies

$$\sum_{j=1}^{n} \overline{a}_{ij} x_j = b_i \qquad \text{for a fixed } i$$

is said to be an exact solution. Any point on the hyperplane is an exact solution.

DEFINITION 9. Any set $\left\{x^P(j,k_j)\right\}$ is said to be a pseudo-solution if

$$(7.2) \qquad \sum_{j=1}^{n} a_{ij} \times x^P(j,k_j) = b_i,$$

where k_j is as yet undefined.

Throughout this chapter, ordinary precision multiplication is used unless specifically stated otherwise.

If one or more numbers of the set $\left\{x^P(j,k_j)\right\}$ also satisfy the condition

$$(7.3) \qquad a_{ij} \times x^P(j,k_j) = a_{ij} \;\; x^P(j,k_j) \qquad j=1,2,\cdots,n$$

then $x^P(j,k_j)$ is denoted by $x^{PC}(j,k_j)$.

It is obvious that there exist exact solutions to a linear equation. Any point on the hyperplane is an exact solution. It should also be noted that the exact solutions are not necessarily pseudo-solutions of the equations. For example,

$$x_1 + x_2 - x_3 = .1$$

has an exact solution (.07,.07,.04), but this point is not a pseudo-solution since $.1 + .1 - 0 \neq .1$. Thus, one sees that, although exact solutions exist, an exact solution may not be a pseudo-solution.

The next question that arises is whether pseudo-solutions exist for a linear equation. The answer to this question is in the affirmative. The following argument is sufficient to show the existence of pseudo-solutions. Pick n - 1 of the $x^P(j,k_j)$ arbitrarily. Next, rewrite equation (7.2) as follows:

$$\sum_{j=1}^{n-1} \bar{a}_{ij} \times x^P(j,k_j) + \bar{a}_{in} \times x^P(n,k_n) = b_i .$$

Since the summation (mod an integer) in the preceding equation is a digital number, this equation can be written

$$\bar{a}_{in} \times x^P(n,k_n) = d_i$$

where d_i (mod an integer) is a digital number. If this equation is solved for $x^P(n,k_n)$, one finds that $x^P(n,k_n)$ may lie anywhere in the open interval

$$\left(\frac{d_i}{\bar{a}_{in}} - \frac{\beta^{-s}}{2\bar{a}_{in}}, \frac{d_i}{\bar{a}_{in}} + \frac{\beta^{-s}}{2\bar{a}_{in}}\right).$$

Therefore, the existence of pseudo-solutions is established. Next, the properties or characteristics of pseudo-solutions and exact solutions are studied.

THEOREM 6. The sets $\left\{x^{PC}(j,k_j)\right\}$ are pseudo-solutions which lie on the hyperplane of exact solutions.

Proof. Since $\left\{x^{PC}(j,k_j)\right\}$ is a pseudo-solution,

$$\sum_{j=1}^{n} \bar{a}_{ij} \times x^{PC}(j,k_j) = b_i \;;$$

but by equation (7.3)

$$\sum_{j=1}^{n} \bar{a}_{ij} \times x^{PC}(j,k_j) = \sum_{j=1}^{n} \bar{a}_{ij}\, x^{PC}(j,k_j) = b_i \;.$$

The last equality shows that $\left\{x^{PC}(j,k_j)\right\}$ is an exact solution. This completes the proof.

THEOREM 7. The points $\left\{x^{PC}(j,k_j)\right\}$ are at the centers of n-dimensional rectangular parallelepipeds, hereafter referred to as n-topes, which contain all the pseudo-solutions.

Proof. If, and only if, the $x^{P}(j,k_j)$ are selected so that

$$x^{PC}(j,k_j) - \beta^{-s}/2\bar{a}_{ij} < x^{P}(j,k_j) < x^{PC}(j,k_j) + \beta^{-s}/2\bar{a}_{ij}$$

is equation (7.2) satisfied. Since only such values as those that lie in the interval indicated above satisfy equation (7.2) then this means that there exists an n-tope with center at $\left\{x^{PC}(j,k_j)\right\}$ which contains pseudo-solutions.

The $\left\{x^{PC}(j,k_j)\right\}$ form a lattice of points. If one considers a particular exact solution of the linear equation or equivalently if one considers a particular point on the hyperplane, there is at least one $\left\{x^{PC}(j,k_j)\right\}$ which is closest to the specified point. If there are several such points, then select one arbitrarily from those that are closest. This point is designated $\left\{ x^{PC}(j,0)\right\}$ and the n-tope with this point as its center is called the principal n-tope.

To recapitulate, $\left\{x_j\right\}$ is an exact solution, $\left\{x^P(j,k_j)\right\}$ is a pseudo-solution, $\left\{x^{PC}(j,k_j)\right\}$ is the center of an n-tope whose interior points are pseudo-solutions and $\left\{x^{PC}(j,0)\right\}$ is the center of the principal n-tope.

The n-topes are $\beta^{-s}/|\bar{a}_{ij}|$ units on a side. To distinguish between the centers of the n-topes and to show their relative positions the following formula is given:

$$x^{PC}(j,k_j) = x^{PC}(j,0) + k_j\beta^{-s}/|\bar{a}_{ij}| \tag{7.4}$$

$$k_j = 0, \pm 1, \pm 2, \cdots$$

$$j = 1,2,3,\cdots,n$$

$$i \text{ fixed.}$$

One should not infer from equation (7.4) that if the k_j are selected arbitrarily that this translates one from the center of the principal n-tope to the center of another n-tope. If

$$x^{PC}(j,k_j) = x^{PC}(j,0) + k_j\beta^{-s}/|\bar{a}_{ij}|$$

then

$$\sum_{j=1}^{n} \bar{a}_{ij} \times x^{PC}(j,k_j) = \sum_{j=1}^{n} \bar{a}_{ij} \times (x^{PC}(j,0) + k_j\beta^{-s}/|\bar{a}_{ij}|).$$

By Theorem 1, the distributive law may be applied to the right member of the equation since $\bar{a}_{ij} \times x^{PC}(j,0)$ has no round-off error. Therefore, one obtains

$$\sum_{j=1}^{n} \bar{a}_{ij} \times x^{PC}(j,k_j) = \sum_{j=1}^{n} \bar{a}_{ij} \times x^{PC}(j,0) + \sum_{j=1}^{n} \bar{a}_{ij} \times k_j \; \beta^{-s}/|\bar{a}_{ij}|.$$

Since the left member and the first term of the right member of this equation are both equal to b_i, this means that the k_j must be selected so that

$$(7.5) \qquad \sum_{j=1}^{n} \bar{a}_{ij} \times k_j/|\bar{a}_{ij}| = 0.$$

The symbols used in $x^{PC}(j,k_j)$ are clarified now. The P in the superscript denotes that a solution is a pseudo-solution. The C in the superscript denotes that the point is the center of one of the n-topes. The j refers to the j^{th} coordinate and the k_j's give the displacement from the center of the principal n-tope. Other properties of exact and pseudo-solutions are now considered.

THEOREM 8. An exact solution of a linear equation lies inside the principal n-tope if and only if

$$\sum_{j=1}^{n} \bar{a}_{ij} \times x_j = b_i \; .$$

Proof. If x_j lies inside the principal n-tope, then

$$\sum_{j=1}^{n} \bar{a}_{ij} \times x_j = b_i$$

by equation (7.2) and Theorem 7. Next if

$$\sum_{j=1}^{n} \bar{a}_{ij} \times x_j = b_i$$

and if one sets

$$x_j = x^{PC}(j,0) + v_j \, \beta^{-s}/2|\bar{a}_{ij}|$$

then one sees that this can be true only if

$$\sum_{j=1}^{n} \bar{a}_{ij} \times v_j \, \beta^{-s}/2|\bar{a}_{ij}| = 0.$$

This is true, however, only if $|v_j|$ is less than or equal to one. If this is true, however, the point lies inside the principal n-tope. This completes the proof.

It has already been shown by means of an example that there are points on the hyperplane of exact solutions that are not pseudo-solutions. In order to give an analytic description of the criterion for an exact solution to be a pseudo-solution the following analysis is given.

The general coordinate of any point on the hyperplane can be written

$$x_j = x^{PC}(j,0) + k_j \, \beta^{-s}/|\bar{a}_{ij}| + m_j \, \beta^{-s}/2|\bar{a}_{ij}|$$

where the k_j are integers which satisfy equation (7.5) and the m_j are any real numbers which satisfy the condition that

$$(7.6) \qquad \sum_{j=1}^{n} m_j \, \bar{a}_{ij}/|\bar{a}_{ij}| = 0.$$

Equations (7.5) and (7.6) assure one that the point $\{x_j\}$ is on the hyperplane.

If the m_j are all numerically less than one, the exact solution obviously lies inside one of the n-topes. If only one of the m_j is numerically greater than one, then the point is outside the n-tope and therefore is not a pseudo-solution. Other cases, where more than one m_j are numerically greater than one, may or may not produce exact solutions which are not pseudo-solutions.

C. Criterion for an Exact Solution to be a Pseudo-Solution

To get a better understanding of when an exact solution is a pseudo-solution, the following definition and theorem are given.

DEFINITION 10. The symbol $[x]_E$ means the closest even integer to x. If x is an odd integer, then the selection of the closest even integer is optional except in the following situation. If x is equal to an m_j which is an odd integer, then the closest even integer is selected, if possible, so as to satisfy the condition that the exact solution be a pseudo-solution.

This condition is given in the following theorem.

THEOREM 9. The exact solution of a linear equation is a pseudo-solution if and only if

$$(7.7) \qquad \sum \left[\text{positive } \frac{\bar{a}_{ij}\, m_j}{|\bar{a}_{ij}|}\right]_E = \sum \left[\left|\text{negative } \frac{\bar{a}_{ij}\, m_j}{|\bar{a}_{ij}|}\right|\right]_E \quad ,$$

where the m_j are the same as those in equation (7.6).

Proof. The summation $\sum_{j=1}^{n} \bar{a}_{ij}\, m_j/|\bar{a}_{ij}|$ is equal to zero since x_j is an exact solution. Let s equal the number of positive and r the number of negative numbers contained in the summation. The positive numbers are designated by $p_1, p_2, \cdots, p_s$ and the negative ones by $n_1, n_2, \cdots, n_r$. Next, set

$$[p_i]_E + q_i = p_i$$

and

$$[|n_j|]_E + t_j = -n_j \;.$$

This means that the q_i and t_j are numerically less than one. Now since

$$\sum_{i=1}^{s} p_i + \sum_{j=1}^{r} n_j = 0,$$

one obtains

$$\sum_{i=1}^{s} [p_i]_E + q_i - \sum_{j=1}^{r} [|n_j|]_E + t_j = 0.$$

But since it is given that

$$\sum_{i=1}^{s} [p_i]_E - \sum_{j=1}^{r} [|n_j|]_E = 0$$

and since q_i and t_j are numerically less than one, then this means that the exact solution is a pseudo-solution.

To prove the only if part, one observes that since the exact solution is a pseudo-solution that

$$x_j = x^{PC}(j,0) + k_j \beta^{-s}/|\bar{a}_{ij}| + m_j \beta^{-s}/2|\bar{a}_{ij}|$$

can be written

$$x_j = x^{PC}(j,0) + k_j \beta^{-s}/|\bar{a}_{ij}| + (2w_j + r_j)\beta^{-s}/2|\bar{a}_{ij}|$$

where the $2w_j$ are integers which satisfy equation (7.5) and the r_j are real numbers numerically less than one. The fact that the m_j can be written as $2w_j + r_j$ is equivalent to saying that

$$\sum \left[\text{positive } \bar{a}_{ij}m_j/|\bar{a}_{ij}|\right]_E = \sum \left[\left|\text{negative } \frac{\bar{a}_{ij}m_j}{|\bar{a}_{ij}|}\right|\right]_E .$$

This completes the proof.

To illustrate, the following examples are cited.

Example 1. If

$$m_1 a_{i1} /|a_{i1}| = 1.7,$$
$$m_2 a_{i2} /|a_{i2}| = 3.2,$$
$$m_3 a_{i3} /|a_{i3}| = 0.3$$

and

$$m_4 a_{i4} /|a_{i4}| = -5.2,$$

then

$$\sum_{j=1}^{4} m_j a_{ij} /|a_{ij}| = 0.$$

Since

$$[1.7]_E + [3.2]_E + [0.3]_E = [|-5.2|]_E$$

this illustrates the case where the exact solution is a pseudo-solution.

Example 2. If

$$m_1a_{i1} / |a_{i1}| = 1.6,$$
$$m_2a_{i2} / |a_{i2}| = 1.7,$$
$$m_3a_{i3} / |a_{i3}| = -2.4$$

and

$$m_4a_{i4}/|a_{i4}| = -0.9,$$

then

$$\sum_{j=1}^{4} m_ja_{ij} / |a_{ij}| = 0.$$

Since

$$[1.6]_E + [1.7]_E \neq [|-2.4|]_E + [|-0.9|]_E$$

this illustrates the case where the exact solution is not a pseudo-solution.

D. Discussion of the Two, Three and Four Dimensional Cases

Since the volumes and ratios of volumes of two and higher dimensional figures are discussed in Section E of this chapter, it is important that the effect upon the volumes and ratios of the volumes be determined when certain transformations are applied to the coefficients of the linear equation. The object is to try to simplify the coefficients of the linear equation and yet not change the ratios of the volumes. For example, it is much more convenient to refer to a linear equation which has all positive coefficients except one. Since any linear equation can be changed to this form by means of an orthogonal transformation which does not affect volumes, all subsequent linear equations discussed are of this form. Although this is a big help, further changes are desired.

To facilitate these changes, volumes and ratios of volumes are briefly discussed. As in Minkowski (4), for example, the volume of an n dimensional convex body is the n-fold integral $\int_R dx_1 \cdots dx_n$. The volume of a convex sub-region of this body would be obtained of course merely by integrating over a different region. It can be readily seen, therefore, that a substitution $x'_{ij} = \frac{x_{ij}}{|a_{ij}|}$ would not change the ratios of the volumes, but would replace each of the coefficients of the linear equation by plus or minus one. It is also readily seen that it is no less general to consider the hyperplane as passing through the origin. Therefore, for the remainder of this chapter, equations of the type

$$(7.8) \qquad x_1 + x_2 + x_3 + \cdots + x_{n-1} - x_n = 0$$

are analyzed. It is important to note that throughout this thesis, equation (7.8) is referred to as the n dimensional hyperplane or the n dimensional case.

It is also important to note that by translation, the entire hyperplane can be generated by a region near the origin. In the analysis that follows, therefore, only this portion of the hyperplane is considered. For example, if

$$x_1 - x_2 = 0$$

is the equation, then the line segment whose end points are (0,0) and (2,2), is the generating region. Using geometrical considerations one sees that all exact solutions are pseudo-solutions for this case.

Next, the equation

$$x_1 + x_2 - x_3 = 0$$

is considered. The generating region is the parallelogram whose vertices are the points (0,0,0),(2,0,2),(0,2,2) and (2,2,4). Since the m_j of equation (7.6) must also satisfy equation (7.8) and since it is easier to refer to the x_j of the equation, for the remainder of the chapter the analysis is made by letting the x_j vary. Before proceeding it seems advisable to make the following definition.

DEFINITION 11. The symbol I_M where I is an integer means all numbers in the closed interval $[I, I + 1]$.

That is, if $x_2 = 3_M$ then

$$3 \leq x_2 \leq 4 .$$

Now going back to the equation

$$x_1 + x_2 - x_3 = 0,$$

one sees that the generating region can be broken up into eight sub-regions which have properties as indicated in the following table:

Table 3

Analysis of the Generating Region of the Plane

	Range of Values			Pseudo-Solution	
Region	x_1	x_2	x_3	yes	no
1	0_M	0_M	0_M	x	
2	0_M	0_M	1_M		x
3	0_M	1_M	1_M	x	
4	0_M	1_M	2_M	x	
5	1_M	0_M	1_M	x	
6	1_M	0_M	2_M	x	
7	1_M	1_M	2_M		x
8	1_M	1_M	3_M	x	

Since each of these regions has the same area, this means that three-fourths of the area in the generating region contain points which are pseudo-solutions and one-fourth of the area contains points which are outside the 3-topes. Such points are not paeudo-solutions.

The hyperplane

$$x_1 + x_2 + x_3 - x_4 = 0$$

is discussed now. The convex body whose vertices are (0,0,0,0),(2,0,0,2),(0,2,0,2),(0,0,2,2),(2,2,0,4),(2,0,2,4), (0,2,2,4) and (2,2,2,6) is the generating region. In the last example, only those sub-regions of the generating region which were possible were listed but to better illustrate the situation, this time all forty-eight sub-regions are listed and the impossible ones classified accordingly.

Table 4

Analysis of the Generating Region of the Four Dimensional Hyperplane

	Range of Values				Pseudo-Solutions		
Region	x_1	x_2	x_3	x_4	Yes	No	Impossible
1	0_M	0_M	0_M	0_M	x		
2	"	"	"	1_M		x	
3	"	"	"	2_M		x	
4	"	"	"	3_M			x
5	"	"	"	4_M			x
6	"	"	"	5_M			x
7	"	"	1_M	0_M			x
8	"	"	"	1_M	x		
9	"	"	"	2_M	x		
10	"	"	"	3_M		x	
11	"	"	"	4_M			x
12	"	"	"	5_M			x
13	"	1_M	0_M	0_M			x

Table 4 (cont.)

	Range of Values				Pseudo-Solutions		
Region	x_1	x_2	x_3	x_4	Yes	No	Impossible
14	0_M	1_M	0_M	1_M	x		
15	"	"	"	2_M	x		
16	"	"	"	3_M		x	
17	"	"	"	4_M			x
18	"	"	"	5_M			x
19	"	"	1_M	0_M			x
20	"	"	"	1_M			x
21	"	"	"	2_M		x	
22	"	"	"	3_M	x		
23	"	"	"	4_M	x		
24	"	"	"	5_M			x
25	1_M	0_M	0_M	0_M			x
26	"	"	"	1_M	x		
27	"	"	"	2_M	x		
28	"	"	"	3_M		x	
29	"	"	"	4_M			x
30	"	"	"	5_M			x
31	"	"	1_M	0_M			x
32	"	"	"	1_M			x
33	"	"	"	2_M		x	
34	"	"	"	3_M	x		
35	"	"	"	4_M	x		

Table 4 (cont.)

	Range of Values				Pseudo-Solutions		
Region	x_1	x_2	x_3	x_4	Yes	No	Impossible
36	1_M	0_M	1_M	5_M			x
37	"	1_M	0_M	0_M			x
38	"	"	"	1_M			x
39	"	"	"	2_M		x	
40	"	"	"	3_M	x		
41	"	"	"	4_M	x		
42	"	"	"	5_M			x
43	"	"	1_M	0_M			x
44	"	"	"	1_M			x
45	"	"	"	2_M			x
46	"	"	"	3_M		x	
47	"	"	"	4_M		x	
48	"	"	"	5_M	x		

Regions 1,3,8,10,14,16,21,23,26,28,33,35,39,41,46 and 48 each have a volume of 1/6. Regions 2,9,15,22,27,34,40 and 47 each have a volume of 2/3.

The writer tabulated the results for the five and six dimensional cases but since it would require 448 lines of type to display them, they are not given here. In general, if one tabulates the k dimensional case, it takes $(k-1)2^k$ lines of type.

E. The Limit of the Ratio of Certain Volumes as the Dimension Increases Without Bound

This section is devoted to determining the distribution of the number of congruent sub-regions in the generating region, the classification of such sub-regions according to whether they contain pseudo-solutions or not and the determination of the ratio of the volume of the sub-regions which contain pseudo-solutions to the total volume of the generating region. The limit of this ratio is determined as the dimension of the hyperplane increases without limit.

THEOREM 10. If the equation

$$(7.9) \qquad x_1 + x_2 + \cdots + x_n - x_{n+1} = 0$$

is considered, the distribution of the number of congruent sub-regions in the generating region which contain pseudo-solutions is C_1^{n+1}, C_2^{n+1}, $\cdots$, C_g^{n+1}, $\cdots$, and C_n^{n+1} according to their position in the unit n-tope. That is, in the g^{th} position of the unit n-topes which make up the generating region, there are C_g^{n+1} congruent regions which contain pseudo-solutions.

Proof. If $x_{n+1} = 0_M$, there is only one way of selecting the $x_i (i = 1,\cdots,n)$ such that the true solution is a pseudo-solution. This can be done only by letting each $x_i = 0_M$ $(i=1,\cdots,n)$. If $x_{n+1} = 1_M$, then the exact solution is a pseudo-solution if and only if one of the $x_i=1_M$ $(i=1,\cdots,n)$ This can be done in C_1^n ways. If $x_{n+1} = 2_M$, then the exact solution is a pseudo-solution if only one of the $x_i = 1_M$ $(i=1,\cdots,n)$. This can be done in C_1^n ways. If $x_{k+1} = 3_M$, then the exact solution is a pseudo-solution if two of the $x_i = 1_M$ $(i=1,\cdots,n)$. This can be done in C_2^n ways. In general, if $x_{n+1} = I_M$, then there are $C^n{}_{[I+\varepsilon^2]_E/2}$ $(\varepsilon^2 < 1)$ ways of getting a pseudo-solution. This means that the total number of sub-regions of the generating n-tope in which the exact solution is a pseudo-solution is

$$C_0^n + C_1^n + C_1^n + \cdots + C^n{}_{[2n-3-\varepsilon]_E/2} + C^n{}_{[2n-1-\varepsilon]_E/2} \, ,$$

where $0 < \varepsilon < 1$.

Now one must show how these numbers are distributed according to their position in the n-tope. For example, one of the sub-regions which is under consideration at this time is the region between the n dimensional hyperplanes

$\sum_{i=1}^{n} x_i = 1$ and $\sum_{i=1}^{n} x_i = 2$ and inside the unit n-tope.

If an integral number of the x_i, say w of them, are picked equal to 1_M, then that part of the unit n-topes between the hyperplanes $\sum_{i=1}^{n} x_i = w$ and $\sum_{i=1}^{n} x_i = w + 1$ is referred to throughout the remainder of this thesis as the first position of the sub-regions. The second position of the sub-regions is that part of the unit n-topes which lies between the hyperplanes $\sum_{i=1}^{n} x_i = w + 1$ and $\sum_{i=1}^{n} x_i = w + 2$.

The g^{th} position is that part of the unit n-topes which lies between the hyperplanes $\sum_{i=1}^{n} x_i = w + g - 1$ and $\sum_{i=1}^{n} x_i = w + g$. This region contains pseudo-solutions only if

$$[x_{n+1}]_E = 2w.$$

But this can occur in $(C_{g-1}^{n} + C_{g}^{n})$ ways. Since $C_{g-1}^{n} + C_{g}^{n}$ is equal to C_{g}^{n+1}, this means that there are C_{g}^{n+1} g^{th} position regions which are in the unit n-topes that contain only pseudo-solutions. This completes the proof.

What portion of the generating region of the hyperplane given by equation (7.9) contains points which are pseudo-solutions? This question has already been answered for the two and three dimensional cases. To answer this question for higher dimensions it is first shown that some of the volumes of the sub-regions are congruent.

By translating any one of the unit n-topes under consideration to the origin it is noted that the coordinates of the vertices are either plus or minus one-half. Also, it is noted that the coordinates of the vertices of the sub-region in the g^{th} position are opposite in sign to the coordinates of the vertices of the sub-region in the $(n - g)^{th}$ position. That is, an orthogonal transformation $x_i = -x_i \quad i = 1,\cdots,n$ maps these regions into each other. Since the Jacobian of this transformation is plus or minus one, the volumes are preserved.

To distinguish between the volumes in the various positions, the symbol V_{ij} is introduced, and it refers to the volume in the i^{th} position of the unit (j - 1)-tope.

In the three dimensional case, there are four unit 2-topes which make up the generating region. Each of these unit 2-topes has its area subdivided into two areas. Of the four areas in the first position, three of them contain points which are pseudo-solutions. Of the four areas in the second position, three of them contain points which are pseudo-solutions. Therefore, the ratio of the area of the pseudo-solutions to the total area is

$$\frac{3V_{13} + 3V_{23}}{4V_{13} + 4V_{23}} = \frac{3}{4} .$$

In the four dimensional case, there are eight unit 3-topes which make up the generating region. Each of these unit 3-topes has its volume subdivided into three volumes. Of the eight volumes in the first position, four of them contain points which are pseudo-solutions. There are six volumes in the second position which contain only points which are pseudo-solutions. There are four volumes in the third position which contain only points which are pseudo-solutions. Therefore, the ratio of the volume containing pseudo-solutions to the total volume is

$$\frac{4V_{14} + 6V_{24} + 4V_{34}}{8V_{14} + 8V_{24} + 8V_{34}} .$$

Since

$$V_{14} = V_{34}$$

and

$$V_{14} + V_{24} + V_{34} = 1,$$

one obtains after substituting in the above ratio and simplifying that the ratio is

$$3/4(1 - 2/3\ V_{14})\ .$$

Since V_{14} is the volume of the corner of a unit cube, its value is one-sixth. Therefore, the ratio is

$$3/4(1 - 1/9) = 2/3\ .$$

This means that 2/3 of the generating region of the four dimensional hyperplane consists of points which are pseudo-solutions.

In the five dimensional case, there are sixteen unit, 4-topes which make up the generating region. Each of these unit 4-topes has its volume divided into four volumes. Of the sixteen volumes in the first position five of them contain points which are pseudo-solutions. There are ten volumes in the second position which contain only points which are pseudo-solutions. There are ten volumes in the third position and five in the fourth position which contain only points which are pseudo-solutions. Therefore, the ratio of the volume containing pseudo-solutions to the total volume of the generating region is

$$\frac{5V_{15} + 10V_{25} + 10V_{35} + 5V_{45}}{16V_{15} + 16V_{25} + 16V_{35} + 16V_{45}} \cdot$$

Now since

$$V_{15} = V_{45}, \qquad V_{25} = V_{35}$$

and

$$\sum_{i=1}^{4} V_{i5} = 1,$$

the ratio becomes

$$5/8\ (1 - V_{15}) \cdot$$

If the number of variables is 2m there are 2^{2m-1} unit (2m - 1)-topes which make up the generating region. Each of these unit (2m - 1)-topes has its volume subdivided into (2m - 1) volumes. Of the 2^{2m-1} volumes of the generating region which lie in the g-th position, C_g^{2m} of them contain points which are pseudo-solutions. Therefore, the ratio of the volume containing pseudo-solutions to the total volume is

$$(7.10)\quad \frac{C_1^{2m}\ V_{1,2m} + C_2^{2m}\ V_{2,2m} + \cdots + C_{2m-1}^{2m}\ V_{2m-1,2m}}{2^{2m-1}\left[V_{1,2m} + V_{2,2m} + \cdots + V_{2m-1,2m}\right]} \cdot$$

Since

$$V_{i,2m} = V_{2m-i,2m},$$

$$\sum_{i=1}^{2m-1} V_{i,2m} = 1$$

and

$$C_1^{2m} = C_{2m-1}^{2m} \quad ,$$

then this ratio can be written

$$(7.11) \quad \frac{1}{2^{2m-1}} \left[2C_1^{2m} V_{1,2m} + \cdots + 2C_{m-1}^{2m} V_{m-1,2m} + C_m^{2m} V_{m,2m} \right] \quad .$$

Now if one eliminates $V_{m,2m}$ from equation (7.11) the ratio becomes

(7.12)

$$\frac{C_m^{2m}}{2^{2m-1}} \left[1 + 2 \frac{(C_1^{2m} - C_m^{2m})}{C_m^{2m}} V_{1,2m} + \cdots + 2 \frac{(C_{m-1}^{2m} - C_m^{2m})}{C_m^{2m}} V_{m-1,2m} \right] .$$

But since $C_i^{2m} - C_m^{2m}$ is negative for $(i=1,2,\cdots,m-1)$ the ratio is always less than $C_m^{2m}/2^{2m-1}$. Although this ratio is computed for the even dimensional cases, it is important to note that the coefficient of the bracket in equation (7.12) is the same value for the 2m-1 dimensional case as it is for the 2m dimensional case. This is due to the fact that the middle terms of the binomial coefficients for n equal to 2m-1 are one-half of the middle term for n equal to 2m.

Now the ratio $C_m^{2m}/2^{2m-1}$ is studied as $m \longrightarrow \infty$. The limit can be evaluated by first writing

$$(7.13) \quad \lim_{m\to\infty} \frac{C_m^{2m}}{2^{2m-1}} = \lim_{m\to\infty} \frac{\dfrac{(2m)!}{\sqrt{2\pi}\, e^{-2m}\,(2m)^{2m+1/2}}}{\left[\dfrac{m!}{\sqrt{2\pi}\, e^{-m} m^{m+1/2}}\right]^2 \sqrt{\dfrac{m\pi}{2}}} ,$$

but this is zero since the asymptotic value of m! is $\sqrt{2\pi}\, e^{-m} m^{m+1/2}$ by Stirling's formula. This leads to the theorem.

THEOREM 11. The ratio of the volumes of the subregions which contain pseudo-solutions to the total volume of the generating region approaches zero as the dimension of the hyperplane increases without bound.

In addition it is of interest to note that the sequence $\left\{ C_m^{2m}/2^{2m-1} \right\}$ is a monotonic decreasing sequence. To show this, one observes that the ratio of two successive terms is equal to

$$\frac{(2m + 1)(2m + 2)}{(m+1)^2 (4)} = \frac{(m + 1/2)}{m+1}$$

which is always less than one since m is a positive integer.

VIII. SUMMARY

Regardless of whether the solution of a linear system of equations is obtained by elimination methods, iterative methods or by first solving for the inverse of the coefficient matrix and then multiplying this inverse times the column matrix of constants, the problem of pseudo-multiplication and pseudo-division always confronts one. This problem is the basis for the research in this thesis.

In Chapters I and II pseudo-operations for scalars are defined and some old and new properties of these pseudo-operations are derived.

Chapter III first gives a description of the Bingham Method for inverting a matrix. The essential difference between this method and well-known elimination methods is that the Bingham Method requires the computation of a finite number of powers of the matrix. To be more explicit the first n-1 powers and the diagonal elements of the n-th power of the matrix must be computed. Since memory storage space is a critical item in performing calculations with automatic calculators, the writer suggests a modification of the Bingham Method. This Modified Bingham Method requires only about 3/n as much storage space as that required by the usual Bingham Method.

Pseudo-operations for matrices are defined and some of their properties developed in Chapter IV. Left and right pseudo-multiplication of matrices, used to determine computed powers of the matrix, are expressed in terms of pseudo-associators.

In Chapter V the number of additions, multiplications and divisions required to compute the inverse of a matrix by the Bingham and Modified Bingham Methods is given. It is observed that if the matrix A is symmetric then the pseudo-product, $A \times A$, is symmetric. It is also shown that if AB is symmetric and if double precision multiplication is used, then $A \times B$ is symmetric. These facts help to reduce by approximately one-half the number of operations necessary to compute the inverse of a symmetric matrix.

If a matrix A is pseudo-multiplied times itself it is often desirable to obtain a matrix whose coefficients are numerically less than one. This can be done by dividing each coefficient of the matrix A by an appropriate scale factor. The question arises as to the value of such a scale factor. In Chapter VI criteria for obtaining appropriate scale factors are given using both strict and probabilistic considerations.

In Chapter VII a comparison is made of exact solutions and pseudo-solutions of a linear equation. Located on the hyperplane of exact solutions are a set of points

$\left\{x = \left\{x^{PC}(j,k_j)\right\}\right\}$ which form a lattice structure. It is shown that each $\left\{x^{PC}(j,k_j)\right\}$ is the center of an n dimensional rectangular parallelepiped, called an n-tope. Inside each of these n-topes are points which are pseudo-solutions of the linear equation. Whether the points on the hyperplane lie inside one of these n-topes is a function of the dimension of the hyperplane. It is proved that as the dimension of the hyperplane increases without bound that the ratio of the volume of the generating region which contains only pseudo-solutions to the total volume of the generating region approaches zero.

IX. SUGGESTIONS FOR FURTHER STUDY

Although this thesis has made considerable progress in the discussion of pseudo-solutions of a linear equation, it is realized by the writer that the questions proposed in Chapter VII were only partially answered. It is suggested that lattice structure and ideals be utilized in any further investigation. Minkowski (4) gives an excellent introduction to these topics. In addition, perhaps one could find classes of matrices for which definite answers to the questions posed in Chapter VII can be given.

X. BIBLIOGRAPHY

A. Literature Cited

1. Von Neumann, J. and Goldstine, H. H. Numerical inverting of matrices of high order. Bull. Amer. Math. Soc. 53:1022-1097. 1947.

2. Cramer, H. On the composition of elementary errors (First paper). Skandinavisk Aktuarietidskrift. 2:13-74. 1928.

3. -------- Mathematical methods of statistics. Princeton Universit y Press, 1951.

4. Minkowski, H. Diophantische approximationen. B. G. Teubner Publisher, Leipzig, 1927.

B. Other Bibliography

Bonnesen, T. and Fenchel, W. Theorie der konvexen Korper. Ergebnisse Der Mathematik und Ihrer Grenzgebiete. 3:1-172. 1934-35.

Ford, W. B. Studies on divergent series and summability. Michigan Science Series Vol. 2. The Macmillan Co. 1916.

Goldstine, H. H. and Von Neumann, J. Numerical inverting of matrices of high order II. Proc. Amer. Math. Soc. 2:188-200. 1951.

Hotelling, H. Some new methods in matrix calculations. Annals Math. Stat. 14:1-33. 1951.

Lonseth, A. T. Systems of linear equations with coefficients subject to error. Annals Math. Stat. 13:332-337. 1942.

Moulton, F. R. On the solution of equations having small determinants. Amer. Math. Monthly. 20:242-249. 1913.

Satterthwaite, F. E. Error control in matrix calculations. Annals Math. Stat. 15:373-387. 1944.

Turing, A. M. Rounding-off errors in matrix processes. Quar. Jour. Mech. Applied Math. 1:287-308. 1948.

Whittaker, E. T. and Watson, G. N. A course of modern analysis. The Macmillan Co. 1915.

Wundheiler, A. W. The necessity of error analysis in numerical computations. Annals Comp. Lab. Harvard Univ. 16:83-90. 1947

XI. ACKNOWLEDGMENT

The author wishes to acknowledge the guidance and many helpful suggestions of Dr. Bernard Vinograde in the preparation of this thesis.