GP-B ERROR MODELING AND ANALYSIS
(Annual Report)

Prepared for

National Aeronautics and Space Administration
George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

by

James C. Hung
The University of Tennessee
Electrical Engineering Department
Knoxville, Tennessee 37996

Under Contract NAS8-34426

30 September 1982

GP-B ERROR MODELING AND ANALYSIS
(Annual Report)

Prepared for

National Aeronautics and Space Administration
George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

by

James C. Hung
The University of Tennessee
Electrical Engineering Department
Knoxville, Tennessee 37996

Under Contract NAS8-34426

30 September 1982

ACKNOWLEDGEMENT

This annual report documents the progress made by the University of Tennessee for the contract NAS8-34426 during the period 24 June 1981 to 31 August 1982. The contract was devoted to GP-B error modeling and analysis, which was supported by the National Aeronautics and Space Administration (NASA) and monitored by the Marshall Space Flight Center (MSFC).

Dr. James C. Hung of the University of Tennessee at Knoxville was the Principal Investigator of the contract, who was assisted by three graduate students Messrs Ali Alouani, Jelel Zrida, and Mounir Laroussi. Mr. James M. McMillion of NASA/MSFC was the Contracting Officer's Technical Representative.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

In 1916, Albert Einstein published his famed thoery of general relativity. Although many contemporary physicists and astronomers based their work on the principles of this theory, proof of Einstein's description of the universe remains inconclusive. In fact, alternatives to Einstein's theory, most notably the Brans-Dicke scalar-tensor theory, have been developed.[1]

The Gravity Probe experiment, conceived by the late Professor Leonard Schiff of Stanford University, seeks to test Einstein's theory, and has been recognized by the U. S. Government as having far-reaching scientific importance. Since 1962, the government has supported the development of this experiment in which academic, industrial, and governmental institutions have participated. Currently, NASA's Marshall Space Flight Center is charged with the overall responsibility of the experiment.

In proposing this experiment, Professor Schiff had shown that a gyroscope in orbit undergoes a relativistic precession in the framework of the fixed stars.[2] In the "Gravity Probe B" (GP-B) satellite, four almost perfectly spherical fused-quartz gyros, each the size of a ping-pong ball, will be coated with niobium, and element made superconductive at temperatures near absolute zero. In this condition, the gyros can be suspended electrostatically in the weightlessness of space. It is expected that the drift rate of these gyros will be held to less than .001 arc-second a year. For gyroscopes in polar orbit, there are two predicted precessional effects which are perpendicular to one another, as shown in Figure I-1. The larger effect is due to the motion of the gyroscope through the Earth's gravitational

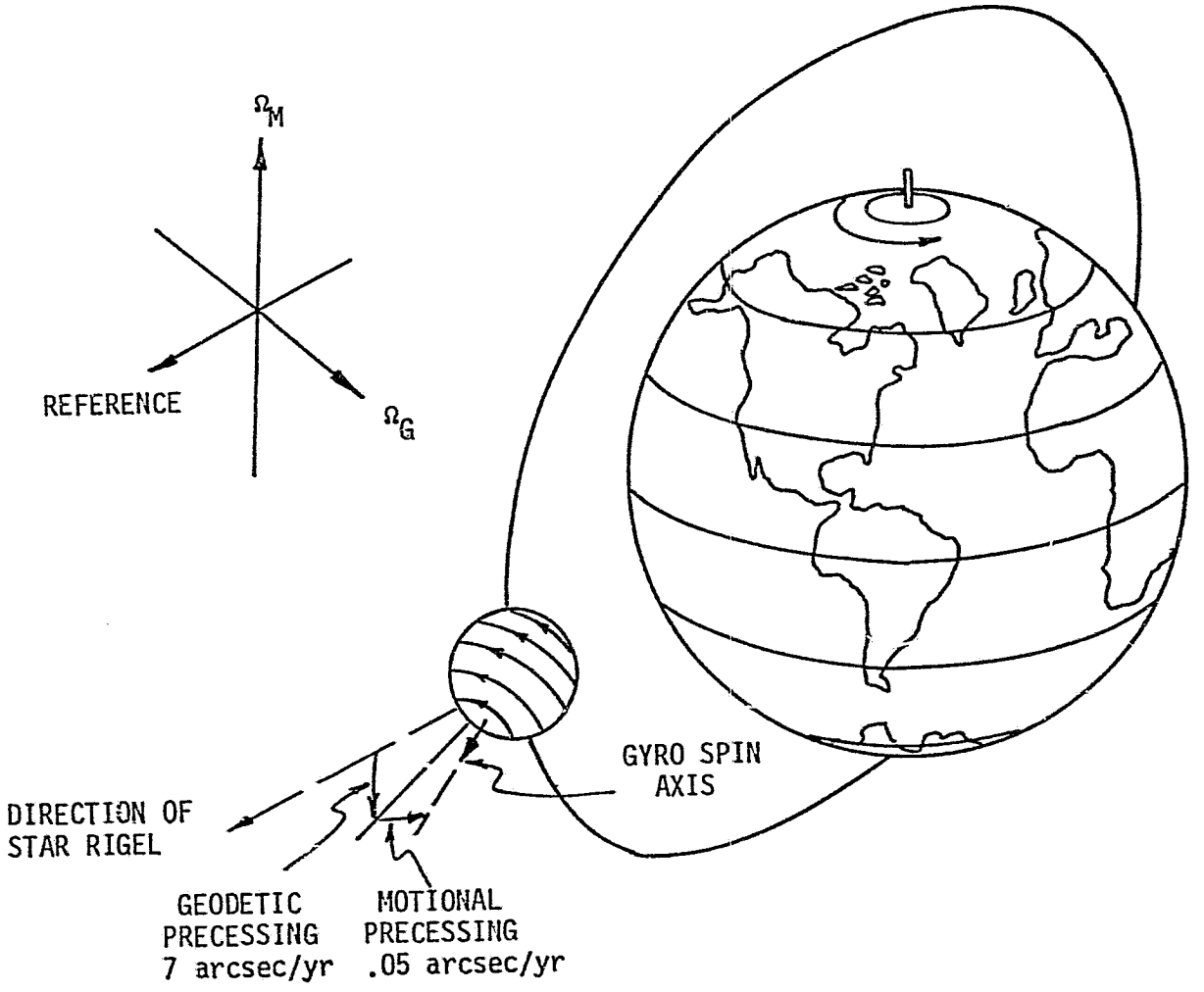


Figure I-1. Relativistic precessing effects of a gyro.

field and is referred to as geodetic precession. The other predicted effect is the spin-spin coupling between the gyroscope and the Earth's rotation and is referred to as the motional effect. As shown in the figure, the geodetic effect is in a plane of orbit approximately parallel to the Earth's axis. The motional effect lies in a plane perpendicular to both the orbit plane and the Earth's spin axis. The predicted magnitudes of procession vary from theory to theory. The purpose of this experiment is to determine the most valid theory by comparing the experimentally measured magnitudes with the various magnitudes obtained from different predictions.

The most important part of the GP-B satellite is a dewar containing the four spherical gyroscopes, a telescope which tracks the star Rigel, and an amount of supercooled helium. Spacecraft attitude and translational control is accomplished by proportional thrusters with the propellant supplied by the boiloff helium from the dewar. The control signals are derived from the telescope when the guide star is visible, and from the gyroscopes when the star is occulted

The experiment design goal is to make drift measurements accurate to .001 arcsecond per year, an unprecedented requirement for drift measurement. To ensure the success of the experiment, every conceivable error source in the GP-B system needs to be identified and its effect on drift measurement investigated.

In April 1980, NASA/MSFC awarded a study contract, NAS8-33849, to the University of Tennessee to initiate a system approach to GP-B error analysis. The study results included a block diagram error model, identification and discussion of source errors, some preliminary error analysis, documentation

of error budget, and several recommendations.[3] These results provided an organized picture of various errors involved.

The present contract, NAS8-34426, is aimed at an in-depth investigation of individual source errors and their effect on the accuracy of the GP-B experiment. This report documents the results of the study performed over a fourteen month period, from 24 June 1981 to 31 August 1982. The main emphases during this report period were placed in the following three areas:

1. Refinement on source error identification and classifications of source errors according to their physical nature. The result is given in Chapter II of this report.
2. Error analysis for the GP-B data processing. The result obtained is given in Chapter III.
3. Measurement geometry for the GP-B experiment, to be in Chapter VI.

It should be mentioned that this is a progress report, therefore the results reported can be further improved.

References

1. C. W. Misner, K. S. Thorne, and J. A. Wheeler, Gravitation, W. H. Freeman and Co., 1973.
2. L. I. Schiff, "Motion of a Gyroscope According to Einstein's Theory of Gravitation," Proc. N.A.S., Vol. 46, 1960, 871-882.
3. J. C. Hung, "Gravity Probe B Error Analysis," Final Report for NASA Contract NAS8-33849, The University of Tennessee, Knoxville 18 February 1981.

CHAPTER II

PROPERTY AND CLASSIFICATION OF GP-ERROR SOURCES

Figure II-1 is a block diagram error model for the GP-B experiment.[1] The model was developed in the previous contract based on the data contained in Reference 2. This diagram gives a "macro" description of errors where error sources are lumped into nine groups entering the system at nine entry points. During this contract report period, further effort was exerted on error source identification using the data available in Reference 3. In addition, these error sources were classified in two different ways based on their physical properties. The purpose of classification is to expose the possibility of suppressing the effects of individual errors and to suggest techniques for such suppression.

The errors can be classified according to there statistical nature as follows:

1. Deterministic type — This type of errors can be compensated.
2. Random variable type — This type of errors are uncertain constants. The effect of some, but not all, of these errors can be eliminated by rolling the spacecraft and by orbit averaging.
3. Random process type — This type of errors are uncertain time-varying quantities. Some of their effect can be reduced by averaging or by filtering.

The errors can also be classified according to their physical forms in the following way:

1. Bias — The effect of some of this type of errors may be reduced by roll averaging and orbit averaging.

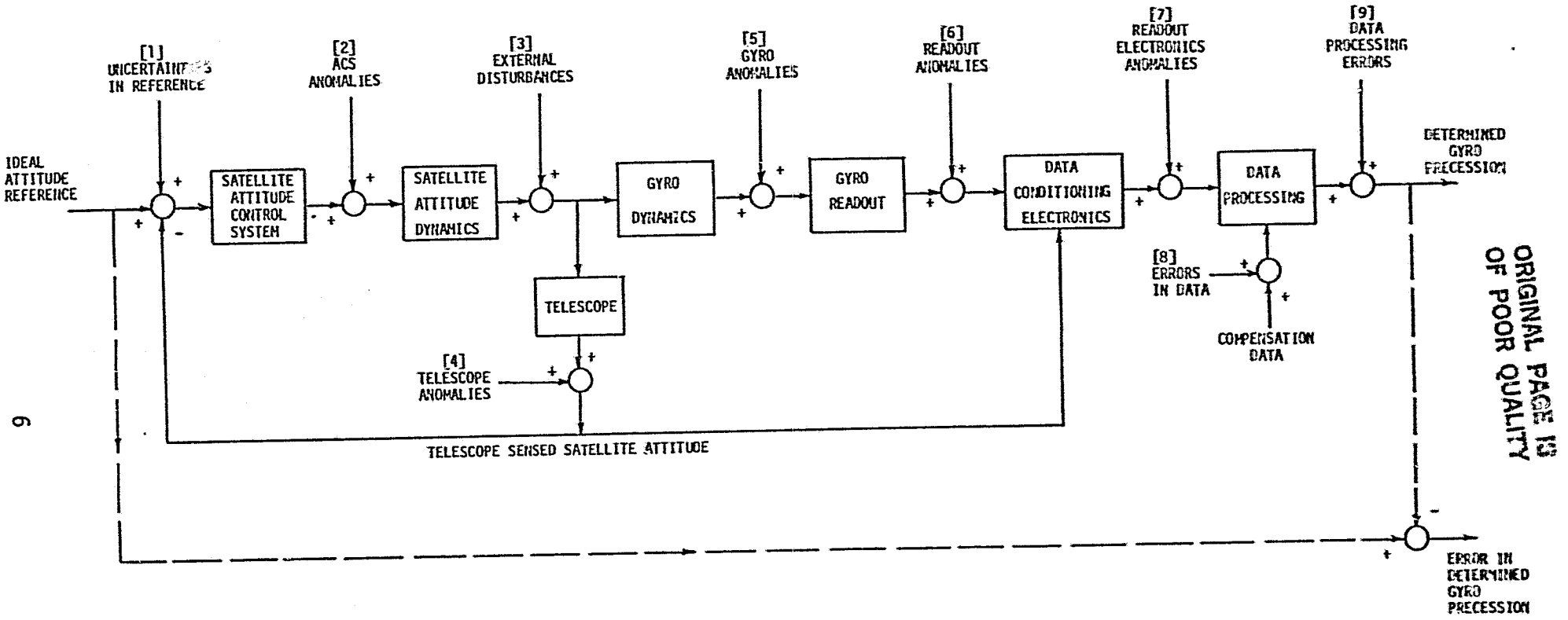


Figure II-1. A block diagram error model.

2. Drift.

3. Scale factor uncertainty — May be identified by Kalman filtering.

4. Noise — May be helped by Kalman filtering.

Table II-1 is a list of error sources, their classification, and methods for the reduction of error effect. In the table, r.v. means random variable, r.p. means random process, 1/f means 1/f noise, and det means deterministics. Notice the estimated errors after data processing, their root-sum-square (RSS) total is 1.253 milliarcseconds assuming zero proper motion for Rigel. It is expected that future progress in hardware and software development will reduce the RSS error. Notice also that the table is open-ended, with blanks, "TBD", and question marks to be taken care of by further study.

Figure II-2 shows the budget curve for the average individual error as a function of the number of equal size error sources. The curve is based on a RSS error of 1 millarcsecond. For example, if the total number of equal size error source is 44, the individual error source should not cause more than .15 milliarcsecond error for the GP-B experiment.

References

1. J. C. Hung, "Gravity Probe B Error Analysis," Final Report for NASA Contract NAS8-33849, University of Tennessee, Knoxville, 18 February 1981.

2. J. R. Parker, "Gravity Probe B Phase A Report," National Aeronautics and Space Administration, Marshall Space Flight Center, March 1980.

3. C. W. F. Everitt, "Report on a Program to Develop a Gyro Test of General Relativity in a Satellite and Associated Control Technology," Stanford University, Hansen Laboratory, June 1980.

Table II-1. List of GP-B error sources.

ERROR	NATURE	SOURCE OF INFORMATION	EFFECT REDUCTION BY ROLLING	EFFECT REDUCTION BY K.F.	EFFECT REDUCTION BY OTHER MEANS	ERROR BEFORE DATA PROC m^3/yr	ERROR AFTER DATA PROC m^3/yr	REMARKS
Gyro								
Mass unbalance	r.v.	Stanford's G.B.	✓				.3	
Suspension	r.v.	Eby's reports	✓				.1	
Gravity gradient	r.v.	Stanford's G.B.	✓				.05	
Magnetic torque	r.v.	NASA TR-R443, 1975					.01	
Residue charge	r.v.	Stanford's G.B.	✓				.15	
Residual gas	r.p.	" "	✓	✓			.1	
Brownian motion	r.p.	" "	✓	✓			.001	
Photon Bombardment	r.p.	" "	✓	✓			.0001	
Cosmic rays	r.p.	" "	✓	✓			.00001	
Mass attraction	r.v.						TBD	
Readout Sensor								
SQUID noise	1/f & r.p.	Stanford's G.B.	✓	✓			.1	
Centering error	r.v.	" "	✓				.01	
Nonlinearity	r.v.	" "	✓				.5	
Readout Electronics								
Null drifts	r.p.	Stanford's G.B.	✓	✓			.1	
Gain matching	r.v. & r.p.	" "		✓			.1	
Nonlinearity	r.v.	" " ; Sheingold					.5	
Roll decoding error	r.p.						TBD	
Digitization error	r.p.		✓	✓			TBD	
Systematic error	r.v.						.002	
Telescope								
Noise	r.p.	Stanford's G.B.	✓	✓			.015	
Chips, curve, nonorthog.	r.v.	" "					.1	
Null drift	r.p.	" "					.02	
Nonlinearity	r.v.	" "	✓				.1	
Distortion due to thermo.	r.p.	" "					TBD	

Table II-1. (Continuation)

ERROR	NATURE	SOURCE OF INFORMATION	EFFECT REDUCTION BY ROLLING	EFFECT REDUCTION BY K.F.	EFFECT REDUCTION BY OTHER MEANS	ERROR BEFORE DATA PROC. m3/yr	ERROR AFTER DATA PROC. m3/yr	REMARKS
Attitude Reference								
Aberrations	det.	Motz & Duveen; Sommerfeld			✓		.076	
Parallax	det.						.6	
Solar bending of LOS	det.						.16	
Variation in star intensity	r.p.						.7	
Proper motion of Rigel	r.p.	Anderson & Everitt			?		1.7	
Rigel's companion	?	Stanford's G.B.			?		TBD	
Data Mgmt & Transmission								
Round-off	r.p.						TBD	
Truncation	r.v.						TBD	
Bit error rate	r.v.						TBD	
A/D & D/A errors	r.v.						TBD	
Algorithm error	r.v.						TBD	
Compensation Data								
Orbit libration	det							
Proper motion	r.p.							
Orbit perturbation	r.p.						TBD	
Steller aberration	r.v.	Stanford's G.B.; Motz & Duveen.					TBD	
Solar bending of LOS	det						TBD	
Gravity anomaly	r.p.						TBD	
Data Processing								
Covariance analysis	r.p.						TBD	
Algorithm accuracy	r.v.						TBD	
Data Integrity	r.p.						TBD	
Truncation error	r.p.						TBD	
Round-off error	r.p.						TBD	
Alignment errors								
Telescope/Gyro	r.v.						TBD	
CG/Proof mass	r.v.						TBD	

Table II-1. (Continuation)

ERROR	NATURE	SOURCE OF INFORMATION	EFFECT REDUCTION BY ROLLING	EFFECT REDUCTION BY K.F.	EFFECT REDUCTION BY OTHER MEANS	ERROR BEFORE DATA PROC. m^2/gr	ERROR AFTER DATA PROC. m^2/gr	REMARKS
Spacecraft Control								
Solar Panel Disturbance	r.p.			✓			TBD	
Temperature	r.p.		✓				TBD	
Total Error							1.253	RSS, without proper motion
							2.112	RSS, with proper motion

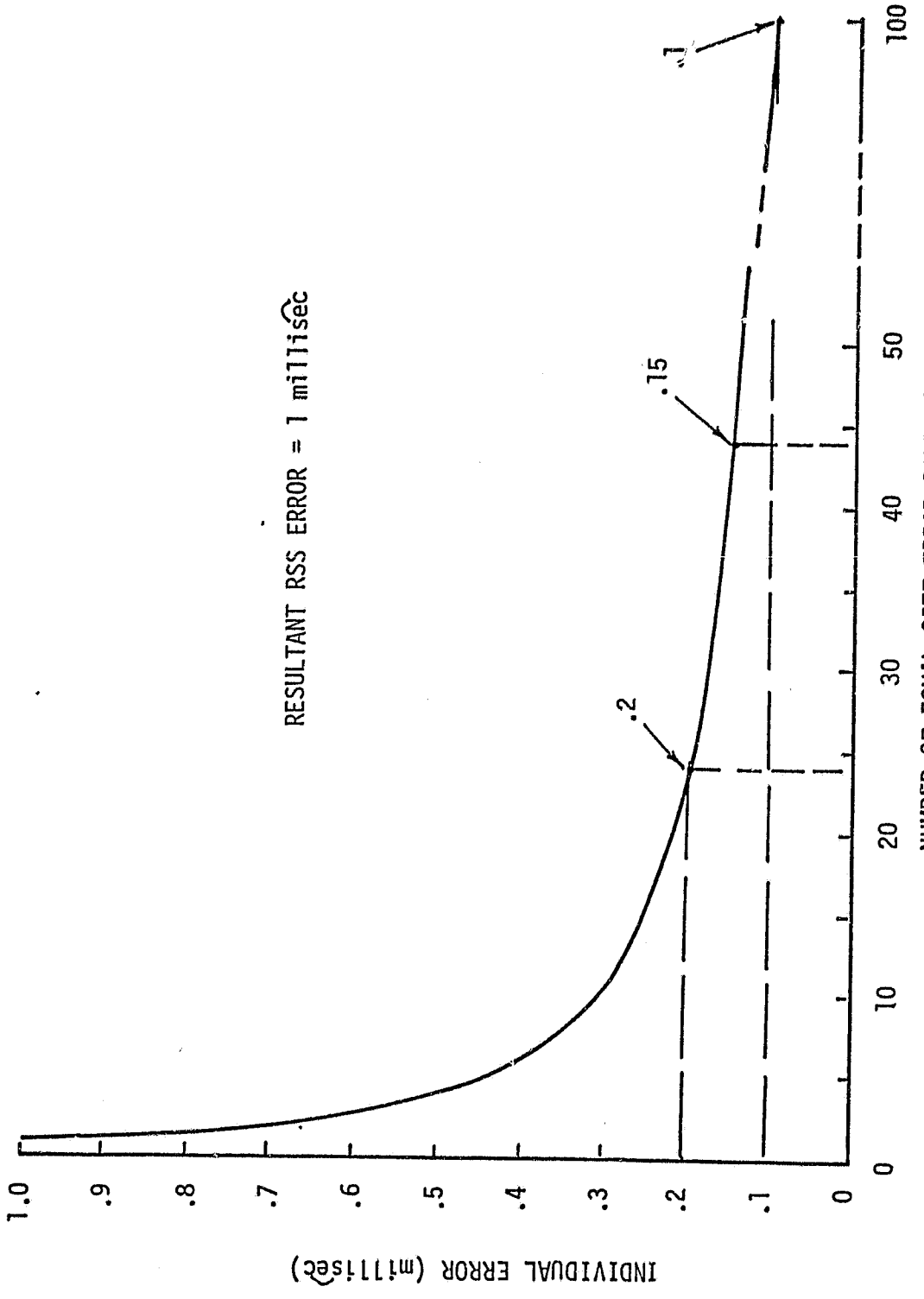


Figure II-2. Budget for average individual error.

4. P. B. Eby, "Electrical Torques on the ELectrostatic Gyro in the Gyro Relativity Experiment," NASA TM-78311, Marshall Space Flight Center, October 1980.

5. NASA TR-R443, "Magnetic Torque on a Rotating Superconducting Sphere," NASA, Washington, D.C., May 1975.

6. D. H. Sheingold, Analog-digital Conversion Handbook, Analog Devices, Inc., Norwood, Mass, 1972.

7. L. Motz and A. Duveen, Essentials of Astronomy, Columbia University Press, New York, 1971.

8. A. Sommerfeld, Mechanis, Academic Press, New York, 1964.

9. J. T. Anderson and C. W. F. Everitt, "Limit on the Measurement of Proper Motion and the Implications for the Relativity Gyroscope Experiment," Stanford University, Hansen Laboratory, November 6, 1979.

CHAPTER III

FINITE-WORDLENGTH INDUCED ERRORS IN KALMAN FILTERING COMPUTATION

1. Introduction

The problem of finite-wordlength effect on digital computations has been investigated extensively during the past twenty years. Finite-wordlength property of a computer requires either rounding or chopping to be used to limit the wordlength of a number. Since most computers use rounding technique, only rounding errors will be considered in the sequel.

There are two approaches to analyze the rounding error, the first approach considers the statistical nature of rounding errors, and treats them as noise generated in the system. This approach has been widely used by those in the field of digital signal processing. In the statistical error analysis one is usually after the ensemble average and standard deviation of the final error based on the estimated characteristics of source errors and their propagation through computation steps. This approach does not seem to be sufficiently reliable for the analysis of GP-B data reduction errors for two reasons. First, GP-B's four experiment gyros represent only a small sample, their combined statistical characteristics may deviate a good deal from those of the population statistics. Thus the use of statistical analysis here may not give a reliable result. Secondly, the GP-B data reduction involves Kalman filtering and other rather complex computations. The exact statistical nature of rounding error generation by and propagation through these computations is not easy to establish. Therefore a more conservative approach is needed.

The second approach is to establish bounds for the rounding errors involved in computation. This approach provides a very conservative, though rather pessimistic, result for rounding error analysis. This approach has often been used by those doing numerical analysis. Because of the unusual precision required of the GP-B and the expensiveness of the experiment the use of error bound approach provide a much more reliable results for the error analysis. Therefore this approach will be used for ensuing rounding error analysis. Since Kalman filtering is the main activity in GP-B data reduction, the present chapter is devoted to the analysis of rounding error in Kalman filtering computation.

2. Rounding Procedure in Floating Point Representation

Let x be a number

$$x = (\pm \cdot d_1 d_2 \text{ ---}) \times b^e \qquad (1)$$

where b is the base of the number system used and e, an integer, is the exponent. In general the mantissa part of the number may have infinite number of digits for an exact representation, such as for $\sqrt{2}$. The number (1) may also be represented in the form

$$x = u \cdot b^e + v \cdot b^{e-t} \qquad (2)$$

where $\frac{1}{b} \leq |u| < 1$, $0 < |u| < 1$, and u contains only t digits.

Examples: Base 10 numbers.

(a) $12.3456 = .1234 \times 10^2 + .56 \times 10^{-2}$

Here b = 10, t = 4, and 3 = 2

(b) $-.0123456 = -.1234 \times 10^{-2} + (-.56) \times 10^{-5}$

Here b = 10, t = 4, and 3 = -1

The rounding procedure drops off the second term on the right side of (2) by appropriately adjusting the value of the first term. Thus,

after rounding, x becomes x_R which has a t-digit mantissa $\cdot d_1 d_2 \text{ --- } d_t$ and an exponent b^e. The conventional round-off procedure for any number is as follows:

$$\hat{x} = \begin{cases} u \cdot b^e & \text{if } |v| < \frac{1}{2} \\ u\, b^e + b^{e-t} & \text{if } v \geq \frac{1}{2} \\ u\, b^e - b^{e-t} & \text{if } v \leq -\frac{1}{2} \end{cases} \qquad (3)$$

Note that u and v always have the same sign.

Examples: b = 10 and t = 4

(a) $x = 765.4567 = .7654 \times 10^3 + .567 \times 10^{-1}$

Here $v \geq \frac{1}{2}$ and e = 3, so

$\hat{x} = u\, b^e + b^{e-t} = .7654 \times 10^3 + 10^{3-4} = .7655 \times 10^3$

(b) $x = 123.426 = .1234 \times 10^3 + .26 \times 10^{-1}$

Here $v < \frac{1}{2}$, so

$\hat{x} = u\, b^e = .1234 \times 10^{-3}$

(c) $x = -765.4567 = -.765 \times 10^3 - .567 \times 10^{-1}$

Here $v \leq -\frac{1}{2}$, so

$x = u\, b^e - b^{e-t} = -.7654 \times 10^3 - 10^{3-4} = -.7655 \times 10^3$

These results are intuitively obvious. The reason for going through the formulations of Equations (1), (2) and (3) is to pave a way for the subsequent analysis of rounding errors.

3. Rounding Errors in Floating Pointing Representation

The "absolute rounding error" in x_R is defined as

$$|\tilde{x}| = |x_R - x| \geq 0 \tag{4}$$

From (2) and (3), it is clear that

$$|\tilde{x}| \leq \frac{1}{2} b^{e-t} \tag{5}$$

Examining (1) shows that $|u\ b^e| \geq b^{e-1}$ because $u \geq b^{-1}$; and $|x| \geq |u\ b^e|$ because the second term, having similar sign, is dropped. Hence

$$|x| \geq |u\ b^3| \geq b^{e-1} \tag{6}$$

DEfine the "absolute relative rounding error" ε as

$$\varepsilon = \frac{|x_R - x|}{|x|} = \frac{|\tilde{x}|}{|x|} \tag{7}$$

By (5), (6) gives

$$\varepsilon \leq \frac{1}{2} b^{1-t} = \beta \tag{8}$$

The quantity β is called the "unit rounding error" which represents the absolute bound of rounding error in the floating point representation of a number of base b and having a t-digit mantissa. It is an important parameter in the analysis of rounding errors.

Example: Consider b = 10 and t = 4

Then $\beta = \frac{1}{2} b^{1-t} = \frac{1}{2} 10^{-3}$

Let $x = 767.4567 = .7654 \times 10^3 + .567 \times 10^{-1}$

then $\hat{x} = .7655 \times 10^3$

$|\hat{x}| = |x_R - x| = .0433$

$$\varepsilon = \frac{.0433}{765.4567} = .56568 \times 10^{-4} < \beta$$

For the sake of comparison, the chopping error in floating point representation of a number will be analyzed next.

4. Chopping Error in Floating Point Representation

For a floating point number in the form of (2), a t-digit chopped number is given by

$$x_c = u\, b^e \tag{9}$$

Define the "absolute chopping error" $\tilde{x}_c$ as

$$|\tilde{x}_c| = |x_c - x| = |v|^{e-t} \tag{10}$$

Since $|x| <$ |

$$|\tilde{x}_c| \leq b^{e-t} \tag{11}$$

Define the "absolute relative chopping error" as

$$\varepsilon_c = \frac{|\tilde{x}_c|}{|x|} \tag{12}$$

clearly,

$$\varepsilon_c \leq \frac{b^{e-t}}{b^{e-1}} = b^{1-t} = \beta_c \tag{13}$$

where β_c is called the "unit chopping error." Comparing (13) and (8) shows

$$\beta_c = 2\beta \tag{14}$$

Example: $b = 10$ and $6 = 4$

Then $\beta_c = 10^{1-4} = 10^{-3}$

Let $x = 765.4567 = .7654 \times 10^3 + .567 \times 10^{-1}$

then $x_c = .7654 \times 10^{-3}$

$$|\tilde{x}_c| = |x_c - x| = .567 \times 10^{-1}$$

$$\varepsilon_c = \frac{.0567}{765.4567} = .74073 \times 10^{-4} < \beta_c$$

5. Rounding Error in Basic Computer Arithmatic Operations

For the convenience of the subsequent analysis, notation for rounded floating foint number is defined here in two equivalent forms. Let x be a floating point number. The rounded value of x is denoted by $\hat{x}$ or $fl(x)$.

Let "*" denote any of the four basic arithmatic operations +, -, x, and / . The computer value of x*y is $fl(x*y)$, which is related to the exact value x*y by

$$fl(x*y) = (x*y)(1+\varepsilon) \tag{15}$$

where ε is the actual relative rounding error. The absolute relative error in (x*y) is bounded by

$$\varepsilon = \left|\frac{fl(x*y) - (x*y)}{(x*y)}\right| \leq \beta \tag{16}$$

where E is the unit rounding error.

6. Rounding Error in Composite Computer Arithmatic Operations

Repeated Additions and subtractions consider the sum

$$s = x_1 + x_2 + x_3 + x_4$$

$$= ((x_1+x_2) + x_3) + x_4$$

The rounded value is

$$\hat{s} = \{[(x_1+x_2)(1+\varepsilon_1) + x_3](1+\varepsilon_2) + x_4\}(1+\varepsilon_3)$$

$$= (x_1+x_2)(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3) + x_3(1+\varepsilon_2)(1+\varepsilon_3) + x_4(1+\varepsilon_3)$$

$$\simeq (x_1+x_2)(1+\varepsilon_1+\varepsilon_2+\varepsilon_3) + x_3(1+\varepsilon_2+\varepsilon_3) + x_4(1+\varepsilon_3)$$

The rounding error is

$$\tilde{s} = (x_1x_2)(\varepsilon_1+\varepsilon_2+\varepsilon_3) + x_3\,(\varepsilon_2+\varepsilon_3) + x_4\,(\varepsilon_3)$$

$$= (x_1+x_2+x_3+x_4)(\varepsilon_1+\varepsilon_2+\varepsilon_2) - x_3\varepsilon_1 - x_4\,(\varepsilon_1+\varepsilon_2)$$

The absolute relative rounding error is founded by

$$\varepsilon \leq 3\beta - \beta x_3 - 2\beta x_4 \leq 3\beta$$

In general, for a sum of n terms

$$s = \sum_{j=1}^{n} x_j \tag{17}$$

the absolute relative error is bounded by

$$\varepsilon \leq (n-1)\beta \tag{18}$$

Repeated Multiplication and Division. Consider the following combination of product and quotient

$$Q = \frac{x_1x_2}{y_1} = (x_1x_2)/y_1$$

the rounded value is

$$\hat{Q} = \frac{x_1x_2(1+\varepsilon_1)}{y_1}(1+\eta_1) \simeq \frac{x_1x_2}{y_1}(1+\varepsilon_1+\eta_1)$$

where ε_1 and η_1 are relative rounding errors due to multiplication and division, respectively. The rounding error and the absolute relative rounding error are, respectively,

$$\tilde{Q} = \frac{x_1x_2}{y_1}(\varepsilon_1+\eta_1)$$

and

$$\varepsilon = \varepsilon_1 + \eta_1 \leq 2\beta$$

For the general case

$$Q = \frac{x_1 x_2 \text{ --- } x_n}{y_1 y_2 \text{ --- } y_m} \tag{19}$$

The absolute relative sounding error is bounded by

$$\varepsilon \leq (n+m-1)\beta \tag{20}$$

7. Norms of Vectors and Matrices

Norms of vectors and matrices are useful in the analysis of rounding errors in matrix operations. The following definitions of norm will be adopted in this study.

For an n-vector $\underline{x}$ with elements x_j, define the vector norm as

$$||x|| = \underset{j}{\text{Max}}|x_j| \tag{21}$$

Clearly, the norm has the following properties:

(i) $||\underline{x}|| \geq 0$

(ii) $||\underline{x}|| = 0$ only if $\underline{x} = 0$

(iii) $||\underline{x}+\underline{y}|| \leq ||\underline{x}|| + ||\underline{y}||$

(iv) $||a\ \underline{x}|| = |a| \cdot ||\underline{x}||$ for any $a \varepsilon R$

For a maxn matrix A with elements a_{ij}, define the matrix norm as

$$||A|| = \underset{j}{\text{Max}} \sum_{i=1}^{n} |a_{ij}| \tag{22}$$

This norm has the following properties:

(i) $||A|| \geq 0$

(ii) $||A|| = 0$ only if $A = 0$

(iii) $||A+B|| \leq ||A|| + ||B||$

(iv) $||a\ A|| = |a| \cdot ||A||$ for any $a \varepsilon R$

(v) $||AB|| \leq ||A|| \cdot ||B||$

8. Rounding Error in Matrix Addition

Let A and B be two nxm matrices, the rounded sum of them is

$$fl[A+B] = A+B+R \tag{23}$$

where R is the rounding error matrix. By the definition of matrix norm (22) and in view of (15), the norm of the rounding error matrix is bounded by

$$||R|| \leq \beta \, ||A+B|| \tag{24}$$

where $\beta = \frac{1}{2} b^{1-t}$ as given by (8). The relative norm of $||R||$ is bounded by

$$\varepsilon = \frac{||R||}{||A+B||} \leq \beta \tag{25}$$

which is the same as the relative rounding error of the sum of two numbers as shown in (16).

9. Rounding Error in Matrix Multiplication

Since elements of a matrix product are inner products of vector pairs, the rounding error associated with an inner product will be analyzed first. The result will then be used to analyze the rounding error in a matrix product.

9.1. Rounding Error in Inner Product

Consider the inner product of two 3-vectors $\underline{a}$ and $\underline{b}$

$$I = \underline{a}^T\underline{b} = a_1b_1 + a_2b_2 + a_3b_3$$

The rounded value of $\underline{a}^T\underline{b}$ is

$$\hat{I} = fl[\underline{a}^T\underline{b}]$$

$$= \{[a_1b_1(1+\varepsilon_1) + a_2b_2(1+\varepsilon_2)](1+\varepsilon_3) + a_3b_3(1+\varepsilon_4)\}(1+\varepsilon_5)$$

$$\simeq a_1b_1\ (1+\varepsilon_1+\varepsilon_3+\varepsilon_5) + R_2b_2\ (1+\varepsilon_2+\varepsilon_3+\varepsilon_5) + a_3b_3\ (1+\varepsilon_4+\varepsilon_5)$$

where ε_i's are relative rounding errors associated with basic arithmatic operations. The rounding error in $\hat{I}$ is

$$\tilde{I} = a_1b_1\ (\varepsilon_1+\varepsilon_3+\varepsilon_5) + a_2b_2\ (1+\varepsilon_2+\varepsilon_3+\varepsilon_5) + a_3b_3\ (\varepsilon_4+\varepsilon_5)$$

The absolute value of this rounding error is bounded by

$$|\tilde{I}| \leq 3\beta\ |a_1b_1| + 3\beta\ |a_2b_2| + 2\beta\ |a_3b_3|$$

In general, the aboslute rounding error of the inner product of two n-vectors is bounded by

$$|\tilde{I}| \leq \beta n\left\{|a_1b_1| + \sum_{j=2}^{n} (n+2-j)\ |a_jb_j|\right\} \tag{26}$$

The expression for the absolute relative rounding error for an inner product appears cumbersome and is not given here.

9.2. Rounding Error in Matrix Products

Consider the matrix product C = AB where A is mxn and B is nxp. The number n will be called "interface dimension" for matrices A and B. Using the result of (26) the absolute error of the elements of C is bounded by

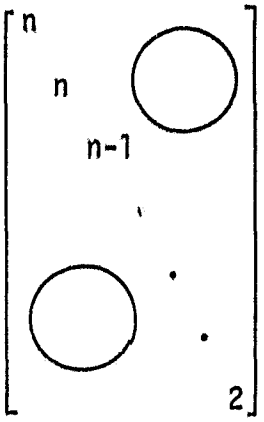
$$|\tilde{c}_{ij}| \leq \beta\ \{n|a_{i1}|\cdot|b_{1j}| + n|a_{i2}|\cdot|b_{2j}| + (n-1)|a_{i3}|\cdot|b_{3j}| + \cdots + 2|a_{in}|\cdot|b_{nj}|\} \tag{27}$$

Let $[\tilde{c}]$ be a matrix whose elements are $|\tilde{c}_{ij}|$, [A] be a matrix whose elements are $|a_{ij}|$, and [B] be a matrix whose elements are $|b_{ij}|$.

Then, based on (27), one has

$$[\tilde{C}] \overset{*}{\leq} \beta[A]D[B]$$

where the symbol "$\overset{*}{\leq}$" means that the comparison is done on element by element basis for the left and right hand matrices, and

$D =$ 

is nxn

Clearly $||D|| = n$. The norm of the rounding error matrix $\tilde{c}$ is therefore bounded by

$$||\tilde{C}|| \leq n\beta||A||\cdot||B|| = \frac{n}{2} b^{1-t}||A||\cdot||B|| \qquad (28)$$

Generalize the above result to a product of N matrices

$$P = M_1M_2 - - - M_N \qquad (29)$$

with interface dimensions $d_1, d_2, - - - d_{N-1}$. Let

$$P_i = M_1M_2 - - - M_i$$

Then the result of (38) implies the following rounded matrices, with ε being the worst error,

$$P_2 = fl[M_1M_2] = M_1M_2(1+d_1\varepsilon)$$

and

$$\hat{P}_N = fl[M_1M_2] = M_1M_2(1+d_1\varepsilon)$$

$$\hat{P}_3 = f\,[\hat{P}_1M_3] = \hat{P}_1M_3(1+d_2\varepsilon)$$

$$= M_1M_2M_3(1+d_1\varepsilon)(1+d_2\varepsilon) \simeq M_1M_2M_3[1+(d_1+d_2)\varepsilon]$$

and

$$\hat{P}_N \simeq M_1 M_2 - - - M_N[1+(d_1+ - - - +d_{N-1})\varepsilon] \tag{30}$$

Rounding errors in $\hat{P}_N$ is

$$\tilde{P}_N = M_1 M_2 - - - M_N(d_1+d_2 + - - -+d_{N-1})\varepsilon \tag{31}$$

The norm of this error matrix is therefore

$$||P_N|| \leq \beta\left(\sum_{i=1}^{N-1} d_i\right) \prod_{j=1}^{N} ||M_j|| \tag{32}$$

The results of (24) and (31) can be used jointly to handle the matrix equation containing both products and sums. This will be demonstrated by the following two examples.

Example1 Compute

$R = ABC+D$

where all matrices are nxn. The rounded R is

$$\hat{R} = ABC(1+2n\varepsilon)(1+\varepsilon) + D(1+\varepsilon)$$

The rounding error of $\hat{R}$ is

$$\tilde{R} = [(2n+1)ABC + D]\varepsilon$$

and its norm is bounded by

$$||\tilde{R}|| \leq \beta[(2n+1)||ABC|| + ||D||]$$

Example 2 Compute

$R = ABC + D$

where A is nxm, B is mxr, c is rxs, and D is nxs. Then

$$\hat{R} = ABC[1+(m+r)\varepsilon](1+\varepsilon) + D(1+\varepsilon)$$

$$\tilde{R} = ABC(1+m+r)\varepsilon + D\varepsilon$$

$$||\tilde{R}|| \leq \beta[(1+m+r)||ABC|| + ||D||]$$

These two examles shows that the rounding error norm of matrix addition does not involve the dimension of the matrices, but that of matrix product involves all the interace dimensions.

10. Rounding Error in Matrix Inversion, First Approach

Let A be nonsingular nxn matrix, its inverse A^{-1} satisfies the relationship

$$A\,A^{-1} = I, \text{ the identify matrix}$$

Let $\underline{u}_j$ be the jth colume vector of I and $\underline{h}_j$ be the jth column of A^{-1}. Then $\underline{h}_j$ is the solution of

$$A\underline{x} = \underline{u}_j \quad j = 1 \text{ to } n \tag{33}$$

Thus A^{-1} can be obtained by solving (29) n times using different $\underline{u}_j$ each time. The solution is usually done by a method based on the Gaussian elimination with partial pivoting. The present concern is the rounding error associated with the conputation of A^{-1}. The analysis will be done in two steps: First, find error in A^{-1} computed from the exact A. Second, find error in A^{-1} computed from $A' = A + \Delta A$ where ΔA is the error in A.

Rounding Error in A^{-1} when A is exact. Let $\hat{\underline{h}}_j$ be computer solution of (33). Define the "residue" associated with $\underline{h}_j$ as

$$\underline{r}_j = A\,\hat{\underline{h}}_j - \underline{u}_j \tag{34}$$

The error in $\hat{\underline{h}}_j$ is

$$\tilde{\underline{h}}_j = \hat{\underline{h}}_j - \underline{h}_j = A^{-1}\underline{r}_j \tag{35}$$

The rounding error matrix for the computer inverse of A is

$$E = [\tilde{\underline{h}}_1\ \tilde{\underline{h}}_2 - - - \tilde{\underline{h}}_n] \tag{36}$$

Define the "residue matrix" for the computer inverse

$$R = [\underline{r}_1\ \underline{r}_2\ - - -\ \underline{r}_n] \tag{37}$$

Then

$$E = A^{-1}R \tag{38}$$

The norm of E is bounded by

$$||E|| \leq ||A^{-1}|| \cdot ||R|| \tag{39}$$

and the relative norm of E is bounded by

$$\varepsilon \leq ||R|| \tag{40}$$

Rounding Error in A^{-1} when $A+\Delta A$ is inverted. Let A be erred to $A+\Delta A$, then the computer solution $\hat{\underline{h}}_j$ for

$$[A+\Delta A]\underline{x} = \underline{u}_j \qquad j = 1 \text{ to } n \tag{41}$$

must satisfy

$$[A+\Delta A]\hat{\underline{h}}_j = \underline{u}_j + \underline{r}_j \qquad j = 1 \text{ to } n$$

where $\underline{r}_j$ is the residue. Then

$$\hat{\underline{h}}_j + A^{-1}\Delta\hat{A}\underline{h}_j = A^{-1}\underline{u}_j + A^{-1}\underline{r}_j = \underline{h}_j + A^{-1}\underline{r}_j$$

The error in $\hat{\underline{h}}_j$ is

$$\tilde{\underline{h}}_j = \hat{\underline{h}}_j - \underline{h}_j = A^{-1}[\underline{r}_j - \Delta A\ \hat{\underline{h}}_j] \tag{42}$$

Using the notation defined in (36) and (37), (42) gives the error matrix

$$E = A^{-1}R - A^{-1}\Delta A\ f\ell[A^{-1}] \simeq A^{-1}R - A^{-1}\Delta A A^{-1}$$

The norm of the error matrix is bounded by

$$||E|| \leq ||A^{-1}|| \cdot ||R|| + ||A^{-1}||^2 \cdot ||\Delta A|| \tag{43}$$

the relative error norm is

$$\varepsilon = \frac{||E||}{||A^{-1}||} = ||R|| + ||A^{-1}|| \cdot ||\Delta A|| \tag{44}$$

Comparing (44) to (40) shows that the latter is a special case of the former where $\Delta A = 0$. Eq.(44) appears elegant, but its practical usefulness is in doubt. The problem is that the residue matrix R cannot easily be obtained. In addition, both (43) and (44) do not explicitly depend on any wordlength related parameter, such as, the unit sounding error β.

11. Rounding Error in Matrix Inversion, Second Approach

The usual method of matrix inversion by a computer is based on a repeated use of Gaussian elimination procedure. The procedure consists of two parts, namely, triangularization of a matrix and back substitution. The rounding error for each part will be analyzed first, followed by the analyze of the resultant error. In the following analysis ε will denote the worst value of any rounding error ε_i. Thus $|\varepsilon| \leq \beta$.

11.1. Rounding Error in Matrix Triangularization

Consider a 3x3 Matrix Equation

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}_{\text{A matrix}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{45}$$

Let $a_{ij}(0) = a_{ij}$ and $b_i(0) = b_i$ for i, j = 1 to n. The first step is to condition the first column. Let

$$m_{21} = -\frac{a_{21}}{a_{11}} = -\frac{a_{21}(0)}{a_{11}(0)} \tag{46}$$

then let

$$a_{22}(1) = a_{22}(0) + m_{21}a_{12}(0) = a_{22}(0) = \frac{a_{21}(0)}{a_{11}(0)}\, a_{12}(0)$$

$$\hat{a}_{22}(1) = f\ell[a_{22}(0) - \frac{a_{21}(0)}{a_{11}(0)} a_{12}(0)]$$

$$= a_{22}(0)(1+\varepsilon) = \frac{a_{21}(0)}{a_{11}(0)} a_{12}(0)\ (1+3\varepsilon)$$

$$= [a_{22}(0) - \frac{a_{21}(0)}{a_{11}(0)} a_{12}(0)] + a_{22}(0) - \frac{a_{21}(0)}{a_{11}(0)} a_{12}(0)(3\varepsilon)$$

$$= a_{22}(1) + a_{22}(1)(3\varepsilon) - a_{22}(0)(2\varepsilon)$$

Similarly

$$\hat{a}_{23}(1) = a_{23}(1) + a_{23}(1)\ (3\varepsilon) - a_{23}(0)\ (2\varepsilon)$$

$$\hat{a}_{32}(1) = a_{32}(1) + a_{32}(1)\ (3\varepsilon) - a_{32}(0)\ (2\varepsilon)$$

$$\hat{a}_{33}(1) = a_{33}(1) + a_{33}(1)\ (3\varepsilon) - a_{33}(0)\ (2\varepsilon)$$

After the first step, the error in the new A, designated $\hat{A}(1)$, is given by $\tilde{A}(1)$ whose elements are

$$a_{ij}(1) = \begin{cases} 0 & i=1;\ j=1,2,3 \\ a_{ij}(1)\ (3\varepsilon) - a_{ij}(1)\ (2\varepsilon) & i,j=2,3 \end{cases} \qquad (47)$$

Define

$$\hat{A}_1(1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & a_{22}(0) & a_{23}(0) \\ 0 & a_{32}(0) & a_{33}(0) \end{bmatrix}$$

and

$$\hat{A}_1(1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & a_{22}(1) & a_{23}(1) \\ 0 & a_{32}(1) & a_{33}(1) \end{bmatrix}$$

Then

$$\tilde{A}(1) = [\tilde{a}_{ij}(1)] = 3\varepsilon A_1(1) - 2\varepsilon A_1(0)$$

$$\simeq \beta[3\hat{A}_1(1) - 2\hat{A}_1(0)] \tag{48}$$

Expressing $\tilde{A}$ in terms of $\hat{A}$ rather than A is important since $\hat{A}$ is available from the computer but not A.

For the b_i coefficients we have

$$b_i(1) = b_i(0) = m_{i1} b_i(0) = b_i(0) - \frac{a_{i1}(0)}{a_{11}(0)} b_i(0)$$

Then

$$b_i(1) = f\ell[b_i(1)] = b_i(0)(1+\varepsilon) - \frac{a_{i1}(0)}{a_{11}(0)} b_1(0)(1+3\varepsilon)$$

$$= b_i(1) + \varepsilon\, b_i(0) - 3\varepsilon \frac{a_{i1}(0)}{a_{11}(0)} b_1(0)$$

$$= b_i(1) + 3\varepsilon\, b_i(1) - 2\varepsilon\, b_i(0)$$

Define

$$\hat{b}_1(0) = \begin{bmatrix} 0 \\ b_2(0) \\ b_3(0) \end{bmatrix} \quad \text{and} \quad b_1(0) = \begin{bmatrix} 0 \\ b_2(1) \\ b_3(1) \end{bmatrix}$$

then

$$\tilde{\underline{b}}(1) = 3\varepsilon\, \underline{b}_1(1) - 2\varepsilon\, \hat{\underline{b}}_1(0) \simeq \beta[3\hat{\underline{b}}_1(1) - 2\, \hat{\underline{b}}_1(0)] \tag{49}$$

Again, expressing $\tilde{\underline{b}}$ in terms of $\hat{\underline{b}}$ rather than $\underline{b}$ is important since $\hat{\underline{b}}$ is available from the computer but $\underline{b}$ is not.

After the first reduction step, one has

$$\begin{bmatrix} a_{11}(0) & a_{12}(0) & a_{13}(0) \\ 0 & \hat{a}_{22}(0) & \hat{a}_{23}(1) \\ 0 & \hat{a}_{32}(0) & \hat{a}_{33}(1) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1(0) \\ \hat{b}_2(1) \\ \hat{b}_3(1) \end{bmatrix}$$

where $\hat{a}_{ij}$ and $\hat{b}_j$ are rounded quantities. Their errors will be compounded to the new rounded quantities in the next step of the reduction process.

The second step of the reduction concerns the second column of the matrix. Let

$$m_{32} = - \frac{\hat{a}_{32}(1)}{\hat{a}_{22}(1)} \tag{50}$$

and let

$$a_{33}(2) = \hat{a}_{33}(1) + m_{32}\, \hat{a}_{23}(1) = \hat{a}_{33}(1) - \frac{\hat{a}_{32}(1)}{\hat{a}_{22}(1)}\, \hat{a}_{23}(1)$$

The rounded value is

$$\hat{a}_{33}(2) = f\ell[a_{33}(2)] = a_{33}(1)(1+\varepsilon) - \frac{\hat{a}_{32}(1)}{\hat{a}_{22}(1)}\, \hat{a}_{23}(1)(1+3\varepsilon)$$

$$= a_{33}(2) + 3\varepsilon\, a_{33}(2) - 2\varepsilon\, \hat{a}_{33}(1)$$

Let

$$\hat{A}_2(1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \hat{a}_{33}(1) \end{bmatrix}$$

and

$$A_2(2) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a_{33}(2) \end{bmatrix}$$

Then

$$\tilde{A}(2) = 3\varepsilon A_2(2) - 2\varepsilon \hat{A}_2(1) \simeq \beta[3\hat{A}_2(\ ) - 2\hat{A}_2(1)] \tag{51}$$

For the b_1 coefficients in the second reduction step,

$$b_3(2) = \hat{b}_3(1) + m_{32}\,\hat{b}_2(1) = \hat{b}_3(1) - \frac{\hat{a}_{32}(1)}{\hat{a}_{22}(1)}\,\hat{b}_2(1)$$

The rounded value is

$$\hat{b}_3(2) = f\ell[b_3(2)] = \hat{b}_3(1)(1+\varepsilon) - \frac{\hat{a}_{32}(1)}{\hat{a}_{22}(1)}\,\hat{b}_2(1)(1+3\varepsilon)$$

$$= b_3(2) + 3\varepsilon\, b_3(2) - 2\varepsilon\, \hat{b}_3(1)$$

Let

$$\underline{\hat{b}}_2(1) = \begin{bmatrix} 0 \\ 0 \\ \hat{b}_3(1) \end{bmatrix} \quad \text{and} \quad \underline{b}_2(2) = \begin{bmatrix} 0 \\ 0 \\ b_3(2) \end{bmatrix}$$

Then

$$\underline{\tilde{b}}(2) = [3\,\underline{b}_2(2) - 2\,\underline{\hat{b}}_2(1)]\,\varepsilon \simeq \beta[3\,\underline{\hat{b}}_2(2) - 2\,\underline{\hat{b}}_2(1)] \tag{52}$$

The resultant errors,

$$\tilde{A} = \tilde{A}(1) + \tilde{A}(2) = \beta[3\,\hat{A}_1(1) - 2\,\hat{A}_1(0)] + \beta[3\,\hat{A}_2(2) - 2\,\hat{A}_2(1)]$$

$$= \beta\left\{ \sum_{i=1}^{2} [3A_i(i) - 2A_i(i-1)] \right\} \tag{53}$$

$$\tilde{\underline{b}} - \tilde{\underline{b}}(1) + \tilde{\underline{b}}(2) = \beta[3\hat{\underline{b}}_1(1) - 2\hat{\underline{b}}_1(0)] + \beta[3\hat{\underline{b}}_2(2) - 2\hat{\underline{b}}_2(1)]$$

$$= \beta\left\{\sum_{i=1}^{2}[3\hat{\underline{b}}_i(i) - 2\hat{\underline{b}}_i(i-1)]\right\} \tag{54}$$

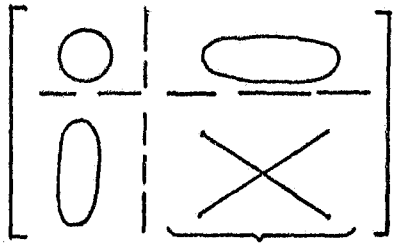
Generalization to an nxn matrix A

$$A = A(0) = \hat{A}(0) = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ | & & | \\ | & & | \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \tag{55}$$

$$\hat{a}_{ij}(0) = a_{ij}(0) = a_{ij} \tag{56}$$

The matrix obtained after the kth reduction step is

$$A(k) = \begin{bmatrix} a_{11}(k) & \cdots & a_{in}(k) \\ | & & | \\ | & & | \\ a_{n1}(k) & \cdots & a_{nn}(k) \end{bmatrix} \tag{57}$$

$A_j(k) =$  (58)

The (n-i)x(n-i) lower right
diagonal block matrix from A(k)

The resultant reduction or triangularization errors in A and $\underline{b}$ are, respectively,

$$A = \beta\left\{\sum_{i=1}^{n-1}[3\hat{A}_i(i) - 2\hat{A}_i(i-1)]\right\} \tag{59}$$

$$\tilde{\underline{b}} = \beta\left\{\sum_{i=1}^{n-1} [3\hat{\underline{b}}_i(i) - 2\hat{\underline{b}}_i(i-1)]\right\} \tag{60}$$

Finally, the norms of errors due to triangularization are given by

$$||\tilde{A}|| = \beta\left\{\sum_{k=1}^{n-1} [3||A_k(k)|| + 2||A_k(k-1)||]\right\} \tag{61}$$

$$||\tilde{\underline{b}}|| = \beta\left\{\sum_{k=1}^{n-1} [3||b_k(k)|| + 2||b_k(k-1)||]\right\} \tag{62}$$

Note that after the (n-1)th reduction step, the original matrix A has been reduced to an upper triangular form. Denote it by $\hat{A}_T = \hat{A}(n-1)$. Thus,

$$\hat{A}_T = \begin{bmatrix} \hat{a}_{11}(0) & \hat{a}_{12}(0) & \cdots & a_{1k}(0) & \cdots & \hat{a}_{1n}(0) \\ & \hat{a}_{22}(1) & \cdots & a_{2k}(0) & \cdots & \hat{a}_{2n}(0) \\ & & \ddots & & & \\ & & & \hat{a}_{kk}(k-1) & \cdots & \hat{a}_{kn}(0) \\ & & & & \ddots & \\ & & & & & a_{nn}(n-1) \end{bmatrix} \tag{63}$$

Denote the associated $\hat{\underline{b}}$ vector $\hat{\underline{b}}_T$, then

$$\hat{b}_T = \hat{b}(n-1) = [b_1(0)\ b_2(1) \cdots b_n(n-1)]^T \tag{64}$$

11.2. Rounding Error in Back Substitution

This problem is approached as follows. Consider the equation

$$A\,\underline{x} = \underline{b} \tag{65}$$

where A is an nxn upper triangular matrix. Let $\hat{\underline{x}}$ be the solution of this equation which contains rounding errors, then find $\hat{A}$ and $\hat{\underline{b}}$

such that

$$\hat{A}\,\hat{\underline{x}} = \hat{\underline{b}} \tag{66}$$

has error-free solution $\hat{\underline{x}}$. Thus the rounding error problem has been transformed to a error problem caused by perturbations in A and $\underline{b}$. Define

$$\Delta A = \hat{A}-A \qquad \Delta\underline{x} = \hat{\underline{x}}-\underline{x} \qquad \Delta\underline{b} = \hat{\underline{b}}-\underline{b} \tag{67}$$

Then (65) gives

$$(\hat{A}-\Delta A)(\hat{\underline{x}}-\Delta\underline{x}) = \hat{\underline{b}} - \Delta\underline{b}$$

$$\Delta\underline{x} = \hat{A}^{-1}[\,\underline{b} - A\,\hat{\underline{x}}] \tag{68}$$

In (68), $\Delta\underline{x}$ is the rounding error in $\hat{x}$, $\hat{A}^{-1}$ and $\hat{\underline{x}}$ are computed by the computer while solving (65), and $\Delta\underline{b}$ and ΔA are computed from formulas to be developed. Assume that A is an upper triangular matrix, the solution of (65) involves only the back substitution operations. Rounding error due to back substitution will now be analyzed.

3x3 Triangular Matrix Equation. Consider the equation

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{69}$$

The equivalent perturbed equation for evaluating rounding errors is

$$\begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} & \hat{a}_{13} \\ 0 & \hat{a}_{22} & \hat{a}_{23} \\ 0 & 0 & \hat{a}_{33} \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{bmatrix} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix} \tag{70}$$

Write (69) as

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \qquad (a)\\ a_{22}x_2 + a_{23}x_3 &= b_2 \qquad (b)\\ a_{33}x_3 &= b_3 \qquad (c)\end{aligned}\right\} \tag{71}$$

From (71c)

$$x_3 = \frac{b_3}{a_{33}}$$

The rounded value is

$$\hat{x}_3 = \frac{b_3}{a_{33}}(1+\varepsilon) = x_3(1+\varepsilon) \tag{72}$$

$$\therefore \quad x_3 = \frac{\hat{x}_3}{1+\varepsilon} \tag{73}$$

Substituting into (71c) and rearranging terms, give

$$\underbrace{a_{33}}_{\hat{a}_{33}}\,\hat{x}_3 = \underbrace{b_3\,(1+\varepsilon)}_{\hat{b}_3} \tag{74}$$

where $\hat{a}_{33}$ and $\hat{b}_3$ are also defined. Next, from (71b)

$$x_2 = \frac{1}{a_{22}}\,[b_2 - a_{23}\hat{x}_3]$$

Its rounded value is

$$\begin{aligned} x_2 &= \frac{1}{a_{22}}\,[b_2 - a_{23}\hat{x}_3(1+\varepsilon)]\,(1+2\varepsilon)\\ &= \frac{1}{a_{22}}\,[b_2 - a_{23}x_3)1+2\varepsilon)]\,(1+2\varepsilon)\\ &= x_2 + \frac{2\varepsilon b_2 - 4\varepsilon\, a_{23}x_3}{a_{22}} \end{aligned} \tag{75}$$

$$\therefore \quad x_2 = \hat{x}_2 - \frac{2\varepsilon b_2 - 4\varepsilon\, a_{23}x_3}{a_{22}}$$

$$= \hat{x}_2 - \frac{2\varepsilon\, b_2 - 4\varepsilon\, a_{23}\hat{x}_3/(1+\varepsilon)}{a_{22}} \tag{76}$$

Substituting (76) and (73) into (71b) and rearranging terms, give

$$\underbrace{a_{22}\hat{x}_2}_{\hat{a}_{22}} + \underbrace{a_{23}(1+3\varepsilon)\hat{x}_3}_{\hat{a}_{23}} = \underbrace{b_2(1+2\varepsilon)}_{\hat{b}_2} \tag{77}$$

where $\hat{a}_{22}$, $\hat{a}_{23}$, and $\hat{b}_2$ are also defined. Next, from (71a)

$$x_1 = \frac{1}{a_{11}}\,[b_1 - a_{12}\hat{x}_2 - a_{13}\hat{x}_3]$$

Its rounded value is

$$\hat{x}_1 = \frac{1}{a_{11}}\,[b_1(1+3\varepsilon) - a_{12}\hat{x}_2(1+4\varepsilon) - a_{13}\hat{x}_3(1+3\varepsilon)]$$

With the help of (72) and (75),

$$\hat{x}_1 = x_1 + \frac{1}{a_{11}}\left\{3\varepsilon b_1 - 4\varepsilon a_{12}x_2 - 2\varepsilon\,\frac{a_{12}}{a_{22}}\,b_2 + 4\varepsilon\,\frac{a_{12}a_{23}}{a_{22}}\,x_3 - 4\varepsilon\, a_{13}x_3\right\}$$

Using the approximations $x_2 \simeq \hat{x}_2$ and $x_3 \simeq \hat{x}_3$, x_1 is expressed in terms of $\hat{x}_1$, $\hat{x}_2$, and $\hat{x}_3$ as

$$x_1 = \hat{x}_1 - \frac{1}{a_{11}}\left\{3\varepsilon b_1 - 2\varepsilon b_2\,\frac{a_{12}}{a_{22}} - 4\varepsilon 9_{12}\hat{x}_2 + \frac{4\varepsilon}{1+\varepsilon}\,(\frac{a_{12}a_{23}}{a_{22}} - a_{13})\hat{x}_3\right\} \tag{78}$$

Substituting (78), (76), and (73) into (71a) and rearranging terms, give

$$\underbrace{a_{11}\,\hat{x}_1}_{\hat{a}_{11}} + \underbrace{a_{12}(1+4\varepsilon)\hat{x}_2}_{\hat{a}_{12}} + \underbrace{a_{13}(1+3\varepsilon)}_{\hat{a}_{13}}\,\hat{x}_3 = \underbrace{b_1(1+3\varepsilon)}_{\hat{b}_1} \tag{79}$$

where $\hat{a}_{11}$, $\hat{a}_{12}$, $\hat{a}_{13}$, and $\hat{b}_1$ are also defined. Finally, put (74), (77) and (79) into a single matrix equation.

$$\underbrace{\begin{bmatrix} \hat{a}_{11}=a_{11} & \hat{a}_{12}=a_{12}(1+4\varepsilon) & \hat{a}_{13}=a_{13}(1+3\varepsilon) \\ 0 & \hat{a}_{22}=a_{22} & \hat{a}_{23}=a_{23}(1+3\varepsilon) \\ 0 & 0 & \hat{a}_{33}=a_{33} \end{bmatrix}}_{\hat{A} = A + \Delta A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\hat{x}} = \underbrace{\begin{bmatrix} \hat{b}_1=b_1(1+3\varepsilon) \\ \hat{b}_2=b_2(1+2\varepsilon) \\ \hat{b}_3=b_3(1+\varepsilon) \end{bmatrix}}_{\hat{b} = b + \Delta b} \tag{80}$$

from which one easily gets

$$\Delta A = \begin{bmatrix} 0 & 4\varepsilon a_{12} & 3\varepsilon a_{13} \\ 0 & 0 & 3\varepsilon a_{23} \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix} \tag{81}$$

$$\Delta b = \begin{bmatrix} 3b_1 \\ 2b_2 \\ b_3 \end{bmatrix} \varepsilon = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \varepsilon \tag{82}$$

Generalization to nxn Triangular Matrix Equation. Consider

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{in} \\ & 2_{22} & \cdots & a_{2n} \\ & & & 1 \\ & & & 1 \\ & & & a_{nn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ 1 \\ 1 \\ x_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ 1 \\ 1 \\ b_n \end{bmatrix}}_{b} \tag{83}$$

Its equivalent perturbation equation for evaluating rounding errors can be obtained by generalizing the result of (80), which is

$$\begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} & \cdots & \hat{a}_{1n} \\ & \hat{a}_{22} & \cdots & \hat{a}_{2n} \\ & & \ddots & 1 \\ & & & 1 \\ & & & \hat{a}_{nn} \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ 1 \\ 1 \\ \hat{x}_3 \end{bmatrix} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ 1 \\ 1 \\ \hat{b}_n \end{bmatrix} \qquad (84)$$

$\hat{A}$ $\quad$ $\underline{\hat{x}}$ $\quad$ $\underline{\hat{b}}$

where

$$\hat{a}_{kj} = \begin{cases} a_{kj} & k=j \\ a_{kj}[1+(n-j+3)\varepsilon] & 1 \le k < j \le n \\ 0 & \text{otherwise} \end{cases} \qquad (85)$$

$$b_k = b_k[1+(n-k+1)\varepsilon] \qquad k=1 \text{ to } n \qquad (86)$$

Thus,

$$\Delta a_{kj} = \begin{cases} a_{kj}(n-j+3)\varepsilon & 1 \le k < j \le n \\ 0 & \text{otherwise} \end{cases} \qquad (87)$$

$$\Delta b_k = b_k(n-k+1)\varepsilon \qquad k = 1 \text{ to } n \qquad (88)$$

$$\Delta A = \begin{bmatrix} 0 & (2n-2)a_{12} & (2n-3)a_{13} & -- & 4a_{1(n-1)} & 3a_{1n} \\ & 0 & (2n-3)a_{23} & -- & 4a_{2(n-1)} & 3a_{2n} \\ & & & & & \\ & & & & & 3a_{(n-1)n} \\ & & & & & 0 \end{bmatrix} \varepsilon$$

$$\leq \begin{bmatrix} 0 & a_{12} & a_{13} & -- & a_{1n} \\ & 0 & a_{23} & -- & a_{2n} \\ & & & & \vdots \\ & & & & a_{(n-1)n} \\ & & & & 0 \end{bmatrix} \begin{bmatrix} 0 & & & \\ & n+1 & & \\ & & & \\ & & 4 & \\ & & & 3 \end{bmatrix} \beta \qquad (89)$$

$$A_D \qquad M_A$$

$$\Delta \underline{b} = \begin{bmatrix} nb_1 \\ (n-1)b_2 \\ \\ 1 \\ 1 \\ b_n \end{bmatrix} \varepsilon \leq \begin{bmatrix} n & & & \\ & n+1 & & \\ & & & \\ & & & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \beta \qquad (90)$$

$$M_b \qquad \underline{b}$$

where matrices A_D, M_A, and M_b are also defined.

11.4. A Numerical Example

Consider inverting the following matrix

$$A = \begin{bmatrix} 3.235 & -1.234 & 3.256 \\ 1.023 & -5.235 & 0.921 \\ 1.336 & 2.120 & -8.235 \end{bmatrix}$$

using computer of different finite decimal wordlength. Then evaluate the corresponding error norms using the procedure of Figure 1. The effectiveness of the procedure is examined by comparing these error norms to the corresponding actual error norms. The actual norms are approximately obtained by using a computer having a much longer decimal wordlength. The result is given in Table 1, which shows that error norms obtained by using the proposed procedure are indeed very conservative. Notice that error norm decreases with increasing wordlength. It is interesting to note that when the proposed method is used all error norms have the same mantissa.

Table I

Matrix Inversion Error Norms

Wordlength: No. of places after decimal point	Error norm	
	By proposed method	Actual value
3	3.706×10^{-2}	7.782×10^{-5}
5	3.706×10^{-4}	4.530×10^{-7}
8	3.706×10^{-7}	4.672×10^{-10}

11.3. Resultant Rounding Error in the Inverse Matrix

Two set of equivalent perturbations for A and $\underline{b}$ have been obtained to account for rounding errors. One set, $\tilde{A}$ and $\tilde{\underline{b}}$ as given by (59) and (60), account for errors from triangularization. The second set, ΔA and $\Delta\underline{b}$ as given by (89) and (90), account for errors from back substitution. The resultant equivalent perturbations for A and are given by the sums

$$\delta A = \tilde{A} + \Delta A \tag{91}$$

$$\delta\underline{b} = \tilde{\underline{b}} + \Delta\underline{b} \tag{92}$$

The resultant rounding error in $\underline{x}$ in the solution of

$$A\,\underline{x} = \underline{u}_j \qquad j = 1 \text{ to } n, \tag{93}$$

where $\underline{u}_j$ is the jth column vector of the identify matrix I, is given by

$$\delta\underline{x}_j = \widehat{A^{-1}}[\delta\underline{b}_j - \delta A\,\hat{\underline{x}}_j] \tag{94}$$

This equation is obtained in a way similar to that of (68). Rounding error in $\hat{A}^{-1}$ is then given by

$$\delta(A^{-1}) = [\delta\underline{x}_1 \text{ - - - } \delta\underline{x}_n]$$

$$= \widehat{A^{-1}}\underbrace{[(\delta\underline{b}_1 - \delta A\hat{\underline{x}}_1) \text{ - - - } (\delta\underline{b}_n - \delta A\hat{\underline{x}}_n)]}_{D} \tag{95}$$

The error norm of the computer's inverse matrix of A is therefore

$$||\delta(A^{-1})|| \leq ||\hat{A}^{-1}|| \cdot ||D|| \tag{96}$$

where the matrix D has been defined in (95). The relative error norm is

$$\varepsilon \leq \frac{||\delta(A^{-1})||}{||\hat{A}^{-1}||} = ||D|| \tag{97}$$

It is obvious that the evaluation of (96) or (97) involves a good deal of computation and should be done by a computer. Figure III-1 is a computation block diagram for this purpose.

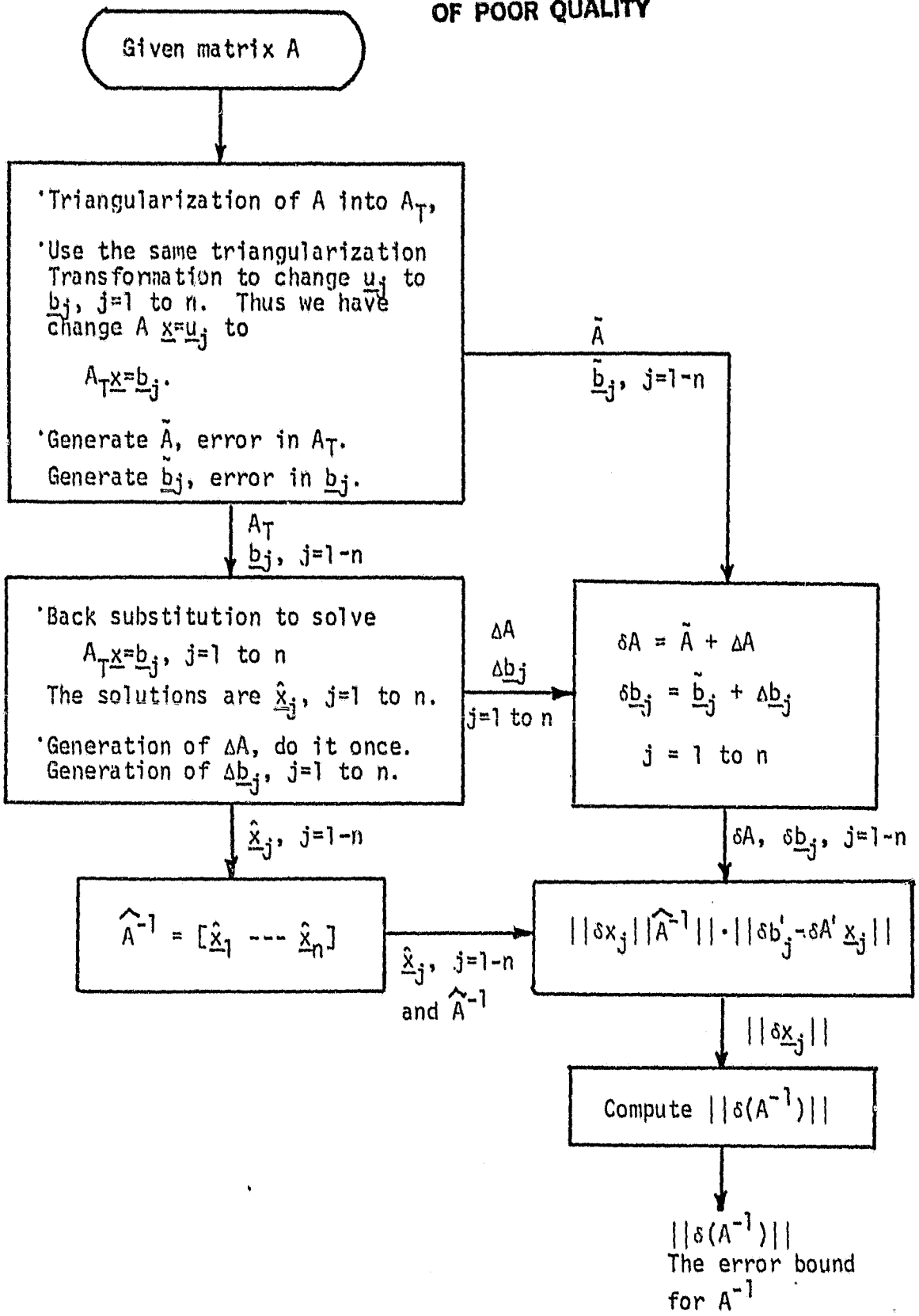


Figure III-1. Flow chart for evaluating rounding error bound of matrix inverse.

12. Rounding Error Bound for Kalman Filtering

Consider a process modeled by the following set of sequations.

$$\underline{x}_k = \phi_{k-1}\,\underline{x}_{k-1} + \underline{\omega}_{k-1} \qquad \dim \underline{x}_k = u$$

$$\underline{z}_k = H_k\,\underline{x}_k + \underline{v}_k \qquad \dim \underline{z}_k = m$$

$$E\,\underline{x}(0) = \underline{x}_0 \qquad E[\tilde{\underline{x}}(0)\,\tilde{\underline{x}}(0)^T] = P_o$$

$$\underline{\omega}_k \sim N(0,Q_k) \qquad V_k \sim N(0,R_k)$$

$$E[\underline{\omega}_k\, v_j] = 0 \qquad \text{all } j,k$$

The Kalman Filter algorithm consists of the following equations.

$$\underline{x}_k^* = \phi_{k-1}\,\underline{x}_{k-1}^* + K_k[\underline{z}_k - H_k\,\phi_{k-1}\,\underline{x}_{k-1}^*] \qquad \underline{x}^*(0) = \underline{x}_0 \tag{98}$$

$$K_k = P_{kp}\,H_k^T[H_k P_{kp} H_k^T + R_k]^{-1} \quad \text{or} \quad K_k = P_k H_k^T R_k^{-1} \tag{99}$$

$$P_{kp} = \phi_{k-1} P_{k-1} \phi_{k-1}^T + Q_k \tag{100}$$

$$P_k = P_{kp} - K_p\,H_k\,P_{kp} \tag{101}$$

where $\underline{x}^*$ is the estimate of $\underline{x}$. The present interact is to find the bound of the rounding error norm for $\underline{x}^*$. For the sake of convenience, the astrick "*" will be dropped, and $\hat{\underline{x}}$ will denote the rounded value of the estimate.

Error in Rounded P_{kp}. Recall (100), that is,

$$P_{kp} = \phi_{k-1} P_{k-1} \phi_{k-1}^T + Q_{k-1}$$

Its rounded result is

$$\hat{P}_{kp} = \phi_{k-1} P_{k-1} \phi_{k-1}^T (1+\overline{2n+1}\varepsilon) + Q_{k-1}(1+\varepsilon)$$

$$= P_{kp}(1+\overline{2n+1}\varepsilon) - 2n\varepsilon Q_{k-1} \tag{102}$$

The rounding error is

$$\tilde{P}_{kp} = (2n+1)\varepsilon P_{kp} - 2n\varepsilon Q_{k-1} \leq (2n+1)\varepsilon P_{kp} \tag{103}$$

Error in Rounded P_k. Recall (101), that is,

$$P_k = [I - K_k H_k] P_{kp}$$

Its rounded value is

$$\hat{P}_k = [I(1+\varepsilon) - K_k H_k (1+\overline{m+1}\varepsilon)]\hat{P}_{kp}(1+n\varepsilon)$$

$$= [I(1+\overline{n+1}\varepsilon) - K_k H_k (1+\overline{n+m+1}\varepsilon)]\hat{P}_{kp}$$

$$= [I(1+\overline{n+1}\varepsilon) - K_k H_k (1+\overline{n+m+1}\varepsilon)][P_{kp} + \tilde{P}_{kp}]$$

$$\underset{\sim}{} P_k + [(n+1)\varepsilon I - K_k H_k (n+m+1)\varepsilon] P_{kp} + [I - K_k H_k]\tilde{P}_{kp}$$

$$= P_k[1+(n+m+1)\varepsilon] - m\varepsilon P_{kp} + [I-K_k H_k]\tilde{P}_{kp} \tag{104}$$

$$\tilde{P}_k = (n+m+1)\varepsilon P_k - m\varepsilon P_{kp} + [I-K_k H_k]\tilde{P}_{kp}$$

$$\leq (n+m+1)\varepsilon P_k - m\varepsilon P_{kp} + \underbrace{[I-K_k P_k](2n+1)}_{(2n+1)\varepsilon P_k} P_{kp}$$

$$\leq (3n+m+2)\varepsilon P_k \tag{105}$$

Error in Rounded K_k. Recall (99), that is,

$$K_k = P_k H_k^T \hat{R}_k^{-1}$$

Its rounded form is

$$\hat{K}_k = \hat{P}_k H_k^T \hat{R}_k^{-1}(1+n+m\epsilon) \tag{106}$$

By (104) and (103)

$$\hat{P}_k \leq (1+\overline{n+m+1}\epsilon)P_k - m\epsilon P_{kp} + \underbrace{[I-K_kH_k](2n+1)\epsilon P_{kp}}_{(2n+1)\epsilon P_k}$$

$$= (1+\overline{3n+m+2}\epsilon)P_k - m\epsilon P_{kp}$$

Assume R_k diagonal

$$\hat{R}_k^{-1} = R_k^{-1}(1+\epsilon)$$

Than (106) becomes

$$\hat{K}_k \leq [(1+\overline{3n+m+1}\epsilon)P_k - m\epsilon P_{kp}]H_k^T R_k^{-1}(1+\epsilon)(1+\overline{n+m}\epsilon)$$

$$= (1+\overline{4n+2m+3}\epsilon)P_k H_k^T R_k^{-1} - m\epsilon P_{kp} H_k^T R_k^{-1} \tag{107}$$

The rounding error is

$$\tilde{K}_k \leq (4n+2m+3)\epsilon K_k - m\epsilon P_{kp} H_k^T R_k^{-1}$$

$$\leq (4n+2m+3)\epsilon K_k \tag{108}$$

Error in Rounded $\underline{x}_k$. Recall (98), which is

$$\underline{x}_k = \phi_{k-1}x_{k-1} + K_k[\underline{z}_k - H_k\phi_{k-1}\underline{x}_{k-1}]$$

$$= [I-K_kH_k]\phi_{k-1}\underline{x}_{k-1} + K_k\underline{Z}_k \tag{98}$$

Let

$$F_k = [I - K_k H_k]\phi_{k-1} \tag{109}$$

$$\underline{u}_k = K_k \underline{z}_k \tag{110}$$

Then (98) becomes

$$\underline{x}_k = F_k \underline{X}_{k-1} + \underline{u}_k \tag{111}$$

Develop the following rounded quantities.

$$\begin{aligned}
\hat{F}_k &= [I(1+\varepsilon) - \hat{K}_k H_k(1+\overline{m+1}\varepsilon)]\phi_{k-1}(1+n\varepsilon) \\
&= [I(1+\overline{n+1}\varepsilon) - K_k(1+\overline{4n+2m+3}\varepsilon)H_k(1+\overline{n+m+1}\varepsilon)]\phi_{k-1} \\
&= [I(1+\overline{n+1}\varepsilon) - K_k H_k(1+\overline{5n+3m+4}\varepsilon)]\phi_{k-1} \\
&= F_k(1+\overline{5n+3m+4}\varepsilon) - (4n+3n+3)\varepsilon I
\end{aligned} \tag{112}$$

$$\begin{aligned}
\hat{\underline{u}}_k &= \hat{K}_k \underline{z}_k(1+m\varepsilon) \\
&= K_k(1+\overline{4n+2m+3}\varepsilon)\underline{z}_k(1+m\varepsilon) \\
&= K_k \underline{z}_k(1+\overline{4n+3m+3}\varepsilon) \\
&= \underline{u}_k(1+\overline{4n+3m+3}\varepsilon)
\end{aligned} \tag{113}$$

$$\begin{aligned}
\hat{\underline{x}}_k &= \hat{F}_k \underline{x}_{k-1}(1+\overline{n+1}\varepsilon) + \hat{\underline{u}}_k(1+\varepsilon) \\
&\leq (1+\overline{5n+3m+4}\varepsilon)F_k \underline{X}_{k-1}(1+\overline{n+1}\varepsilon) + \underline{u}_k(1+\overline{4n+3m+3}\varepsilon)(1+\varepsilon) \\
&= (\overline{6n+3m+5}\varepsilon+1)F_k \underline{X}_{k-1} + (\overline{4n+3m+4}\varepsilon+1)\underline{u}_k
\end{aligned} \tag{114}$$

Define

$$\hat{\hat{F}}_k = (1+\overline{6n+3m+5}\varepsilon)F_k \tag{115}$$

$$\hat{\hat{u}}_k = (1+\overline{4n+3m+4}\varepsilon)u_k \tag{116}$$

Then (114) can be written as

$$\hat{x}_k = \hat{\hat{F}}_k \hat{x}_{k-1} + \hat{\hat{u}}_k \tag{117}$$

Case of k = 3 Eq. (111) gives the exact x_3 as

$$x_3 = F_3F_2F_1 x_0 + F_3F_2u_1 + F_3u_2 + u_3 \tag{118}$$

The rounded x_3 is

$$\hat{x}_3 = \hat{\hat{F}}_3\hat{\hat{F}}_2\hat{\hat{F}}_1x_0(1+\overline{3n+3}\varepsilon) + \hat{\hat{F}}_3\hat{\hat{F}}_2\hat{\hat{u}}_1(1+\overline{2n+3}\varepsilon)$$

$$+ \hat{\hat{F}}_3\hat{\hat{u}}_2(1+\overline{n+2}\varepsilon) + \hat{\hat{u}}_3(1+\varepsilon)$$

Using (115) and (116), and combining terms,

$$x_3 = F_3F_2F_1x_0[1+3(7n+3m+6)\varepsilon] + F_3F_2u_1[1+\overline{18n+9m+17}\varepsilon]$$

$$+ F_3u_2[1+\overline{11n+6m+11}\varepsilon] + u_3[1+\overline{4n+3m+5}\varepsilon]$$

$$= x_3[1+3(7n+3m+6)\varepsilon] - (3m+1)\varepsilon F_3F_2u_1$$

$$- (10n+3m+7)\varepsilon F_3u_2 - (17n+6m+13)\varepsilon u_3 \tag{119}$$

The rounding error is

$$\tilde{x}_3 = 3(7n+3m+6)\varepsilon x_3 - (3n+1)\varepsilon F_3F_2\, u_1$$

$$- (10n+3m+7)\varepsilon F_3\, u_2 - (17n+6m+13)\varepsilon u_3 \tag{120}$$

Its norm is bounded by

$$||\tilde{\underline{x}}_3|| \leq 3(7n+3m+6)\varepsilon||x_3|| - (3n+1)\varepsilon||F_3F_2\underline{u}_1|| - (10n+3m+7)\varepsilon||F_3\underline{u}_2|| - (17n+6m+13)\varepsilon||u_3|| \quad (121)$$

Assume $||F_k|| \leq F$, $||\underline{u}_k|| \leq U$, and $F^i\underline{u} \leq B$ for $i = 0$ to 2, then

$$||\tilde{\underline{x}}_3|| \leq 3(7n+3m+6)\varepsilon||x_3|| + (30n+9m+21)\varepsilon B \leq \beta\left\{3(7n+3m+6)||\underline{x}_3|| + \left[3(3n+1) + \frac{3(3-1)}{2}(7n+3m+6)\right]B\right\} \quad (122)$$

Case of $k=4$ Again, by (111),

$$\underline{x}_4 = F_4F_3F_2F_1\,\underline{x}_0 + F_4F_3F_2\,\underline{u}_1 + F_4F_3\,\underline{u}_2 + F_4\,\underline{u}_4 \quad (123)$$

Following similar derivation, gives the norm of rounding error as

$$||\tilde{\underline{x}}_4|| \leq \beta\left\{4(7n+3m+b)||\underline{x}_4|| + \left[4(3n+1) + \frac{4(4-1)}{2}(7n+3m+b)\right]B\right\} \quad (124)$$

The general case k. From the equation pattern of (122) and (124) for k=3 and 4, the general case is found to be

$$\boxed{||\tilde{\underline{x}}_k|| \leq \beta\left\{k(7n+3m+b)||\underline{x}_k|| + \left[k(3n+1) + \frac{k(k-1)}{2}(7n+3m+6)\right]B\right\}} \quad (125)$$

Eq. (125) is the main result of this chapter, which gives the bound of error norm for the rounded state estimate. The bound depends on β, the unit rounding error; k, the number of iterations; n and m, the dimension parameters of the process; $||\underline{x}_k||$, the norm of the estimated state; and B, a quantity depends on K_k, H_k, ϕ_k and $\underline{u}_k$. The usefulness of this equation is at providing a general idea on the desired number of digits for the mantissa of the computer's floating number system. The following example

will illustrate this point

Example. Consider a one year GP-B operation where relativistic data are taken every 10 seconds. Assume that the Kalman filtering involved in data reduction is also operated at 10 second iteraction period. Then, at the end of one year period, the value of k would be $k=365 \times 24 \times 3600/10 = 3.1536 \times 10^6$. Assume $\underline{x}_k$ be a 10-vector and $\underline{z}_k$ be a 2-vector, so n=10 and m=2. Then,

$$k(7n+3m+b) = 2.6490 \times 10^8$$

$$k(3n+1) + 0.5k(k-1)(7n+3m+6) = 1.2949 \times 10^{16}$$

Just for the sake of discussion, assume only one term at the right-hand side of (125) dominates the result. If the first term dominates, one may estimate the desired β from

$$k(7n+3m+6)\ \beta=1$$

so

$$\beta = 0.3775 \times 10^{-8}$$

comparing to (8), the formula $\beta = 0.5 \times 10^{1-t}$, gives $t \underline{\sim} 9$, therefore 9 digits are desired for the mantissa of the floating point number system. On the other hand, if the second term of (125) dominates, one may estimate the desired β from

$$[k(3n+1) + 0.5k(k-1)(7n+3m+6)]\ \beta=1$$

so

$$\beta = .7723 \times 10^{-16}$$

comparing to (8), gives $t \underline{\sim} 17$. Hence 17 digits are desired for the mantissa.

13. Remarks

1. The main result of this chapter, given by (125), is a "first-cut" result which can probably be further refined to tighten the predicted bound while maintaining its reliability. This result was obtained after several different approaches to the problem had been attempted.

2. It is desirable to find out the effectiveness of (125) by a computer emulation of the GP-B data reduction. This has not been done for three reasons. First, there is no time and resource allocated in the contract period covered by this report. Secondly, this emulation work is a major activity not foreseen at the start of the contract, therefore it was not planned. Finally, the computation details of the GP-B data reduction, being developed by the GP-B group at the Stanford University, is not completely available to the present contract.

3. Besides Kalman filtering the GP-B data reduction also contains other types of computations which should also be included in the rounding error analysis. Attention to a more complete error analysis is intended for for the next contract period.

4. During the present contract period only the finite wordlength induced computation errors were investigated. The truncation error involved in the chosen method of numerical integration for the GP-B data reduction has not been studied, thus its improtance in the overall error analysis is not known at this time. The effect of Truncation error will be investigated in the future.

5. Other indirect errors exists, which also enter the result through GP-B data reduction. These errors are caused by imperfection in compensation data. Effect of this type of errors has not been investigated, but is intended for future study.

14. References

1. J. H. Wilkinson, Rounding Errors in Algebraic Process, Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

2. G. E. Forsythe, M. A. Malcolm, and C. B. Moler, Computer Methods for Mathematical Computations, Prentice-Hall, ENglewood Cliffs, New Jersey, 1977.

3. J. B. Mankin and J. C. Hung, "On Rounding Errors in the Computation of Transition Matricies," Proceeding of the Joint Automatic Control Conference, American Society of Chemical Engineers, 1969, pp. 60-64.

4. Pat. H. Sterbenz, Floating-Point Computation, Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

5. James S. Vandergraft, Introduction to Numerical Computation, Academic Press, New York, 1978.

6. D. I. Steinberg, Computational Matrix Algebra, McGraw-Hill, New York, 1974.

CHAPTER IV

MEASUREMENT GEOMETRY FOR GP-B EXPERIMENT

This chapter describes in detail the development of a measurement model for the GP-B experiment. The model represents the relationships among the orientations of various components of the experiment, including the directions of the reference star, gyro spin axis, readout normal, and the spacecraft's roll axis. With this model, the effect of spacecraft rolling and of component misalignments can be investigated through analysis and computer simulation.

For the convenience of analysis, six coordinate frames are defined with the help of Figure IV-1. All frames are co-origined at the spacecraft's center of rotation.

1. Absolute frame [a-frame: x_a, y_a, z_a] — This frame is stationary in orientation with respect to the universe, i.e. With respect to the ideal distant stars. Its x_a-axis is along the line-of-sight (LOS) to the reference star Rigel.
2. Intermediate frame [i-frame: x_i, y_i, z_i] — This frame is stationary in orientation with respect to the universe. Its x_1 axis is along the roll axis of the spacecraft. This frame is related to the absolute frame by Euler angles α_1, α_2, and α_3.
3. Roll frame [ϕ-frame: x_ϕ, y_ϕ, z_ϕ] — This frame is fixed to the spacecraft with its x_ϕ axis along the roll axis of the spacecraft. Therefore x_ϕ and x_i axes coincide.

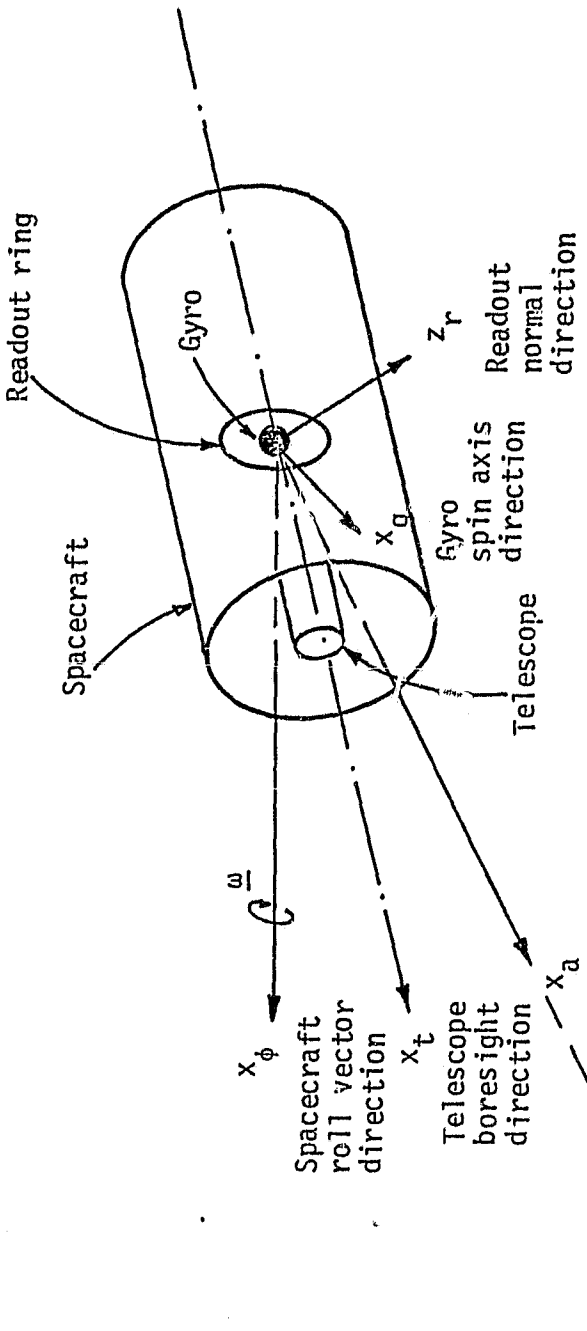


Figure IV-1. Definition of directions.

4. Telescope frame [t-frame: x_t, y_t, z_t] —— This frame is also fixed to the spacecraft, with its x_t axis along the boresight of the telescope. The misalignment between this frame and the roll frame is represented by Euler angles β_1, β_2, and β_3.
5. Readout frame [r-frame: x_r, y_r, z_r] — This frame is again fixed to the spacecraft, which represents the orientation of the gyro readout ring. Its z_r axis is normal to the plane of readout ring. The misalignment between this frame and the telescope frame is represented by Euler angles γ_1, γ_2, and γ_3.
6. Gyro frame [g-frame: x_g, y_g, z_g] — THis frame is inertially fixed, but not absolutely fixed. The relative angular motion of this frame with respect to the absolute frame is the relativistic drift of the inertial space around the GP-B gyro. The orientation of this frame with respect to the readout frame is represented by Euler angles δ_1, δ_2, and δ_3, while that with respect to the intermediate frame is represented by Euler angles θ_1, θ_2, and θ_3.

Relationships among all frames are also depicted in Figure IV-2. Ideally, β's, γ's, and δ's are zero, meaning that the roll, telescope, and readout frames are all lined up. The nonzero values for α's represent the deviation of telescope's boresight from the LOS to star. The θ's will not be zero, they are the relativistic drift angles. It is assumed that angles, α's, β's, γ's, δ's, and θ's are all sufficiently small that small angle approximations of trigonometry apply.

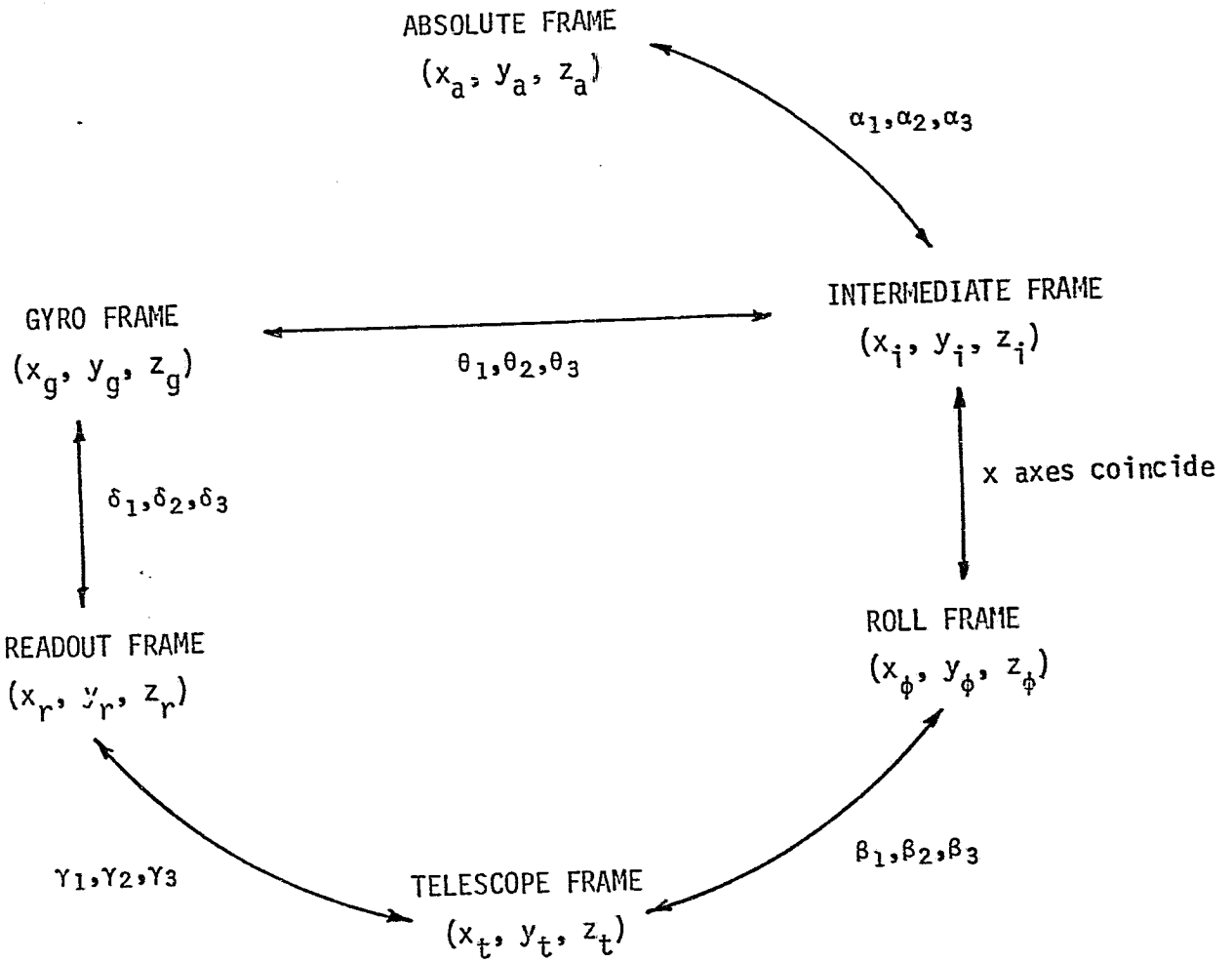


Figure IV-2. Relationships among coordinate frames.

In the experiment, an optical system meaures the angle between the LOS to star and the telescope's boresight. This amounts to measuring the projection of x_a axis of the absolute frame on the telescope frame. With all misalignments included, this relationship is modeled by the following matrix equation.

$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & \beta_3 & -\beta_2 \\ -\beta_3 & 1 & \beta_1 \\ \beta_2 & -\beta_1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & c\omega t & S\omega t \\ 0 & -s\omega t & c\omega t \end{bmatrix} \begin{bmatrix} 1 & \alpha_3 & -\alpha_2 \\ -\alpha_3 & 1 & \alpha_1 \\ \alpha_2 & -\alpha_1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad (1)$$

star LOS unit vector in t-frame

Transformation to t-frame. Accounts for misalignments β_1, β_2, β_3.

Transformation to ϕ-frame. Accounts for rolling about x_1 axis

Transformation to i-frame. Accounts for telescope offset α_1, α_2, α_3.

star LOS unit vector in a-frame

The gyro and its readout ring measure the relative angular displacement between the gyro spin axis x_g and the readout normal z_r. This amounts to measureing the projection of x_g on the readout frame. Note that the readout frame rolls with the spacecraft since it is body fixed. With all misalignment included, the projection of x_g on (x_r, y_r, z_r) is modeled by

$$\begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \gamma_3 & -\gamma_2 \\ -\gamma_3 & 1 & \gamma_1 \\ \gamma_2 & -\gamma_1 & 1 \end{bmatrix}}_{\text{Transformation to r-frame. Accounts for misalignments } \gamma_1, \gamma_2, \gamma_3.} \underbrace{\begin{bmatrix} 1 & \beta_3 & -\beta_2 \\ -\beta_3 & 1 & \beta_1 \\ \beta_2 & -\beta_1 & 1 \end{bmatrix}}_{\text{Transformation to t-frame. Accounts for misalignments } \beta_1, \beta_2, \beta_3.} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & C\omega t & S\omega t \\ 0 & -S\omega t & C\omega t \end{bmatrix}}_{\text{Transformation to } \phi\text{-frame. Accounts for rolling about } x_i \text{ axis}} \underbrace{\begin{bmatrix} 1 & \theta_3 & -\theta_2 \\ -\theta_3 & 1 & \theta_1 \\ \theta_2 & -\theta_1 & 1 \end{bmatrix}}_{\text{Transformation to i-frame. Accounts for gyro drift angles } \theta_1, \theta_2, \theta_3} \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{\text{Gyro spin axis unit vector in g-frame}} \quad (2)$$

Gyro spin axis unit vector in r-frame

Under the ideal condition, β's and γ's are all zero. Then (1) and (2) reduct to

$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & C\omega t & S\omega t \\ 0 & -S\omega t & C\omega t \end{bmatrix} \begin{bmatrix} 1 & \alpha_3 & -\alpha_2 \\ -\alpha_3 & 1 & \alpha_1 \\ \alpha_2 & -\alpha_1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & C\omega t & S\omega t \\ 0 & -S\omega t & C\omega t \end{bmatrix} \begin{bmatrix} 1 & \theta_3 & -\theta_2 \\ -\theta_3 & 1 & \theta_1 \\ \theta_2 & -\theta_1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad (4)$$

Solving (3) and (4) for α's and θ's, gives

$$\begin{bmatrix} C\omega t & -S\omega t \\ S\omega t & C\omega t \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} -\alpha_3 \\ \alpha_2 \end{bmatrix} = \text{offset of star's LOS from } x_t \text{ axis} \quad (5)$$

$$\begin{bmatrix} C\omega t & -S\omega t \\ S\omega t & C\omega t \end{bmatrix} \begin{bmatrix} y_r \\ z_r \end{bmatrix} = \begin{bmatrix} -\theta_3 \\ \theta_2 \end{bmatrix} = \text{offset of gyro spin axis from } x_r\text{-axis} \tag{6}$$

Notice that when small angle approximations are used α_1 and θ_1 cannot be determined. Since x_t and x_r axes are colinear, subtracting (5) from (6) yields the gyro's relativistic drift angles ψ's with respect to the intermediate frame. Because of the absence of α_1, the transformation from the intermediate frame to the absolute frame is simply via an identity matrix. Therefore the relativistic drift angles of the gyro with respect to the obsolute frame are given by the same ψ's.

$$\begin{bmatrix} \psi_3 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} \alpha_3-\theta_3 \\ \theta_2-\alpha_2 \end{bmatrix} = \underbrace{\begin{bmatrix} C\omega t & -S\omega t \\ S\omega t & C\omega t \end{bmatrix}}_{\Omega^{-1}} \begin{bmatrix} y_r-y_t \\ z_r-z_t \end{bmatrix} \tag{7}$$

The Ω^{-1} matrix in (7) represents the deroll operation to be done by the data reduction computer.

To study the effect of misalignments on the accuracy of the experiments, (1) and (2) are used. In there, β's and γ's are unknown quantities, since the effect of known misalignment can be compensated computationally. Angles α's and θ's are desired, and they are estimated from the measurement by the deroll computation, giving

$$\begin{bmatrix} \hat{\alpha}_3 \\ -\hat{\alpha}_2 \end{bmatrix} = \begin{bmatrix} C\omega t & S\omega t \\ -S\omega t & C\omega t \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} \tag{8}$$

$$\begin{bmatrix} \hat{\theta}_3 \\ -\hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} C\omega t & S\omega t \\ -S\omega t & C\omega t \end{bmatrix} \begin{bmatrix} y_r \\ z_r \end{bmatrix} \tag{9}$$

The errors due to misalignments are

$$\begin{bmatrix} \Delta\alpha_2 \\ \Delta\alpha_3 \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_2 - \alpha_2 \\ \hat{\alpha}_3 - \alpha_3 \end{bmatrix} \tag{10}$$

$$\begin{bmatrix} \Delta\theta_2 \\ \Delta\theta_3 \end{bmatrix} = \begin{bmatrix} \hat{\theta}_2 - \theta_2 \\ \theta_3 - \theta_3 \end{bmatrix} \tag{11}$$

The effect on the relativistic drift angle determination is

$$\begin{bmatrix} \Delta\psi_2 \\ \Delta\psi_3 \end{bmatrix} = \begin{bmatrix} \Delta\theta_2 - \Delta\alpha_2 \\ \Delta\alpha_3 - \Delta\theta_3 \end{bmatrix} \tag{12}$$

This completes the present development of the measurement model for the GP-B experiment. The model will be used for investigating the quantitative effect of various misalignments on the experiment data.

References

1. H. Goldstein, Classical Mechanis, Addison-Wesley, Reading, Mass., 1950.

2. C. W. F. Everitt, Final Report on NASA Grant 05-020-019, Stanford University, Hansen Laboratory, July 1977.