Memoirs of the School of Engineering, Okayama University, Vol. 19, No. 1, November 1984

Error Analysis of Band Matrix Method

Takeo TANIGUCHI*and Akira SOGA*

(Received September 17, 1984)

SYNOPSIS

Numerical error in the solution of the band matrix method based on the elimination method in single precision is investigated theoretically and experimentally, and the behaviour of the truncation error and the round-off error is clarified. Some important suggestions for the useful application of the band solver are proposed by using the results of above error analysis.

1. INTRODUCTION

The application of the finite element or the finite difference method to engineering problems generally causes the appearance of a large sparse set of linear equations. For this sparse set we find some effective solvers, for example, the band matrix method, the profile method and the wavefront method, and all of them are generally based on the elimination method. On the other hand it is well known that the reliability of the solution by the elimination method becomes low with the increase of the number of variables, and actually systems with several thousand variables are often solved in the engineering field.

The study on the numerical error appearing at the application of the elimination method was firstly began by Wilkinson[1], and excellent investigation was presented by Roy[2]. According to Roy's results it is recommended that the solution procedure for a large sparse set of linear equations should be done in double precision.

---

* Department of Civil Engineering

Among the solvers described above the band solver is preferably used for medium-size problems in the field of civil engineering by the reasons of the memory size neccessary for the solver and of the simplicity of the programming and its data structure. Therefore, for the user the behaviour and the tendency of the numerical error at the application of the band solver is interesting and important.

It is already known that the number of correct decimal places in the solution is estimated by using the conditioning number which is the ratio of the maximum and minimum eigenvalues. But, since the calculation of these eigenvalues requires a lot of numerical operations, for general purposes it is difficult to calculate exact conditioning number from the economical viewpoint.

The purpose of this investigation is to show the degree of the reliability of solution by the band matrix method in single precision applied to general civil engineering problems. For this aim a number of numerical experiments are done in order to obtain numerical errors appearing through the elimination method, and the data are theoretically analyzed and we clarify the appearance and the growth of numerical error. In the last section some proposals for the main purpose of this investigation are given by using above results of error analysis.

2. NUMERICAL ERROR

Let

$$Ax = b \tag{1}$$

be a set of linear equations in which A is a symmetric, sparse and positive-definite coefficient matrix, and x and b are unknown and known vectors, respectively. In this section we consider on the numerical error appearing in the solution x from the theoretical viewpoint.

We can define following three different sources of numerical error which appear in the computed solution by digital computer.

(1). Error included in original data.

The error or the uncertainty included in original data introduces the numerical error in the solution vector x. Assume that the matrix A subjects to uncertainty dA, and that resulting uncertainty dx appears in the solution x. That is,

$$(A + dA)(x + dx) = b \tag{2}$$

By taking norms of both sides of eq.2 and assuming $||dA||/||A|| << 1$

we obtain following relation;

$$\frac{||dx||}{||x||} \leqslant \mathrm{cond}(A)\ \frac{||dA||}{||A||} \tag{3}$$

in which cond(A) is the conditioning number of A and it is defined by using the maximum and minimum eigenvalues of A.

$$\mathrm{cond}(A) = \lambda_{max}/\lambda_{min} \tag{4}$$

Note that cond(A) becomes large if the ratio $\lambda_{max}/\lambda_{min}$ becomes large. Thus, eq.3 suggests that the uncertainty of the solution is governed by cond(A) and the uncertainty in original data. Generally, it can be said that cond(A) becomes large in accordance with the increase of the dimension of A. Therefore, for a large system of eq.1 the number of places of all data in A must be increased in order to keep the accuracy of the solution. In this paper we assume that the original data have no uncertainty.

(2) Rounding error.

As far as we use any digital computer, any numeral must be expressed by finite number of decimal places, and, therefore, most of resulting data treated in the computer become approximate values of original ones. That is, the digital computation is not exact but approximate. Since the rounding of data occures at any stage of the computation, we find that the resulting solution subjects to the rounding from the last decimal place in accordance with the increasing of the numerical operations. This suggests that the rounding error becomes so large as the dimension of A becomes large.

For the prevention of early rounding of data more decimal places are required, and we can compute in double precision. Since the rounding error is introduced at the expression or operation of data in finite decimal places, the method of rounding becomes important and we find two kind of rounding methods; chopped operations and rounded operations. Probabilistically it is expected that the latter gives accurate solution comparing to the former.

Rounding error is easily found by comparing the solutions of single and double precision computations. Note that the rounding error in the solution does not depend on the physical properties of the problems.

(3) Truncation error.

Assume that the dimension of A in eq.1 is sufficiently large. Then, cond(A) defined in eq.4 has a large number, and the information of some eigenvalues from the smallest one are neccessarily lost from the solution

as far as finite and constant word length is used. That is, the truncation error is governed by the physical properties of the problem.

An upper bound for the truncation error on the solution is easily estimated using eq's 1 and 2. Assume that dA in eq.2 represents the truncation error due to the restriction of each datum in A to the constant word length of the computer. Then, according to Roy's study

$$\frac{||dA||}{||A||} = 10^{-p} \tag{5}$$

, where p is number of decimal places in which any datum is stored. Thus, the number of correct decimal places in the solution, s, is obtained as following[2];

$$s \geq p - \log[\ \mathrm{cond}(A)\ ] \tag{6}$$

Eq.6 suggests that the reliable number of decimal places in the solution depends on the conditioning number of A. We must note that through the derivation of eq.6 the truncation error of the right-hand side of eq.1 is not taken into consideration. Actually, the effect has the same magnitude of the truncation in the left-hand side. However, since eq.6 overestimates the effect of the truncation error, we may apply eq.6 for estimating the reliable number of decimal places in the solution[2].

Rozanoff's study suggests that the iterative improvement of the solution subjecting the truncation can't recover the truncation error[2]. Thus, the complete numerical operations must be done in double precision except the in-put data. If we use single precision input and problem solver in double precision, the number of correct decimal places in the solution, s, is given as followings;

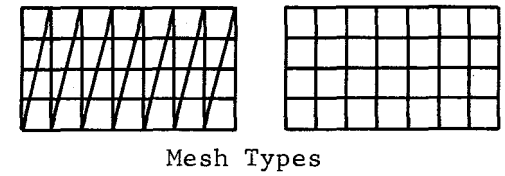
$$\left.\begin{array}{l} \text{if } \log[\ \mathrm{cond}(A)\ ] \leq p \text{ then } s = p \\ \text{if } p < \log[\ \mathrm{cond}(A)\ ] < 2p \text{ then } s \geq 2p - \log[\mathrm{cond}(A)] \end{array}\right\} \tag{7}$$

Note that eq's 6 and 7 do not include the effect of round-off error appearing at the numerical operations.

If a problem is given, its conditioning number is automatically determined. Thus, the effect of the truncation error to the solution is a priori estimated if cond(A) is obtained. On the contrary, the effect of the round-off error is determined by the procedure to obtain the solution. That is, the error according to rounding is different for each matrix solver, and, therefore, we examine the effect of round-off error appearing at the application of the band matrix method based on the elimination.

3. NUMERICAL TESTS AND RESULTS

3.1 Preliminaries



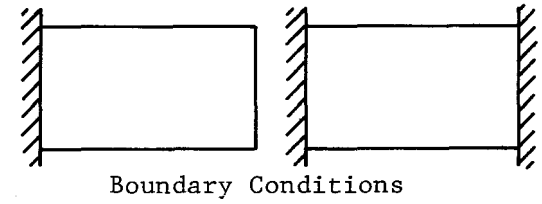


Fig. 1 Models for Numerical Experiments

(1) Model

The models are illustrated in Fig.1. The purpose of this investigation is the survey of the numerical error in the solution of the band matrix method based on the elimination method, and, therefore, two types of boundary conditions which may generate big numerical error are used for these studies(see Fig.1). The coefficient matrix is generated by using the topology of the meshes shown in Fig.1, and non-zero entry in the off-diagonal element is equated to be "-1" and the main diagonal, a_{ii}, is equal to $\Sigma|a_{ij}|$ for j=1 to n. In actual engineering problem this type of matrices appears when the problem region in two dimensional space is subdivided into regular finite elements or it is expressed by using regular finite difference method.

We assume that all nodes in the problem area subject to unit load.

(2) Definition of numerical error

As the numerical error we calculate the maximum relative error expressed as following;

$$\text{error} = \max_{i=1}^{n} \frac{|x_i(D) - x_i(S)|}{|x_i(D)|} \tag{8}$$

in which $x_i(D)$ and $x_i(S)$ are the i-th element of the solution vectors calculated in double and single precision, respectively. In eq.8 the solution in double precision is treated as a strict one, and the propriety of this assumption was proved by numerical experiments for large-size problems. That is, it was proved that the solutions of problems with more than 8000 unknowns obtained by different elimination orderings coincided each other till more decimal places than the decimal places used in single precision calculation.

(3) Floating-point arithmetic operation

It is well known that the magnitude of the numerical error depends on the floating-point arithmetic operation. As the operation we may use "rounded operations" or "chopped operations". According to Forsythe and Moler[3] the rounded operations give better result comparing to the chopped one. In order to prove their result numerical experiments are done and the results are summarized in Fig.2. The difference of numerical errors of these two methods is quite large, and we use "rounded operation" for our numerical experiments.

The machine used for the numerical experiments in this paper is ACOS 1000 Model 20, whose effective decimal places in single and double precisions are 8.12 and 17.8, respectively.

3.2 Numerical Experiment Method

The models shown in Fig.1 are by using the band matrix method in double and single precisions with rounded operations, and by using eq.8 the errors are obtained. The variables in the numerical model are

(a) total number of nodal points, n,
(b) the width, b,
(c) the length, a,
(d) number of boundary nodes,
(e) mesh type,
(f) boundary condition,
(g) elimination order.

A number of numerical experiments are done by changing the combination of above variables. The results and the consideration for the results are summarized in successive section.

3.3 Results of Experiments

The main purpose of the numerical experiments is to know the behaviour of

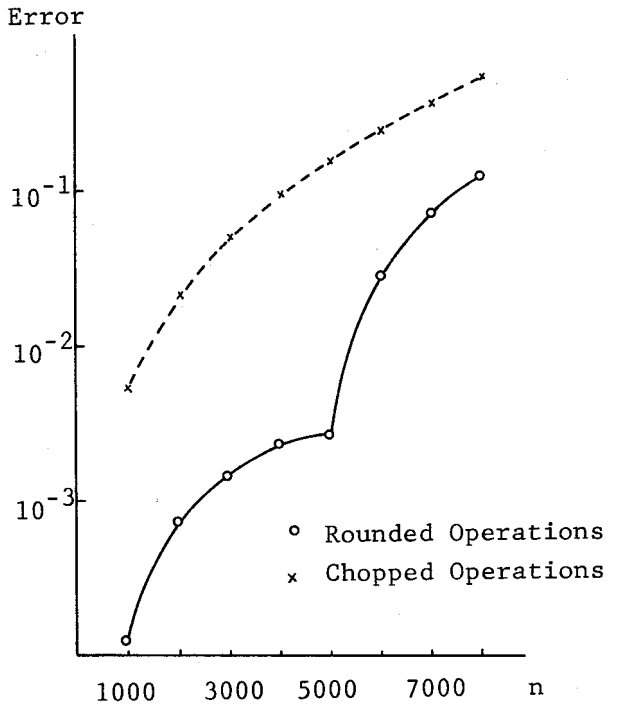


Fig.2 Relative Error in Single Precision

the truncation error and the round-off error which are neccessarily observed as numerical error. Therefore, the numerical results are examined from this aspect.

Fig.3 shows the relation between the correct decimal places and the total number of nodes for different width, b. The same behaviour is observed in Fig.4 which represents the relation between the error of the main diagonals and n, and we remark that in accordance with the increase of the number of nodes the change of the values of last diagonal entries becomes slight and, finally, the successive forward eliminations can't treat significant operations and produce only

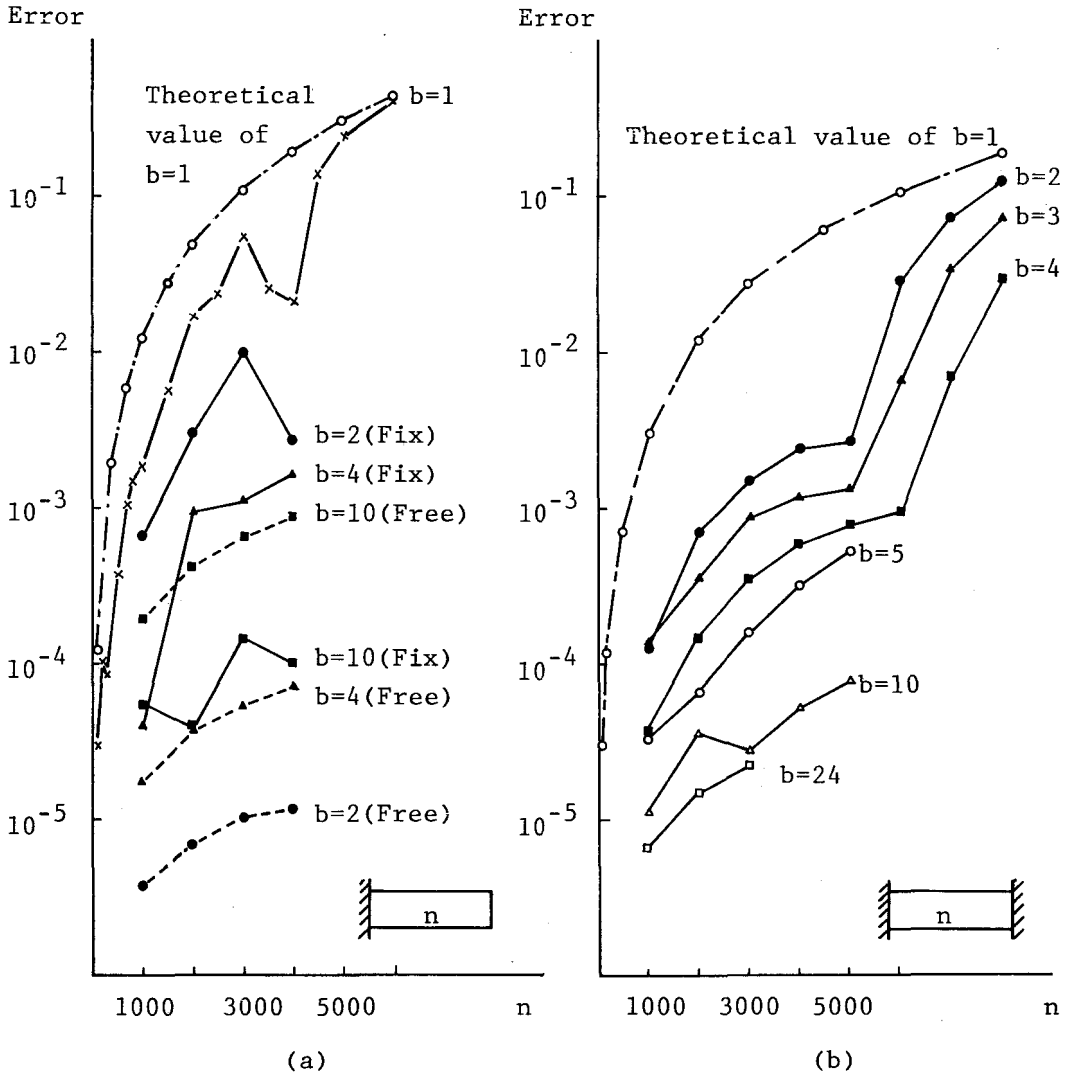


Fig.3 Correct Decimal Places for Different "b"

round-off errors. Therefore, the phenomenon in Fig.1 is recognized as the reflection of insignificant forward elimination calculation. This suggests that in worst case the calculation in single precision may become insignificant for $n > 4000$. Then, we obtain following result;

[Result 1]

The calculation in single precision may be applied for systems with $n < 4000$.

We denote this critical number of nodes as N_{cr}. The conditioning number of a cantilever type structure with n nodes is theoretically obtained by following equation;

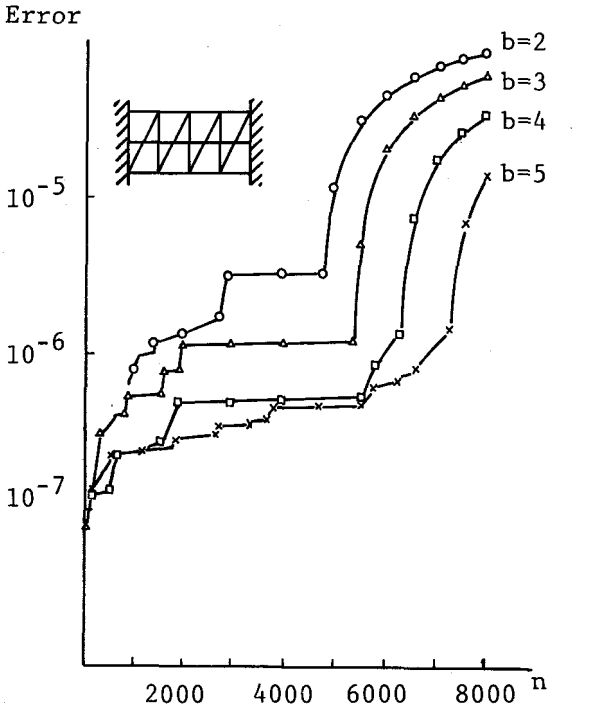


Fig.4 Numerical Error of Main Diagonal

$$\mathrm{cond}(A) = \lambda_{max}/\lambda_{min}$$

in which λ_{max} and λ_{min} are eigenvalues and they are obtained as

$$\lambda_i = 2\left(1 - \cos\frac{2i - 1}{2n + 1}\pi\right), \quad i = 1, 2, \ldots, n \tag{10}$$

The correct decimal places obtained from eq's 6 and 9 is presented in Fig.1.

Since the calculation of eigenvalues for large systems is very troublesome, we approximate cond(A) as

$$\mathrm{cond}(A) = \frac{\max \mathrm{PIVOT}}{\min \mathrm{PIVOT}} \tag{11}$$

, where PIVOT is the diagonal values in the upper triangular matrix which is obtained after the forward elimination. The responsible decimal places obtained by using eq's 11 and 6 are represented in Fig. 5, and from these curves we can remark that even though eq.11 always overestimates the correct decimal places, the characteristics of the curves seem to be very similar to the ones from eq.9. Therefore, in our consideration we use only this characteristics.

Comparing the curves shown in Fig's 1 and 3, we can remark that for

$n < N_{cr}$ the observed correct decimal places have the same or similar tendency of the theoretical ones by eq's 9 and 11. Since the truncation error is determined by cond(A), we conclude as following;

[Result 2]

For $n < N_{cr}$ the behaviour of the numerical error is almost governed by the truncation error.

From the characteristics of the function of cond(A), i.e.eq.9, the curve must be smooth. But, the curve showing the actual correct decimal places is not smooth for $n < N_{cr}$. Thus,

[Result 3]

For $n < N_{cr}$ the round-off error afects the correct decimal places a little.

From the results in Fig's 1 and 3 it is recognized that cond(A) decreases in accordance with the increasing of "b" for constant "n". That is, for constant n (=a*b)

cond(A) increases if b decreases, and

cond(A) decreases if b increases.

Then, we obtain another result;

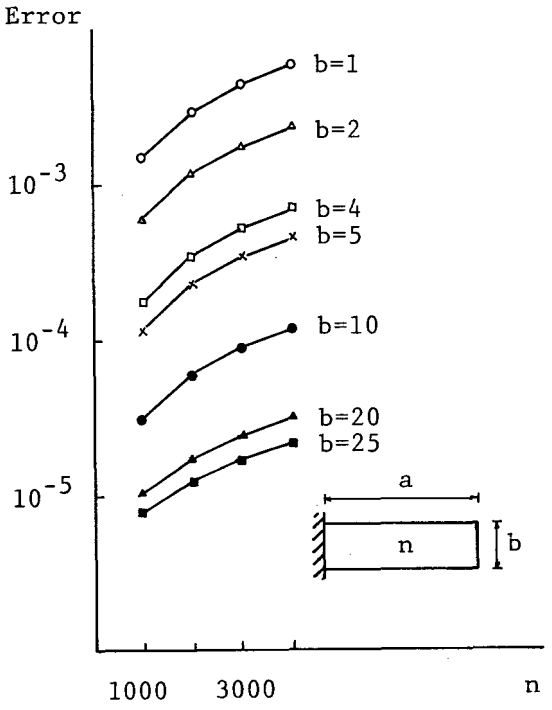


Fig.5 Correct Decimal Places obtained by Eq's 6 and 11.

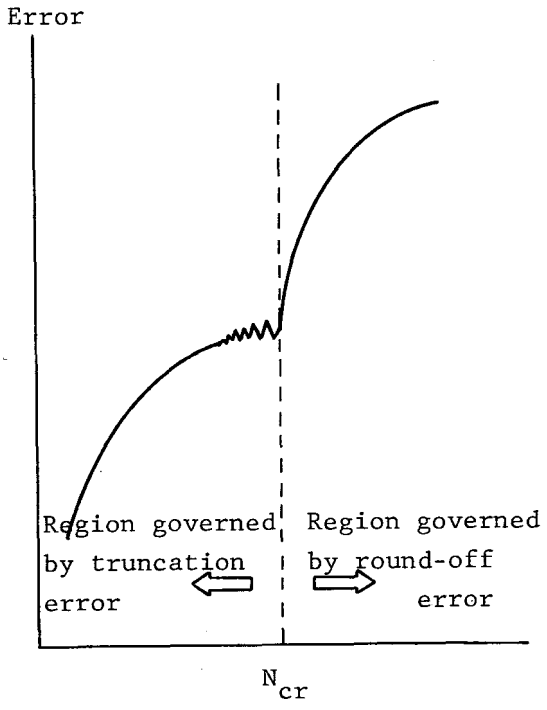


Fig.6 Model of Numerical Error

[Result 4]

The critical number of nodes, N_{cr}, in single precision calculation depends on the value of cond(A), and the value of N_{cr} increases for the decreasing of cond(A).

Summarization of results leads to the representation of the modes of numerical error as shown in Fig.6.

The effect of the round-off error on the truncation error is observed in the results of other numerical experiments which are illustrated in Fig.7. This figure suggests that if b is increases for constant n (=a*b), cond(A) decreases and, therefore, the magnitude of truncation error becomes small. At the same time, since the numerical operations for the band matrix method ($\propto nb^2$) increases according to the increasing of b, the magnitude of round-off error becomes relatively large. Then,

[Result 5]

In accordance with the decreasing of cond(A) the magnitude of round-off error becomes large comparing to the one of truncation error. This explains the oscillation of the curve representing the correct decimal places.

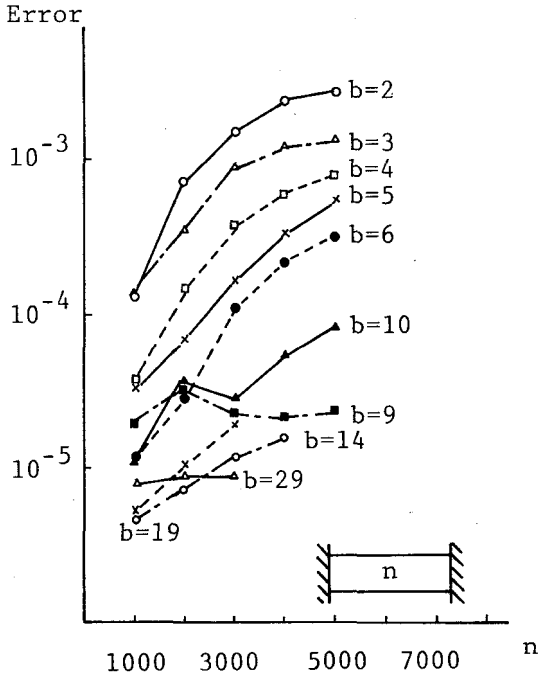


Fig.7 Numerical Error according to the width b.

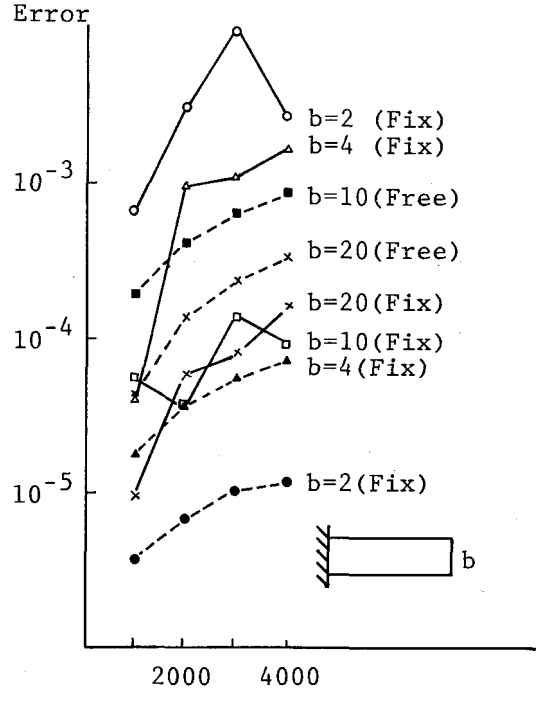


Fig.8 Numerical Error according to Elimination Orderings

Now, we show another result of numerical experiments. Fig.8 indicates that the numerical error depends on the elimination ordering of any system. Since a sparse system has only one conditioning number, the difference of numerical error in Fig.8 is not explained by truncation error but by round-off error. This leads to following result;

[Result 6]

The difference of numerical errors according to the different elimination orderings is caused by round-off error.

[Result 7]

The width of difference of round-off errors for a system (n=a*b) becomes large in accordance with the decreasing of b.

In above consideration we assumed that round-off error depends on the number of numerical operations. Actually, since the numerical operation is done by using rounded operation, the appearance of the round-off error is not deterministic but probabilistic. Then, we obtain

[Result 8]

The magnitude of actual round-off error is small comparing to the one which is deterministically estimated by using the total number of numerical operations.

By summarizing above results we may propose two important models representing the influence of truncation error and round-off error (see Fig's 9 and 10).

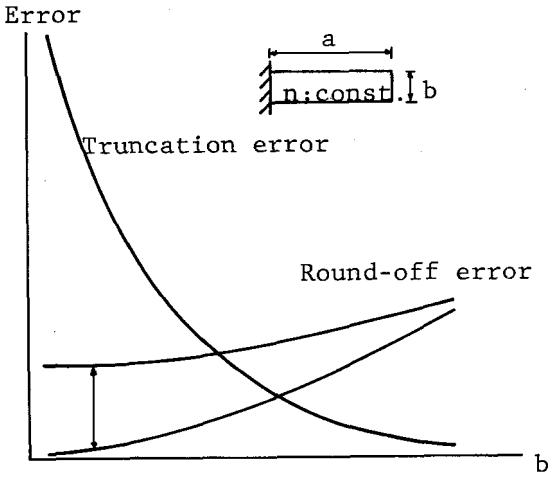


Fig.9 Magnitude of Truncation Error and Round-off Error for Constant "n"

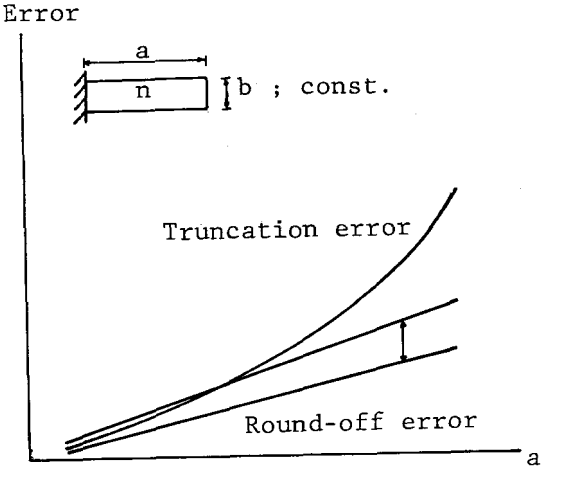


Fig.10 Magnitude of Truncation Error and Round-off Error for Constant "b"

4. CONCLUDING REMARKS

The results obtained in this investigation are summarized as follows.

(1) The behaviour of truncation error and round-off error is clarified.

(2) By using only truncation error the upper bound of the correct decimal places may be estimated, but generally it is overestimated as indicated in ref.2.

(3) The existence of the round-off error becomes clear when a system has large n and small cond(A).

(4) The arithmetic operation in single precision becomes insignificant for a certain number of unknowns, and the number is governed by the value of cond(A).

Now, we consider on the application of the band solver to problems in civil engineering field by using above results of numerical error analysis.

If we require, at least, three correct decimal places for solution, the band solver in single precision may be applied to systems which include less than 1000 unknowns without any cares.But, if it is applied for systems with more variables, then the user should pay attention to the elimination ordering, and it may be applied for problems with more than 3000 unknowns.

REFERENCES

(1) J.H. Wilkinson,'Rounding Errors in Algebraic Processes', Her Britannic Majesty's Stationary Office (1963)

(2) J.R. Roy,"Numerical Error in Structural Solutions", Journal of Structural Division, ASCE (1971), ST4, pp.1039-1054

(3) G.E. Forsythe and C.B. Moler,'Computer Solution of Linear Algebraic Systems', Prentice-Hall, Inc.(1967)