Mathematica Balkanica

Mathematical Society of South-Eastern Europe
A quarterly published by
the Bulgarian Academy of Sciences – National Committee for Mathematics

---

Mathematica Balkanica

New Series Vol. 6, 1992, Fasc. 4

Round-off Error Analysis of the Bordering Method

Plamen Y. Yalamov

Presented by P. Kenderov

An application of the new method of round-off error analysis proposed by Voevodin and Yalamov [3] is discussed. The basic instrument of the analysis is the graph of the algorithm. On this base one step of the bordering method is studied. It is shown that backward analysis of this method is possible.

1. Introduction

A new method of round-off error estimation is proposed by V. K. Voevodin and P. Y. Yalamov [3]. This method is deeply connected with the graph and the parallel structure of the algorithm. The notion of equivalent perturbation is defined for every piece of data in contrast to the generally used backward analysis where the equivalent perturbations are introduced only for the input data. This fact allows us to investigate the error propagation. Under these assumptions V. V. Voevodin and P. Y. Yalamov [3] derived a linear system of equations with respect to the equivalent perturbations:

$$B\varepsilon = \eta, \tag{1}$$

where η is the vector of absolute errors of all the operations and matrix B consists of the Frechet-derivatives of all the vector operations and of -1 and 0 (see [3]). Now, if we give some values to the equivalent perturbations of the output data, then we can obtain approximately the equivalent perturbations of the input data by solving system (1). The structure of matrix B is deeply connected with the parallel structure of the algorithm.

2. Round-off Error Analysis of the Bordering Method

This method is suitable for Ritz and Bubnov-Galerkin systems when it is necessary to make more precise an already found solution. If only one coordinate function is needed for the more precise solution then the new system of linear equations is obtained from the previous one by bordering. For this reason we consider in this section one step of the bordering method. Besides, some methods

for solving linear systems with Toeplitz and Hankel matrices are based on this method (see [2]).

Let us have $A_N = A$. Assume that matrix A_N is the result of bordering of matrix A_{N-1} of order $N-1$, and the inverse A_{N-1}^{-1} of A_{N-1} is known. Let us also have

$$A = A_N = \left[\begin{array}{ccc|c} a_{11} & a_{12} \ldots\ldots & a_{1,N-1} & a_{1N} \\ a_{21} & a_{22} \ldots\ldots & a_{2,N-1} & a_{2N} \\ \ldots & \ldots\ldots\ldots & & \ldots\ldots \\ a_{N-1,1} & a_{N-1,2} \ldots\ldots & a_{N-1,N-1} & a_{N-1,N} \\ \hline a_{N1} & a_{N2} \ldots\ldots & a_{N,N-1} & a_{NN} \end{array}\right] = \begin{bmatrix} A_{N-1} & u_N \\ v_N & a_{NN} \end{bmatrix}.$$

Here A_{N-1} denotes the above mentioned matrix of order $N-1$, and $v_N = (a_{N1}, \ldots, a_{N,N-1})$, $u_N = (a_{1N}, \ldots, a_{N-1,N})^T$. We look for a matrix A_N^{-1} of the following kind:

$$D_N = A_N^{-1} = \begin{bmatrix} P_{N-1} & r_N \\ q_N & \dfrac{1}{\alpha_N} \end{bmatrix},$$

where P_{N-1} denotes a square matrix of order $N-1$, r_N denotes a column, q_N denotes a row and α_N is a number to be defined. From the equality $AA^{-1} = I$, where I is the corresponding unity matrix, it is not difficult to derive the expressions defining P_{N-1}, r_N, q_N, α_N (see [1]). So, we have

$$P_{N-1} = A_{N-1}^{-1} + \frac{A_{N-1}^{-1} u_N v_N A_{N-1}^{-1}}{\alpha_N},$$

$$r_N = -\frac{A_{N-1}^{-1}}{\alpha_N},$$

$$q_N = -\frac{v_N A_{N-1}^{-1}}{\alpha_N},$$

$$\alpha_N = a_{NN} - v_N A_{N-1}^{-1} u_N. \tag{2}$$

These equalities are the base for matrix inversion by means of sequential bordering. The inverses of the following matrices are computed successively:

$$[a_{11}], \quad \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \ \ldots .$$

Each macrostep is based on (2). The algorithm of every macrostep can be written as follows:

1. Compute the column $\beta_N = -A_{N-1}^{-1} u_N$ with components $\beta_{1N}, \ldots, \beta_{N-1,N}$.
2. Compute the row $\gamma_N = -v_N A_{N-1}^{-1}$ with components $\gamma_{N1}, \ldots, \gamma_{N,N-1}$.
3. Compute the number

$$\alpha_N = a_{NN} + \sum_{i=1}^{N-1} a_{Ni}\beta_{iN} = a_{NN} + \sum_{i=1}^{N-1} a_{iN}\gamma_{Ni}.$$

4. Compute the elements d'_{ik} of the inverse D_N as follows:

4.1. Compute matrix P_{N-1} with elements

$$d'_{ik} = d_{ik} + \frac{\beta_{iN}\gamma_{Nk}}{\alpha_N}, \quad i, k = 1, \ldots, N-1.$$

4.2. Compute r_N with components

$$d'_{iN} = \frac{\beta_{iN}}{\alpha_N}, \quad i = 1, \ldots, N-1.$$

4.3. Compute q_N with components

$$d'_{Nk} = \frac{\gamma_{Nk}}{\alpha_N}, \quad k = 1, \ldots, N-1.$$

4.4. Compute $d'_{NN} = \frac{1}{\alpha_N}$.

Here d'_{ik} are the elements of A_{N-1}^{-1}. It is very difficult to construct the graph of this algorithm from the Fortran-like language as it was done by V. V. Voevodin and P. Y. Yalamov in [3]. Therefore, let us consider the macrograph in which each vertex corresponds to one of the macrooperations 1-4.4. This macrograph is given in Fig. 1. Some of the arcs carry vectors and matrices depending on the result in every vertex. It is clear that input and intermediate data is multiplied. As far as there are no long paths in this graph we shall try to do backward analysis.

Further on the subscript of ε means that this is the equivalent perturbation of the corresponding matrix, vector or number. For example, $\varepsilon_{A_{N-1}^{-1}}$ is the matrix of equivalent perturbations of A_{N-1}^{-1}, ε_{u_N} is the vector of equivalent perturbations of u_N. Analogously, the subscript of η shows that this is the local absolute round-off error of the corresponding matrix, vector or number. Besides, if v_N is a row then ε_{v_N} and η_{v_N} are also rows. Let us assume that $\varepsilon_{P_N} = 0, \varepsilon_{r_N} = \varepsilon_{q_N} = 0, \varepsilon_{d_{NN}} = 0$. Then system (1) in this case looks like this:

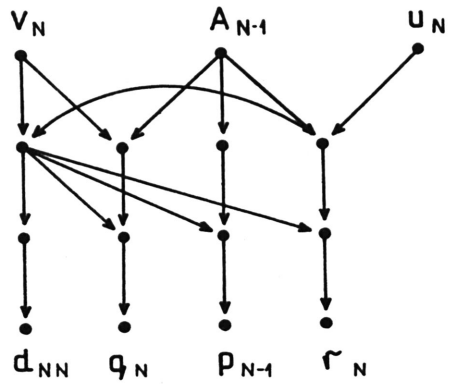

Figure 1. The macrograph of the bordering method

$$
\begin{array}{llll}
-\varepsilon_{A_{N-1}^{-1}} u_N - A_{N-1}^{-1}\varepsilon_{u_N} & -\varepsilon_{\beta_N} & & = \eta_{\beta_N} \\
-v_N \varepsilon_{A_{N-1}^{-1}} & -\varepsilon_{v_N} A_{N-1}^{-1} & -\varepsilon_{\gamma_N} & = \eta_{\gamma_N} \\
 & \varepsilon_{v_N}\beta_N + v_N\varepsilon_{\beta_N} + \varepsilon_{\alpha_{NN}} & -\varepsilon_{\alpha_N} & = \eta_{\alpha_N} \\
\varepsilon_{A_{N-1}^{-1}} & +\dfrac{1}{\alpha_N}\varepsilon_{\beta_N}\gamma_N & +\dfrac{1}{\alpha_N}\beta_N\varepsilon_{\gamma_N} - \dfrac{1}{\alpha_N^2}\beta_N\gamma_N\varepsilon_{\alpha_N} & = \eta_{P_N} \\
 & \dfrac{1}{\alpha_N}\varepsilon_{\beta_N} & -\dfrac{1}{\alpha_N^2}\beta_N\varepsilon_{\alpha_N} & = \eta_{r_N} \\
 & & \dfrac{1}{\alpha_N}\varepsilon_{\gamma_N} - \dfrac{1}{\alpha_N^2}\gamma_N\varepsilon_{\alpha_N} & = \eta_{q_N} \\
 & & -\dfrac{1}{\alpha_N^2}\varepsilon_{\alpha_N} & = \eta_{d_{NN}}
\end{array} \tag{3}
$$

From the last four equations we define ε_{α_N}, ε_{γ_N}, ε_{β_N}, $\varepsilon_{A_{N-1}^{-1}}$ by back substitution as follows:

$$
\begin{aligned}
&\varepsilon_{\alpha_N} = -\alpha_N^2 \eta_{d_{NN}}, \\
&\varepsilon_{\gamma_N} = \alpha_N \eta_{q_N} + \frac{1}{\alpha_N}\gamma_N \varepsilon_{\alpha_N} = \alpha_N \eta_{q_N} - \alpha_N \eta_{d_{NN}} \gamma_N, \\
&\varepsilon_{\beta_N} = \alpha_N \eta_{r_N} + \frac{1}{\alpha_N}\beta_N \varepsilon_{\alpha_N} = \alpha_N \eta_{r_N} - \alpha_N \eta_{d_{NN}} \beta_N, \\
&\varepsilon_{A_{N-1}^{-1}} = \eta_{P_N} - \eta_{r_N}\gamma_N - \beta_N \eta_{q_N} + 3\eta_{d_{NN}}\beta_N\gamma_N.
\end{aligned} \tag{4}
$$

For every element $(\varepsilon_{A_{N-1}^{-1}})_{ik}$ of matrix $\varepsilon_{A_{N-1}^{-1}}$ we have:

$$(5)\qquad (\varepsilon_{A_{N-1}^{-1}})_{ik}=(\eta_{P_N})_{ik}-(\eta_{r_N})_i\gamma_{Nk}-\beta_{iN}(\eta_{q_N})_k+3\eta_{d_{NN}}\beta_{iN}\gamma_{Nk}.$$

Further on assume that the operations are realized with accumulation (see [4]) where it is possible. Round-off error computation gives

$$\tilde{d}'_{ik}=[d_{ik}+\frac{1}{\tilde{\alpha}_N}\tilde{\beta}_{iN}\tilde{\gamma}_{Nk}(1+\rho_{ik}^{(1)})](1+\rho_{ik}^{(2)}),\ |\rho_{ik}^{(s)}|\leqq p^{-t+1},\ s=1,2,$$

where p is the radix and t is the number of digits. Hence, neglecting terms of second order in p^{-t+1} the following equality is valid:

$$(6)\quad (\eta_{P_N})_{ik}=d_{ik}\rho_{ik}^{(2)}+\frac{1}{\tilde{\alpha}_N}\tilde{\beta}_{iN}\tilde{\gamma}_{Nk}(\rho_{ik}^{(1)}+\rho_{ik}^{(2)}),\ |\rho_{ik}^{(s)}|\leqq p^{-t+1},\quad s=1,2.$$

From 4.1 we have

$$(7)\qquad \frac{1}{\alpha_N}\beta_{iN}\gamma_{Nk}=d'_{ik}-d_{ik}.$$

Then from (6), neglecting terms of second order in p^{-t+1} again, it follows that

$$(8)\qquad |(\eta_{P_N})_{ik}|\leqq(1.5\,|d_{ik}|+|d'_{ik}|)p^{-t+1}.$$

Now, we turn to $(\eta_{r_N})_i$. We have

$$\tilde{d}'_{iN}=\frac{1}{\tilde{\alpha}_N}\tilde{\beta}_{iN}(1+\sigma_i),\quad |\sigma_i|\leqq 0.5p^{-t+1},$$

hence

$$(9)\qquad (\eta_{r_N})_i=\frac{1}{\alpha_N}\beta_{iN}\sigma_i,\quad |\sigma_i|\leqq 0.5p^{-t+1}.$$

Analogously we obtain that

$$(10)\qquad (\eta_{q_N})_k=\frac{1}{\alpha_N}\gamma_{Nk}\tau_k,\quad |\tau_k|\leqq 0.5p^{-t+1}.$$

At last we have

$$(11)\qquad \eta_{d_{NN}}=\frac{1}{\alpha_N}\xi,\quad |\xi|\leqq 0.5p^{-t+1}.$$

From (5), (7)–(11) we have the following estimates:

$$(12)\qquad |\varepsilon_{A_{N-1}^{-1}})_{ik}|\leqq(4\,|d_{ik}|+3.5\,|d'_{ik}|)p^{-t+1}.$$

If the range of the results is not so wide these estimates are acceptable.

Further on by $\|\cdot\|$ we mean one of the following norms $\|\cdot\|_1$, $\|\cdot\|_\infty$. From (12) using these norms we have

$$\|\varepsilon A_{N-1}^{-1}\| \leqq (4\|A_{N-1}^{-1}\| + 3.5\|P_N\|)p^{-t+1}. \tag{13}$$

Thus we have estimated the equivalent perturbations of one part of the input data, namely A_{N-1}^{-1}. Now let us estimate the equivalent perturbations of the other part of the input data $-$ u_N, v_N and a_{NN}. From the first three equations of (3) we can define $\varepsilon_{u_N}, \varepsilon_{v_N}, \varepsilon_{a_{NN}}$. We have

$$\begin{aligned} \varepsilon_{v_N} &= -(\eta_{\gamma_N} + \varepsilon_{\gamma_N} + v_N \varepsilon_{A_{N-1}^{-1}}) A_{N-1}^{-1} \\ \varepsilon_{u_N} &= -A_{N-1}^{-1}(\eta_{\beta_N} + \varepsilon_{\beta_N} + \varepsilon_{A_{N-1}^{-1}} u_N), \\ \varepsilon_{a_{NN}} &= \eta_{\alpha_N} + \varepsilon_{\alpha_N} - \varepsilon_{v_N}\beta_N - v_N\varepsilon_{\beta_N}. \end{aligned} \tag{14}$$

These equivalent perturbations are estimated with the help of the norms. Let us start with ε_{v_N}. For this purpose we have to estimate the norms of $\eta_{\gamma_N}, \varepsilon_{\gamma_N}$ at first. We have

$$(\eta_{\gamma_N})_k = (v_N A_{N-1}^{-1})_k \rho_k, \quad |\rho_k| \leqq 0.5p^{-t+1}.$$

Then

$$\|\eta_{\gamma_N}\| \leqq 0.5\|A_{N-1}^{-1}\|\ \|v_N\| p^{-t+1}. \tag{15}$$

Analogously, from (6), (10) and (11) we have

$$\|\varepsilon_{\gamma_N}\| \leqq \|A_{N-1}^{-1}\|\ \|v_N\| p^{-t+1}. \tag{16}$$

About ε_{v_N} from (13)$-$(16) it follows that

$$\|\varepsilon_{v_N}\| \leqq (5.5\varkappa_{N-1} + 3.5\|P_N\|\ \|A_{N-1}\|)\|v_N\| p^{-t+1}. \tag{17}$$

Here $\varkappa_{N-1}$ is the condition number of A_{N-1}, i. e. $\varkappa_{N-1} = \|A_{N-1}^{-1}\|\ \|A_{N-1}\|$. As far as P_N and A_{N-1} are principal submatrices of A_{N-1}^{-1} and A_N correspondingly we have that

$$\|P_N\|\ \|A_{N-1}\| \leqq \|A_{N-1}^{-1}\|\ \|A_{N-1}\| = \varkappa_N.$$

Then (17) can be rewritten as follows:

$$\|\varepsilon_{v_N}\| \leqq (5.5\varkappa_{N-1} + 3.5\varkappa_N)\|v_N\| p^{-t+1}, \tag{18}$$

and the relative estimates look like this:

$$\frac{\|\varepsilon_{v_N}\|}{\|v_N\|} \leqq (5.5\varkappa_{N-1} + 3.5\varkappa_N) p^{-t+1}.$$

It is seen that the last estimates depend only on the condition of A_{N-1}, A_N. Analogously we can obtain the estimates of ε_{u_N}:

$$\|\varepsilon_{u_N}\| \leqq (5.5\varkappa_{N-1} + 3.5\varkappa_N)\ \|u_N\| p^{-t+1},$$

$$\frac{\|\varepsilon_{u_N}\|}{\|u_N\|} \leqq (5.5\varkappa_{N-1} + 3.5\varkappa_N) p^{-t+1}.$$

At last from the inequalities

$$|\eta_{\alpha_N}| \leqq 0.5|\alpha_N| p^{-t+1},\ |\varepsilon_{\alpha_N}| \leqq 0.5|\alpha_N| p^{-t+1},$$

and from (4), (14), (18) we obtain the estimate of $|\varepsilon_{a_{NN}}|$:

$$|\varepsilon_{a_{NN}}| \leqq |\alpha_N| p^{-t+1} + (5.5\varkappa_{N-1} + 3.5\varkappa_N + 1)\ \|A_{N-1}^{-1}\|\ \|u\|\ \|v\| p^{-t+1}.$$

Thus it was shown that backward analysis of one macrostep is possible and the corresponding estimates were obtained. Some other applications of the new method of round-off error analysis are described in [5].

References

1. D. K. Fadeev, V. N. Fadeeva. Numerical methods of linear algebra. Moscow. Fizmatgiz, 1960 (In Russian.)
2. V. V. Voevodin, E. E. Tyrtyshnikov. Computational processes with Toeplitz matrices. Moscow, Nauka, 1987 (In Russian.)
3. V. V. Voevodin, P. Y. Yalamov. A new method of round-off error estimation. – In: Proc. Workshop on Parallel and Distributed Processing, 27-29 March, 1990, Sofia, Bulgaria, Elsevier, Amsterdam, 1990.
4. J. H. Wilkinson. The algebraic eigenvalue problem. Oxford, Clarendon Press, 1965.
5. P. Y. Yalamov. A new method of round-off error analysis. Dissertation, Moscow State University, 1990. (In Russian.)

Department of Mathematics
Technical University
7017 Russe
BULGARIA

Received 07. 02. 1992