Milan Práger; Emil Vitásek

Stability of numerical processes

In: Valter Šeda (ed.): Differential Equations and Their Applications, Proceedings of the Conference held in Bratislava in September 1966. Slovenské pedagogické nakladateľstvo, Bratislava, 1967. Acta Facultatis Rerum Naturalium Universitatis Comenianae. Mathematica, XVII. pp. 315--322.

Persistent URL: http://dml.cz/dmlcz/700202

ACTA FACULTATIS RERUM NATURALIUM UNIVERSITATIS COMENIANAE
MATHEMATICA XVII — 1967

STABILITY OF NUMERICAL PROCESSES

M. PRÁGER and E. VITÁSEK, Praha

The high speed of modern computers and the corresponding employment of ever increasing number of arithmetic operations have brought to the foreground in recent years the importance of the problems of numerical stability. What does it mean numerical stability? It is the sensitivity of a computation on the errors, which are during the computation on the computer necessarily performed. The source of these errors are the round-off errors in single arithmetic operations, substitution of general analytic expressions by rational ones, etc. This problem and the necessity to solve it is now generally accepted. There are different ways of treatment of the problem. In the first place it is the estimation of the accumulated round-off error, which is e.g. done in the well known books of WILKINSON [1] and HENRICI [2, 3], further the method of closure of processes suggested by SOBOLEV [4], recently the techniques of the interval arithmetic by R. MOORE [5] or simply the intuitive use of multiple arithmetic.

In this paper we will consider the problem of the numerical stability for problems of mathematical analysis, particularly for some problems for differential equations. A characteristic feature of such problems is that the arithmetical operations predominate over the logical ones (the influence of the lasts will be therefore neglected in what follows) and that numerical methods for solution of these problems involve always a certain parameter (e.g. the step of the mesh, the number of approximating functions) and the exact solution of the given problem is then obtained by passing to the limit with this parameter.

This paper is a development of problems of numerical stability for initial-value problems for differential equations which was reported by the Authors on the first Equadiff conference [6]. The introducing of the β_s-solution of the sequence of numerical processes will here be essentially new and will be utilized for the study of numerical stability of some other problems. The main results are published in the book [7].

Now we can pass on the exact formulation of concepts and results. The basic concept is the concept of a numerical process. The numerical process (it is possible to say the computing algorithm, too) is a sequence of arithmetic operations, which transforms the set of the initial data in the set of results. It has been just mentioned that in problems of mathematical analysis we have always a sequence of such processes. We introduce consequently.

Definition 1. *Let there be given a sequence of normed vector spaces*

$$X_0^{(j)}, X_1^{(j)}, \ldots, X_{Nj}^{(j)}, \qquad j = 1, 2, \ldots$$

and a sequence of continuous operators

$$A_i^{(j)}, \qquad i = 0, 1, \ldots, N_j - 1; \qquad j = 1, 2, \ldots,$$

which map the Cartesian product $X_0^{(j)} \times \ldots \times X_i^{(j)}$ *into* X_{i+1}.

Then the sequence of equations

$$x_{i+1}^{(j)} = A_i^{(j)}(x_0^{(j)}, \ldots, x_i^{(j)}), \qquad i = 0, 1, \ldots, N_j - 1, \tag{1}$$

where $x_0^{(j)} \in X_0^{(j)}$ *is given and* $x_i^{(j)} \in X_i^{(j)}$ *is called a numerical process.*

The sequence of elements $x_i^{(j)}$ *is called the solution of the numerical process with the initial value* $x_0^{(j)}$.

Thus, by Definition 1 we have introduced the sequence of numerical processes, the results of which converge for $j \to \infty$ (j is the parameter of the sequence) in some or other sense to the exact solution of the given problem. In practical computations, we have obviously $X_i^{(j)} \in R_1$, $i = 1, 2, \ldots$ and $X_0^{(j)} = R_n$, where R_n is the n-dimensional Euclidean space and $A_i^{(j)}$ are the operators of elementary arithmetic operations. However, in order to simplify the study in many cases, it is convenient to introduce more general objects such as vectors, matrices or others.

The numerical process has an algorithmic, i.e., explicit character. For example, Euler's method for the solution of an initial-value problem for the differential equation $y' = f(x, y)$, i.e., the formula

$$y_{n+1} = y_n + hf(x_n, y_n),$$

where y_0 is given, represents a sequence of numerical processes in dependence of the number of subintervals as parameter. On the contrary, the method of finite differences for solution of the boundary-value problem $y'' = f$, $y(0) = y(1) = 0$ i.e.

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f_n$$

with $y(0) = y(1) = 0$ is not a sequence of numerical processes since it is not indicated any method for solution of the obtained system of algebraic equations. If we add that this system will be solved, e.g., by elimination, then we have defined a sequence of numerical processes. In detail, it will be shown later.

Numerical process (in sense of Def. 1) cannot be in any case realized exactly on the computer because of the errors resulting of the finite character of the work of the computer. We introduce therefore

Definition 2. *Let $X_i^{(j)}$ and $A_i^{(j)}$ satisfy assumptions of Definition 1, let $\delta_i^{(j)} \in$ $\in X_i^{(j)}$ and let there be given an initial element $x_0^{(j)}$. Then the sequence of equations*

$$\tilde{x}_{i+1}^{(j)} = A_i^{(j)}(\tilde{x}_0^{(j)}, \ldots, \tilde{x}_i^{(j)}) + \delta_{i+1}^{(j)}, \qquad i = 0, 1, \ldots, N_j - 1 \tag{2}$$

with $\tilde{x}_0^{(j)} = \tilde{x}_0^{(j)} + \delta_0^{(j)}$ will be called the perturbed numerical process (1).

The behaviour of the solution of the numerical process (1) with respect to numerical stability will be tested on the basis of comparison with perturbed numerical processes.

Definition 3. *We shall say that the solution of the numerical process* (1) *corresponding to the initial value $x_0^{(j)}$ is a β_s-solution, if*

$$\limsup_{\Delta \to 0} \frac{1}{\Delta} \sup_{\|\delta_i^{(j)}\| \leq \Delta} \sup_{i=1,\ldots,N_j} \|\tilde{x}_i^{(j)} - x_i^{(j)}\| \leq C_j^s$$

where C is a constant independent of j. We say that the given solution is a B_{s_0}-solution if $s_0 = \inf s$.

The subscript s in the concept of a B_s-solution indicates the character of the stability of the given numerical process. It is useful to note here that in accordance with practical experience the constant C in Definition 3 depends on the type of computer used whereas the constant s is independent on the special type of computer and therefore, it is a universal characterization of the given numerical process. It is quite obvious from Def. 3 that such numerical processes which have B_s-solutions with smallest possible s are most favourable from the point of view of numerical stability.

After this short survey of the general theory of numerical stability we shall pay our attention to concrete examples.

Let us investigate the numerical stability of the method of finite differences for a boundary-value problem for a second-order ordinary differential equation. In this connection, we shall also utilize the method of closure of processes.

Thus, let there be given the differential equation

$$y'' - qy = f$$

with the boundary conditions $y'(a) = \alpha$, $y'(b) = \beta$. All what will be said holds also for general self-adjoint equation and other types of boundary conditions and can be used even for fourth-order equations. For the sake of simplicity, we restrict us to this very simple case. By utilizing of the most simple method of finite differences, we obtain for the unknown approximate values y_n the following system of equations, written in matrix form

(3) $$\begin{pmatrix} 1, & -1, & 0, \dots & \\ -1, & 2+h^2q_1, & -1, \dots & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & \dots, 0, & -1, & 1 \end{pmatrix} \begin{pmatrix} y_0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{pmatrix} = \begin{pmatrix} -\alpha h \\ -h^2f_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta h \end{pmatrix}$$

where $q_i = q(x_i)$ etc, N is the number of subintervals and $h = (b - a)/N$.

We have already mentioned that this system is not a numerical process until we indicate some method for its solution. Thus, we shall connect it here with the elimination method.

This method leads in our case of tridiagonal matrix to three recurrence relations:

(4) $$d_{i+1} = 2 + h^2q_{i+1} - \frac{1}{d_i}, \qquad \text{for } i = 0, 1, \dots, N-2,$$
$$d_0 = 1, \quad d_N = 1 - \frac{1}{d_{N-1}}$$

(5) $$C_{i+1} = -h^2f_{i+1} + \frac{C_i}{d_i}, \qquad \text{for } i = 0, 1, \dots, N-2,$$
$$C_0 = -\alpha h, \quad C_N = \beta h + \frac{C_{N-1}}{d_{N-1}}.$$

The backward substitution for computing the unknown y_n yields

(6) $$y_i = \frac{C_i + y_{i+1}}{d_i}, \quad i = N-1, \dots, 0, \quad y_N = \frac{C_N}{d_N}.$$

Before investigate these reccurence relations, we shall demonstrate their connection with the factorization method by the method of closure of processes. Thus, let there be φ_i and z_i defined by the following relations

(7) $$d_i = 1 + h\varphi_i, \qquad C_i = -hz_i.$$

Then, for the quantities φ_i and z_i we obtain

(8) $$\varphi_{i+1} = \varphi_i + h\left(\frac{-\varphi_i^2}{1 + h\varphi_i} + q_{i+1}\right), \qquad i = 0, 1, \dots, N-2,$$
$$\varphi_0 = 0,$$

(9) $$z_{i+1} = z_i + h\left(-\frac{z_i\varphi_i}{1 + h\varphi_i} + f_{i+1}\right), \qquad i = 0, 1, \dots, N-2,$$
$$z_0 = \alpha,$$

and, after an analogical arrangement, for the y_i

$$
(10)\qquad \begin{aligned} y_i &= y_{i+1} - h\left(\frac{z_i + \varphi_i y_{i+1}}{1 + h\varphi_i}\right), \qquad i = N-1, \ldots, 0,\\ y_N &= \frac{\beta - z_{N-1} + h\beta\varphi_{N-1}}{\varphi_{N-1}}. \end{aligned}
$$

From here it is seen that formulae (8) to (10) represent an approximate method for solution of following initial-value problems

$$
\begin{aligned} &\varphi' + \varphi^2 = q, &&\varphi(a) = 0,\\ &z' + \varphi z = f, &&z(a) = \alpha,\\ &y' - \varphi y = z, &&y(b) = \frac{\beta - z(b)}{\varphi(b)}. \end{aligned}
$$

The last equation is solved from right to left. It can be easily shown that the solution of the last equation is the solution of the original equation with corresponding boundary conditions. Thus we have obtained the so-called factorization method which represents a transformation of the boundary-value problem into numerically stable initial-value problems.

From here it is obvious that the quantities φ_i and z_i are bounded independently of h, that consequently d_i is also bounded and that the quantity C_i is of order h. Analysis of equations (4) to (6) yields then easily that the finite-difference method in connection with the elimination method gives β_2-solution of numerical processes in dependence on the number of subintervals used. Obviously, it is also B_2-solution.

The performed analysis suggests a possibility of a convenient modification of the process of elimination, namely so that we replace the recurrence relation (4) by (8) and then we use the relation (10). By this we obtain for the process even a B_1-solution.

Further, let us consider the stability of the solution of parabolic equation by the method of finite differences. For the sake of simplicity, we shall consider the equation

$$
(11)\qquad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - q(x, t)\, u + f(x, t)
$$

with the initial condition $u(x, 0) = g(x)$, $x \in \langle 0, 1\rangle$ and boundary conditions $u(0, t) = \gamma^0(t)$, $u(1, t) = \gamma^1(t)$, $t \in (0, T)$ and we shall investigate the stability of the Crank-Nicholson formula for the case that the space- and time-steps are related by $\tau = \omega h$, where ω is a constant.

The corresponding system will be written in the matrix form

$$
A^{(l)}u^{(l)} = B^{(l)}u^{(l-1)} + (f^{(l)} + f^{(l-1)}), \qquad l = 1, 2, \ldots, r,
$$

where the tridiagonal matrices $A^{(l)}$ and $B^{(l)}$ are given as follows

$$A^{(l)} = \begin{bmatrix} \frac{2h}{\omega} + 2 + h^2 q_{1,l}, & -1, & 0, \ldots & & \\ -1, & \frac{2h}{\omega} + 2 + h^2 q_{2,l}, & -1, \ldots & & \\ & & \ddots & & \\ & & \ldots, 0, & -1, & \frac{2h}{\omega} + 2 + h^2 q_{n-1,l} \end{bmatrix}$$

$$B^{(l)} = \begin{bmatrix} \frac{2h}{\omega} - 2 - h^2 q_{1,l-1}, & 1, & 0, \ldots & & \\ 1, & \frac{2h}{\omega} - 2 - h^2 q_{2,l-1}, & 1, \ldots & & \\ & & \ddots & & \\ & & \ldots, 0, & 1, & \frac{2h}{\omega} - 2 - h^2 q_{n-1,l-1} \end{bmatrix}$$

the vector $u^{(l)}$ is the vector of the unknown solution at the l-th time level and $f^{(l)}$ is the right-hand side vector

$$f^{(l)} = \{h^2 f_{1,l} + \gamma^0(t_l), h^2 f_{2,l}, \ldots, h^2 f_{n-1,l} + \gamma^1(t_l)\}$$

The numerical process by which $u^{(l)}$ is computed consists in the following recurrence procedure. Assuming that $u^{(l-1)}$ is known we compute

$$v^{(l)} = B^{(l)} u^{(l-1)} + f^{(l)} + f^{(l-1)}$$

and then the equation

(12) $$A^{(l)} u^{(l)} = v^{(l)}$$

is solved. The method for solution of the last system (not yet indicated) is an essential part of the numerical process.

The perturbed process is given by

$$\tilde{u}^{(l)} = B^{(l)} \tilde{u}^{(l-1)} + f^{(l)} + f^{(l-1)} + \delta^{(l)},$$

where $|\delta^{(l)}| < K\delta$, δ is the error of the elementary operation and by the solution of the system

$$A^{(l)} u^{(l)} = \tilde{v}^{l}.$$

We shall assume that the actual solution of the system $u^{(l)}$ fulfils the equation

$$A^{(l)}\tilde{u}^{(l)} = \tilde{v}^{(l)} + r^{(l)}.$$

In order to be able to say something about the stability of the method in question, it is not necessary to specify the method for solution (12) completely. It is sufficient to make certain assumptions about the magnitude of the residue $r^{(l)}$. We have the theorem

The numerical process of solution of the equation (11) *by the Crank-Nicholson formula is a $B_{2+\varrho}$-solution (with respect to the parameter $1/h$) under the assumption that the method for solution of* (12) *is such that the residual vector fulfils the estimate*

$$\max_k |r_k^{(l)}| \leqq K \frac{\delta}{h^\varrho}.$$

This theorem describes the stability of the Crank-Nicholson formula under the assumption that for the chosen method the asymptotic behaviour of residues arising by solving (12) is known. Analogically, the stability of a general parabolic equation with general boundary conditions may be investigated.

In the case, when for the solution of (12) the elimination method is used, one can prove by considering the concrete form of the matrix $A^{(l)}$ and the vector $v^{(l)}$ that the residue fulfils

$$\max_k |r^{(l)}| \leqq K\delta,$$

where K is independent on h. In this case, we obtain a B_2-solution for the entire process. And this is a rather favourable result.

We have investigated in some cases the numerical stability by the concept of β_s-solution. This assesment of numerical stability is of an asymptotic, essentially qualitative character. Our approach, maximalistic in essence, shows the trend of accumulated errors rather than their accurate bounds. The characterization of methods by β_s-solutions may be utilized in different ways. First of all a comparison of different methods will most conveniently be based not only on computer time and memory capacity required, but also their numerical stability. Another example is in some cases utilized combination of methods. It is no sense in utilizing some, e.g., iterative method in order to get a more acsurate solution, if its numerical stability is equal or even worse than that of the original method. Occasionally, we may use the conclusions concerning β_s-solutions even in a quantitative way, for example by comparison with some simple case, where the error is known.

Naturally there are many other such possibilities which depend on a person's experience, intuition and skill.

REFERENCES

[1] J. H. Wilkinson: *Rounding errors in algebraic processes, H. M. S. O., London, 1963.*

[2] P. Henrici: *Discrete variable methods in ordinary differential equations, J. Wiley & Sons, Inc., New York, London, 1962.*

[3] P. Henrici: *Error propagation for difference methods, J. Wiley & Sons, New York, London, 1963.*

[4] С. Л. Соболев: *Некоторые замечания о численном решении интегральных уравнений, Изв. АН СССР, Сер. матем., 20 (1956), 413—436.*

[5] *Error in digital computation, T. I, II, edited by L. B. Rall, J. Wiley & Sons, New York, London, Sydney, 1965.*

[6] *Differential equations and their applications, Proceedings of the conference held in Prague in September 1962, NČSAV, Praha, 1963.*

[7] I. Babuška, M. Práger and E. Vitásek: *Numerical processes in differential equations, J. Wiley & Sons, New York, London, Sydney and SNTL, Praha, 1966.*