Milan Práger; Emil Vitásek

Stability of numerical processes

In: (ed.): Differential Equations and Their Applications, Proceedings of the Conference held in Prague in September 1962. Publishing House of the Czechoslovak Academy of Sciences, Prague, 1963. pp. 123--130.

Persistent URL: http://dml.cz/dmlcz/702179

STABILITY OF NUMERICAL PROCESSES

M. PRÁGER, E. VITÁSEK, Praha

In this paper, we shall deal with questions of the stability of numerical methods for the solution of initial value problems for ordinary differential equations. Let us first mention the problem of the stability of a numerical process in general. We can imagine a numerical process by means of which we get an approximate solution of a differential equation or, in general, of an arbitrary mathematical problem, in the following way.

Definition 1. *Suppose there is given a sequence of vector spaces* X_i $(i = 1, 2, \ldots)$, *a sequence of operators* A_i, *mapping the cartesian product* $X_1 \times X_2 \times \ldots \times X_i$ *into* X_{i+1} $(i = 1, 2, \ldots)$ *and an element* $x_1 \in X_1$; *we call the sequence of equations*

$$x_{i+1} = A_i(x_1, x_2, \ldots, x_i), \quad i = 1, 2, \ldots \tag{1}$$

a numerical process.

Remark. Although the result of every arithmetic operation is only a single number, it is often useful to imagine a numerical process as a process upon more general elements (vectors, matrices etc.), i.e. to consider always a group of arithmetic operations as a single operation. This yields certain specific difficulties which will be mentioned later.

The essential problem, though not the only one, to be solved in numerical methods, is the question of convergence of the values computed from equation (1) towards the exact solution of the original problem. The affirmative answer to this question does not yet guarantee a successful carrying out of the numerical process. For example in constructing sequence (1) in practice, we do not use exact values for computing, but we always round them off on some way. And this circumstance can influence the whole numerical process in a substantial way, so that it becomes completely worthless. Let us explain this by two simple examples.

Example 1. We are to compute the integrals

$$u_n = \int_0^1 x^n \sinh x \, dx, \quad v_n = \int_0^1 x^n \cosh x \, dx, \quad n = 0, 1, \ldots, N \tag{2}$$

where N is a large number. For these integrals the well known recurrent formulas obtained by integration by parts hold:

$$\begin{aligned} u_n &= \cosh 1 - n v_{n-1}, \quad n = 1, 2, \ldots, \quad u_0 = \cosh 1 - 1, \\ v_n &= \sinh 1 - n u_{n-1}, \quad n = 1, 2, \ldots, \quad v_0 = \sinh 1. \end{aligned} \tag{3}$$

If we use these recurrent formulas as a numerical process to compute the integrals (2), the convergence obviously holds because formulas (3) are exact. The following table contains the results of the computation of v_n (the results for u_n are similar): The table shows that the results obtained by means of the recurrent formulas become quite absurd after only a few steps and that this situation does not improve much if we increase the number of decimal digits.

Tab. 1

n	Values computed to		Exact values
	5 decimals	8 decimals	
0	1·17520	1·17520119	1·17520119
1	0·63212	0·63212056	0·63212056
2	0·43944	0·43944231	0·43944231
3	0·33868	0·33868266	0·33868265
4	0·27616	0·27618639	0·27618637
5	0·23340	0·23345124	0·23345087
6	0·20152	0·20230911	0·20230908
7	0·17644	0·17858886	0·17857299
8	0·11568	0·15986631	0·15986485
9	− 0·00884	0·14587344	0·1447310 5
10	− 3·84440	0·13236279	0·1322314 3
11	− 16·77108	0·24739266	0·1217302 0
12	−524·80256	0·13012191	0·1127816 9
13		19·7084081	0·1050639 8

Example 2. Let us solve the differential equation

$$y' = 1 - y^2, \quad y(0) = 5 \tag{4}$$

by using mid-point rule

$$y_{n+2} = y_n + 2h\, y'_{n+1}\,. \tag{5}$$

(This method can be found in all usual handbooks on numerical methods, see e.g. Milne [1].) The results can be seen in the table 2 (p. 125) which shows the error of the approximate solution computed with the aid of formula (5).

The approximate solution oscillates arround the exact solution and the amplitude of these oscillations grows simultaneously with the interval where the solution is sought.

These very simple examples show that there can be a substantial difference between the real numerical process carried out by the computer and its mathematical idealization represented e.g. by equations (1). It seems indisputable that an exact analysis of the real numerical process will be at least for a long time impossible, and therefore

it is necessary to use some idealized model and to study those of its properties which enable us to draw conclusions about the real numerical process. This also gives an intuitive meaning to the notion of a stable numerical process. It will be a numerical process where the undesirable properties mentioned above do not occur.

Tab. 2

$y' = 1 - y^2$, $y(0) = 5$, $h = 0.01$		
x_n	y_n	10^4 . error
0,00	5,0000000	0
0,01	4,7714360	0
0,02	4,5646667	10
0,03	4,3747114	6
0,04	4,2019043	15
0,05	4,0415816	10
0,06	3,8952026	19
0,07	3,7581191	10
. . .	. . .	. . .
1,30	1,1313933	272
1,31	1,0739150	− 281
1,32	1,1283264	284
1,33	1,0684523	− 293
1,34	1,1254882	297
1,35	1,0631117	− 307
1,36	1,1228790	310
1,37	1,0578933	− 321
. . .	. . .	. . .
1,93	0,9230361	−1055
1,94	1,1317138	1038
1,95	0,9174209	−1100
1,96	1,1348724	1081
1,97	0,9116530	−1146
1,98	1,1382446	1125
1,99	0,9057326	−1195
2,00	1,1418304	1171

Suppose a numerical process (1) is given and let $\tilde{x}_{i+1}$ be the value really computed by the computer in the i-th step (i.e. by using inexact values $\tilde{x}_1$, $\tilde{x}_2, \ldots, \tilde{x}_i$ and by inexactly carrying out the operations given by the operator A_i), hence

$$\tilde{x}_{i+1} = A_i^*(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_i) \tag{6}$$

where A_i^* denotes the inexact value of the operator A_i. One of the possibilities of idealization of equations (6) consists in the assumption $A_i^*(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_i) = A_i(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_i) + \delta_i$. Then the numerical process can be described by the equation

$$\tilde{x}_{i+1} = A_i(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_i) + \delta_i\,. \tag{7}$$

We shall base our considerations on this model of the numerical process. It is possible to define its stability in the following way:

Definition 2. *Suppose there is given a numerical process in the sense of definition 1 and let its solution be identically zero. We call this trivial solution numerically stable if for every* $\varepsilon > 0$ *there exists a* $\delta > 0$ *so that every solution of equation* (7) *which satisfies*

$$|\tilde{x}_1|_1 < \delta\,, \quad |\delta_i|_{i+1} < \delta \tag{8}$$

satisfies the inequalities

$$|\tilde{x}_i|_i < \varepsilon\,, \quad i = 1, 2, \ldots \tag{9}$$

($|\ |_i$ *denotes the norm in the space* X_i).

Definition 3. *Suppose there is given a sequence of numerical processes*

$$x_{i+1}^{(m)} = A_i^{(m)}(x_1^{(m)}, \ldots, x_i^{(m)})\,, \quad i = 1, 2, \ldots\,, \quad m = 1, 2, \ldots \tag{10}$$

in the sense of definition 1 and suppose that the solution of each of these numerical processes is identically zero. We call these solutions uniformly stable if for every $\varepsilon > 0$ *there exists a* $\delta > 0$ *(independent on* m*) so that all the solutions of the equations*

$$\tilde{x}_{i+1} = A_i^{(m)}(\tilde{x}_1^{(m)}, \ldots, \tilde{x}_i^{(m)}) + \delta_i^{(m)} \tag{11}$$

which satisfy

$$|\tilde{x}_1^{(m)}|_1 < \delta\,, \quad |\delta_i^{(m)}|_{i+1} < \delta \tag{12}$$

satisfy the inequalities

$$|\tilde{x}_i^{(m)}|_i < \varepsilon\,, \quad i = 1, 2, \ldots\,, \quad m = 1, 2, \ldots \tag{13}$$

These definitions are based on experiences resulting from systematic experiments with many numerical methods carried out in the Mathematical Institute of the Czechoslovak Academy of Sciences. According to our experience the knowledge of whether or not their assumptions are fulfilled gives very good information about the possibility or impossibility of carrying out the numerical process.

Let us now turn to an investigation of the stability of numerical methods for the solution of initial value problems for ordinary differential equations.

Let us first study in detail the difference methods. Suppose there is given a differential equation

$$y' = f(x, y)\,, \quad x \in \langle 0, a\rangle \tag{14}$$

with the initial condition $y(0) = y_0$. Let us divide the interval $\langle 0, a\rangle$ into N parts of length h. Our task is to study the stability of the difference equation

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f((n + \nu)\, h, y_{n+\nu})\,, \quad n = 0, 1, \ldots, N - k \tag{15}$$

with initial conditions

$$y_\kappa = y_{0,\kappa}\,, \quad \kappa = 0, 1, \ldots, k - 1\,, \tag{16}$$

provided that the coefficients α_ν, β_ν satisfy $p + 1$ conditions:

$$\sum_{\nu=0}^{k} \alpha_\nu = 0\,, \quad \sum_{\nu=0}^{k} \frac{\alpha_\nu \nu^s}{s!} = \sum_{\nu=0}^{k} \frac{\beta_\nu \nu^{s-1}}{(s-1)!}\,, \quad s = 1, 2, \ldots, p\,. \tag{17}$$

Conditions (17) guarantee that the differential equation (14) is locally approximated by the difference equation (15), i.e. that a sufficiently smooth solution $y(x)$ of equation (14) satisfies

$$\sum_{\nu=0}^{k} \alpha_\nu\, y(x + \nu h) - h \sum_{\nu=0}^{k} \beta_\nu\, y'(x + \nu h) = 0(h^{p+1})\,. \tag{18}$$

In this case we shall say that the difference equation (15) is of degree p.

The problem of the convergence of the solution of equation (15) was studied in detail by Dahlquist [2]. He proved a theorem which for our purpose can be formulated as follows:

Theorem 1. *Let the solution $y(x)$ of the differential equation* (14) *have in the interval interval* $\langle 0, a\rangle$ $p + 1$ *continuous derivatives and let* $f(x, y)$ *be continuous and satisfy a Lipschitz condition with respect to* y *in a neighbourhood of the solution* $y = y(x)$. *Further suppose a difference formula* (15) *of degree* $p \geqq 1$ *is given such that for every root* ζ_i *of the characteristic polynomial* $p(\zeta) = \sum_{\nu=0} \alpha_\nu \zeta^\nu$ *we have* $|\zeta_i| \leqq 1$ *and all roots of absolute value 1 are simple; let* $\tilde{y}_n$ *be a solution of the difference equation*

$$\sum_{\nu=0}^{k} \alpha_\nu \tilde{y}_{n+\nu} = h \sum_{\nu=0}^{k} \beta_\nu f((n + \nu) h, h, \tilde{y}_{n+\nu}) + o(h) \tag{19}$$

with initial conditions (16). *Then for every* $\varepsilon > 0$ *there exists a* $\delta > 0$ *and a* $h_0 > 0$ *such that*

$$|\tilde{y}_{0,\kappa} - y(\kappa h)| < \delta\,, \quad \kappa = 0, 1, \ldots, k - 1,\ h < h_0 \Rightarrow \tag{20}$$
$$\Rightarrow \sup_{n \leqq N} |\tilde{y}_n - y(nh)| < \varepsilon\,.$$

This theorem, in the first place, completely solves the problem of the convergence of a multistep method. In the second place, it implies that a solution of equation (15) is stable in the sense of definition 2 for every fixed sufficiently small h and in every fixed interval $\langle 0, a\rangle$. It would seem now that the problems of stability of a multistep method are completely solved by this theorem. However, let us look at the statement of theorem 1 in more detail. Firstly, the rate of improvement in the accuracy of carrying out separate operations (or groups of operations, as mentioned above) necessary for the required accuracy of the result may be of higher order than the rate of decrease of h. In the second place, this accuracy can very substantially depend on the length of the interval where we seek the solution. It can easily happen that a solution of equation (15) is stable in every finite interval but is not stable in an infinite interval. Then, of course, it is very difficult to decide which interval for practical computing can still be regarded as a finite one, because with the increase of the interval the process becomes less stable.

Now it could seem convenient to judge the stability of a method with respect to the infinite interval. This, of course, is not immediately possible because the vector field of a given differential equation outside of the interval $\langle 0, a\rangle$ can be unknown to us and, as a matter of fact, we are not interested in it when we are solving the equation in the interval $\langle 0, a\rangle$. Consequently, it will be a matter of finding a convenient way to judge the tendency of the solution outside of the interval $\langle 0, a\rangle$. This can be done by the following asymptotic process based on definition 3. Instead of the stability of the solution of the differer.ce equation (15), we shall investigate the uniform stability of

the system of difference equations

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu}^{(m)} = h \sum_{\nu=0}^{k} \beta_\nu f\left(\frac{n+\nu}{m} h,\, y_{n+\nu}^{(m)}\right), \tag{21}$$

$$m = 1, 2, \ldots,$$

$$n = 0, 1, \ldots, mN - k\,.$$

Each equation of this system represents a numerical solution of the differential equation

$$y' = f\left(\frac{x}{m},\, y\right), \quad x \in \langle 0, ma\rangle\,, \quad m = 1, 2, \ldots \tag{22}$$

Consequently, our process consists in gradually extending the vector field of a given differential equation to the interval $\langle 0, \infty)$, assuming only a knowledge of the vector field in the interval $\langle 0, a\rangle$, and requiring that the stability remains invariable according to this extension. According to our experience, we can assert that the requirement of uniform stability agrees very well with our intuitive conception of a good method, i.e. if a uniformly stable method is used no indesirable properties (such as a rapid less of decimals) occur. Apparently, this concept is not entirely exhausting; it is substantially based on the requirement of a small absolute error and thus it is to a certain degree adapted to the class of problems which make this requirement natural. If, for example, the solutions we are looking for are rappidly increasing, then it is evident that the requirement of a small absolute error in the approximate solution would not be reasonable. In this case, it is more reasonable to require a small relative error. These problems, however, are much more complicated; they depend to a high degree on the character of the solution looked for and it is usually necessary to solve them individually.

A theorem on the uniform stability of the solution of the system of difference equations (21) follows:

Theorem 2. *Let $f(x, y)$ be a continuous function of two variables satisfying a Lipschitz condition with respect to y (with a constant independent of x) in a neighbourhood of the segment $0 \leqq x \leqq a$, $y = 0$ and let $f(x, 0) \equiv 0$. Suppose that for every fixed $\mu \in \langle 0, a\rangle$ the following condition holds: to every $\varepsilon > 0$ and to every (sufficiently small) $\eta > 0$ there exist a $\delta > 0$ and a $N_0 > 0$ so that every solution $z_n^{(\mu)}$ of the difference equation*

$$\sum_{\nu=0}^{k} \alpha_\nu z_{n+\nu}^{(\mu)} = h \sum_{\nu=0}^{k} \beta_\nu f(\mu, z_{n+\nu}^{(\mu)}) + \delta(n), \quad n = 0, 1, \ldots \tag{23}$$

satisfying $|z_\kappa^{(\mu)}| < \eta$, $\kappa = 0, 1, \ldots, k-1$ and $|\delta(n)| < \delta$ satisfies the inequality $|z_n^{(\mu)}| < \varepsilon$ for $n \geqq N_0$ (we shall call this condition which is stronger than the stability in sense of definition 2 asymptotic stability). Then the trivial solution of equation (21) is uniformly stable for every fixed sufficiently small h. Inversely, if the trivial

solution of equation (21) *is uniformly stable, then the trivial solution of equation*

$$\sum_{\nu=0}^{k} \alpha_\nu z_{n+\nu}^{(\mu)} = h \sum_{\nu=0}^{k} \beta_\nu f(\mu, z_{n+\nu}^{(\mu)}) \tag{24}$$

is stable for every fixed $\mu \in \langle 0, a \rangle$.

The main importance of this theorem consists in the fact that it reduces the investigation of the uniform stability or instability of the system (21) to the investigation of the asymptotic stability or instability of the difference equation (24) which does not contain the independent variable in an explicit form. This kind of problem is obviously much simpler.

With the help of theorem 2, we can derive a very simple sufficient condition for the uniform stability of a system (21) if the right hand side of the given differential equation is of a special type.

Theorem 3. *Suppose that the function* $f(x, y)$ *satisfies the assumptions of theorem 2 and suppose that it also satisfies*

$$f(x, y) = A(x)\, y + f_0(x, y) \tag{25}$$

where $A(x) > 0$ *on the interval* $\langle 0, a \rangle$ *and* $f_0(x, y) = o(|y|)$ *for* $|y| \to 0$. *Further suppose that all the roots of the characteristic polynomial* $p(\zeta) = \sum_{\nu=0}^{k} \alpha_\nu \zeta^\nu$ *with the exception of one root equal to one lie inside the unit circle. Then the trivial solution of the system* (21) *is uniformly stable for an arbitrary sufficiently small* h.

This sufficient condition enables us to assert that if we use formulas which satisfy the assumptions of Dahlquist's convergence theorem and whose characteristic polynomial has more than one root lying on the unit circular line, then an inadmissible less of decimals can occur. This is also verified by example 2.

Let us point out the fact that the uniform stability depends not only on the formula itself, but also on the differential equation whose solution we are seeking. This is obviously natural and it follows from the concept of a small absolute error mentioned above.

Until now, we have dealt with difference methods. As to the methods of the Runge-Kutta type, analogous theorems can be derived.

Let us now notice a very important fact concerning the model of a numerical process (7). The realization of a real numerical process evidently does not consist in computing the exact value of the operator A_i at the point $(\tilde{x}_1, \ldots, \tilde{x}_i)$ and then adding the error δ_i, but on the contrary, the error δ_i arises from the inaccuracy with which the operations determined by the operator A_i are performed. The stability of the model then means that the error of the result will be small if the δ_i are small. So if we are using this model, there still remains the question of the realization of small δ_i with the aid of a given type of computer. However, this is usually easier than to choose the

second alternative, namely to consider the question of realization as a part of the model of the numerical process.

The notion of stability used in our investigation has been based on the model of a numerical process given by equation (7). This model, as we point out again, did not come into existence by chance but arose from our experience with various numerical methods and it is available for investigating many other numerical methods, as e.g. methods for solving systems of linear algebraic equations, methods for solving boundary problems etc.

REFERENCES

[1] W. E. MILNE: Numerical solutions of differential equations. Wiley, New York-London 1953.
[2] G. DAHLQUIST: Convergence and stability in the numerical integration of ordinary differential equations. Math. Scand. *4* (1956), 1, 33—53.