Aplikace matematiky

Emil Vitásek

Numerical stability in solution of ordinary differential equations

Aplikace matematiky, Vol. 13 (1968), No. 2, 203–207

Persistent URL: `http://dml.cz/dmlcz/103158`

NUMERICAL STABILITY IN SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

E. VITÁSEK

1. INTRODUCTION

In this paper, I will deal with the problems of the numerical stability in solution of an initial-value problem for a special-type ordinary differential equation of higher order which right-hand term does not depend on the derivatives of the function sought. The numerical stability will be understood in the sense of [1] and [2], i.e., the dependence of the accumulated roundoff error on the number of subintervals on which the interval in which the solution is saught is divided will be investigated. Particularly, we will deal with the multistep difference method and, especially, we will be interested in the manner how the properties of the stability of this method can be influenced in positive direction. We also mention very briefly the above problems for the Runge-Kutta-type methods.

For the sake of brevity, we restrict ourselves in what follows only on the second-order differential equation of the above mentioned type, i.e., on the equation

$$y'' = f(x, y)\,, \quad x \in \langle a, b \rangle \tag{1}$$

with initial conditions

$$y(a) = \eta_0\,, \quad y'(a) = \eta_1\,. \tag{2}$$

This restriction is only formal and all essential what will be said in what follows can be directly generalized on the general case.

2. MULTISTEP DIFFERENCE METHOD

The general multistep difference formula for the solution of the equation (1) can be written in the following form:

$$\sum_{\nu=0}^{k} \alpha_\nu y_{n+\nu} = h^2 \sum_{\nu=0}^{k} \beta_{\nu.} f_{n+\nu}\,, \quad \alpha_k \neq 0\,, \quad f_n = f(x_n, y_n) \tag{3}$$

or symbolically,

$$\varrho(E)\, y_n = h\, \sigma(E) f_n \tag{4}$$

where E is the translation operator $(Ey_n \equiv y_{n+1})$, $\varrho(\zeta)$ and $\sigma(\zeta)$ are polynomials defined by

$$\varrho(\zeta) = \sum_{\nu=0}^{k} \alpha_\nu \zeta^\nu\,, \quad \sigma(\zeta) = \sum_{\nu=0}^{k} \beta_\nu \zeta^\nu \tag{5}$$

$h = (b - a)/N$, y_n is the approximate solution in the point $x_n = a + nh$ and N is the number of subintervals of the interval $\langle a, b\rangle$. Without any loose of gnerality, it can be assumed that the polynomials (5) have no common factors for, in the opposite case, the difference equation (3) can be reduced on the equation of lower order. Note that for using of (3), k initial values of y_n must be available and they must be obtained by some other method. We will assume in what follows that these values are known. The following two definitions will be useful in further investigation:

Definition 1. The formula (3) (or, briefly, the polynomial $\varrho(\zeta)$) will be said to be r-stable in the sense of Dahlquist if the polynomial $\varrho(\zeta)$ does not have zeros outside the unit circle and if all roots lying on unit circumference are at most of the multiplicity r.

Definition 2. The formula (3) will be said to be r-consistent if

$$\varrho(1) = \varrho'(1) = \ldots = \varrho^{(r-1)}(1) = 0\,, \quad \varrho^{(r)}(1) = r!\, \sigma(1)\,. \tag{6}$$

It is well known (cf., for example, [3]) that under the above assumptions the 2-stability and 2-consistency are necessary and sufficient conditions for the convergence. It is therefore natural to assume that the formula (3) is 2-stable and 2-consistent.

Denote now by $\tilde{y}_n$ the solution of

$$\varrho(E)\, \tilde{y}_n = h^2\, \sigma(E) \tilde{f}_n\,, \quad \tilde{f}_n = f(x_n, \tilde{y}_n)\,. \tag{7}$$

Then it is possible to prove (analogical problems cf., e.g., [1], [2]).

Theorem 1. *Let the formula* (4) *be 2-stable in the sense of Dahlquist. Further, let the right-hand term of* (1) *be continuous and satisfying the Lipschitz condition with respect to y in the domain $a \leqq x \leqq b$, $-\infty < y < \infty$. Finally, let y_n be the solution of* (4) *and $\tilde{y}_n$ of* (7) *with the same initial values. Then*

$$|\tilde{y}_n - y_n| \leqq K_1 \frac{\delta}{h^2} \tag{8}$$

where $\delta = \max |\delta_n|$ *and K_1 is a constant not depending on h.*

Consequently, in terms of [2], this theorem expresses that the numerical process defined by (4) has B_2-solution (with N as parameter). It is clear that under the natural assumption of 2-consistency, the assumptions of Theorem 1 cannot be modified in such a manner to gain the numerical process with B_s-solution with $s < 2$. (It is, however, useful to note here that the 2-consistency is not essential for the assertion of Theorem 1.) This result is not satisfactory in comparison with the numerical stability of the method arising when the given differential equation is replaced by a system of first-order equations and this system is then solved by some difference method for solution of first-order equations. Such numerical processes lead to B_1-solutions. It is therefore natural to ask if it is not better to use always the last method and to omit formulae (4). But the positive answer on this question would mean that one resignes on an advantage of the methods of the type (4) which consists in that fact that these formulae have the local truncation errors of higher orders than the methods arising by solving (1) as a system and using the same number of points (cf., for example, [3] and [4]). Let us investigate therefore if it is not possible to profit by the more favourable properties of the stability of difference formulae for solution of first-order equations in some other way. One possibility is to change the equation (4) leading to evaluation of y_n in such a manner that the new equations will be formaly of the form of difference equations for solution of first-order systems. We will deal therefore with the possibility of replacing of (4) by

$$\varrho_1(E)\, y_n = h\, \sigma_1(E)\, z_n\,, \quad \varrho_2(E)\, z_n = h\, \sigma_2(E) f_n \tag{9}$$

where the degrees of $\sigma_1(\zeta)$ and $\sigma_2(\zeta)$ are not greater than the degrees of $\varrho_1(\zeta)$ and $\varrho_2(\zeta)$, respectively. It is ovious, that the equations (4) and (9) will be equivalent if

$$\varrho(\zeta) = \varrho_1(\zeta)\, \varrho_2(\zeta)\,, \quad \sigma(\zeta) = \sigma_1(\zeta)\, \sigma_2(\zeta) \tag{10}$$

and if it will be possible to choose the initial values of the auxiliary variable z_n so that

$$\tag{11} \begin{aligned} \sigma_1(E)\, z_\nu &= \frac{1}{h}\, \varrho_1(E)\, y_\nu\,, \quad \nu = 0, \dots, k_2 - 1 \\ \varrho_2(E)\, z_\nu &= h\, \sigma_2(E) f_\nu\,, \quad \nu = 0, \dots, k_1 - 1 \end{aligned}$$

where k_1 and k_2 are the degrees of $\varrho_1(\zeta)$ and $\varrho_2(\zeta)$, respectively. But the determinant of this system of linear algebraic equations is the resultant of the polynomials $\sigma_1(\zeta)$ and $\varrho_2(\zeta)$ and, consequently, different of zero ($\varrho(\zeta)$ and $\sigma(\zeta)$ have no common factors).

Let now $\hat{y}_n$ be the solution of

$$\tag{12} \begin{aligned} \varrho_1(E)\, \hat{y}_n &= h\, \sigma_1(E)\, \hat{z}_n + \delta_n^{(1)}\,, \\ \varrho_2(E)\, \hat{z}_n &= h\, \sigma_2(E) \hat{f}_n + \delta_n^{(2)}\,, \quad \hat{f}_n = f(x_n, \hat{y}_n)\,. \end{aligned}$$

Then the following theorem holds:

Theorem 2. *Let the polynomials $\varrho(\zeta)$ and $\sigma(\zeta)$ have no common factors. Further, let it be possible to write $\varrho(\zeta)$ and $\sigma(\zeta)$ in the form* (10) *and let $\varrho_i(\zeta)$ and $\sigma_i(\zeta)$ satisfy*

(i) *$\varrho_i(\zeta)$ and $\sigma_i(\zeta)$ have real coefficients;*

(ii) *the degree of $\sigma_i(\zeta)$ is not greater than the degree of $\varrho_i(\zeta)$;*

(iii) *$\varrho_i(\zeta)$ is* 1-*stable in the sense of Dahlquist. Let y_n be the solution of* (4) *and $\hat{y}_n$ of* (12) *where $\hat{y}_\nu = y_\nu$, $\nu = 0, \ldots, k - 1$ and $\hat{z}_\nu$ are determined from* (11). *Finally, let the right-hand term of* (1) *satisfy the assumptions of Theorem* 1. *Then*

$$|y_n - \hat{y}_n| \leqq K_2 \frac{\delta}{h}, \tag{13}$$

where $\delta = \max\left(|\delta_n^{(1)}|, |\delta_n^{(2)}|\right)$ *and K_2 is a constant not depending on h.*

Thus, the equations (9) define a new numerical process having B_1-solution. The only thing which is still to be discussed is the existence of the decomposition of $\varrho(\zeta)$ and $\sigma(\zeta)$ with the properties (i), (ii), (iii) from Theorem 2. About this problem, it can be proved the following

Theorem 3. *Let the formula characterized by $\varrho(\zeta)$ and $\sigma(\zeta)$ be* 2-*stable in the sense of Dahlquist and* 2-*consistent. Denote further by a and $2b$ the number of real zeros different from the unity and the number of complex zeros of $\varrho(\zeta)$, respectively. Finally, denote by $2c$ the number of complex zeros of $\sigma(\zeta)$.* (*Each zero is computed so many times as it is its multiplicity.*) *Then if $a \geqq 1$, the decomposition* (10) *of $\varrho(\zeta)$ and $\sigma(\zeta)$ with properties* (i), (ii), *and,* (iii) *always exists, and if $a = 0$ it exists if and only if $c \leqq b$.*

This theorem thus answers the question when it is possible modifying the algorithm (4) to achieve more convenient properties of the numerical stability. Note that the above described construction of the decomposition of $\varrho(\zeta)$ and $\sigma(\zeta)$ generally does not lead to 1-consistent polynomials $\varrho_i(\zeta)$ and $\sigma_i(\zeta)$ and, consequently, these polynomials cannot be used for the solution of general systems of differential equations of the first order.

3. RUNGE-KUTTA-TYPE FORMULAE

The general Runge-Kutta-type formula for the solution of (1) can be written as follows (cf., e.g., [5]):

$$y_{n+1} = y_n + hz_n + h^2 \sum_{s=0}^{m} w_s(1 - a_s)\, k_s(x_n, y_n, z_n, h)\,, \tag{14}$$

$$z_{n+1} = z_n + h \sum_{s=0}^{m} w_s k_s(x_n, y_n, z_n, h)\,, \quad y_0 = \eta_0\,, \quad y_1 = \eta_1\,, \tag{15}$$

$$k_s(x_n, y_n, z_n, h) = f(x_n + a_s h, y_n + a_s h z_n + h^2 \sum_{i=0}^{s-1} b_{si} k_i(x_n, y_n, z_n, h)\,,$$

and a_s, $s = 0, \ldots, m$, b_{si}, $s = 0, \ldots, m$, $i = 0, \ldots, s - 1$, w_s, $s = 0, \ldots, m$ are constants, $a_0 = 0$. The convenience of this formula in comparison with standard formulae using the corresponding system is based on that fact that this formula saves one evaluation of the right-hand term retaining the same degree of truncation error (cf., e.g. [6]). From the point of view of the numerical stability, the situation is here more convenient than above. It holds

Theorem 4. *Under the same assumptions as above about the function $f(x, y)$, the equations* (14) *and* (15) *define a numerical process with B_1-solution.*

This more convenient result is caused by the fact that (14) is one-step formula and, we are here, as a matter of fact, obliged to compute the approximation of the derivative of the function saught.

References

[1] *I. Babuška, M. Práger, E. Vitásek:* Numerical Processes in Differential Equations, Interscience Publishers, 1966.

[2] *I. Babuška:* Problems of Minimization and Numerical Stability in Computations, Liblice 1967.

[3] *G. Dahlquist:* Convergence and Stability in the Numerical Integration of Ordinary Differential Equations, Math. Scand., *2* (1954), 91—102.

[4] *G. Dahlquist:* Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations, Trans. Royal Inst. of Techn., Stockholm, *130* (1959).

[5] *R. E. Scraton:* The Numerical Solution of Second-Order Differential Equations Not Containing the First Derivative Explicitly, Comp. J., *6* (1964), 368—370.

[6] *L. Collatz:* The Numerical Treatment of Differential Equations, Springer Verlag Berlin, 1960.

E. Vitásek, MÚ ČSAV, Opletalova 45, Praha 1, ČSSR.