Simulation of Dynamic Systems Using Global Calculation Error

Yaroslav Shmygelsky
Department of radiophysics and computer technology
Ivan Franko National University of Lviv,
107 Tarnavsky Str, 79017 Lviv, Ukraine
shmygelsky@ukr.net

Volodymyr Brygilevych
The Bronisław Markiewicz State Higher School of Technology
and Economics in Jarosław (PWSTE in Jarosław)
Jarosław, Poland

Abstract—**A numerical value of the estimate of global discretization error is used in the simulation process of transient and steady-state modes in dynamic systems. This allowed the calculation accuracy to by increased significantly with low computational expanses.**

Index Terms—**dynamic systems; transient and steady-state modes; numerical methods; global discretization error**

I. Introduction

One of the most important characteristics that determine the reliability of the process of numerical simulation of dynamic systems is the global calculation error, that is, the difference between accurate and calculated solutions over the whole range of integration.

The significance of such a characteristic for the researcher is undeniable. In general, it is believed that a good matrix for solving the Cauchy problem for ordinary differential equations should, at the request of the user, provide information about the global calculation error at each starting point [1].

Naturally, the exact definition of global error in modeling programs is not real. We can only talk about some of its approximation. In many cases, the investigator can only satisfy the order of the magnitude of the error and its sign. But if the global error is determined by the corresponding algorithms sufficiently reliable, then its calculated value can be tried to use to correct the results of the calculation conducted with low accuracy.

II. Preliminary rezults

Consider the mathematical model of a dynamical system in the form implicit differential-algebraic equations:

$$\boldsymbol{F}(\boldsymbol{x},\dot{\boldsymbol{x}},t)=0, \tag{1}$$

where $\boldsymbol{x}\in\boldsymbol{R}^n$ - is the vector of independent variables, $\boldsymbol{F}:\boldsymbol{R}^{2n+1}\to\boldsymbol{R}^n$ - is the vector-function, continuous by t and continuously differentiated with respect to $\boldsymbol{x}$ and $\dot{\boldsymbol{x}}$.

Let $\boldsymbol{x}(t)=\boldsymbol{\varphi}(t,t_0,\boldsymbol{x}_0),\ t\in[t_0,t_{fin}]$ - the solution of the system (1) satisfying the initial condition $\boldsymbol{x}(t_0)=\boldsymbol{x}_0$, and $\boldsymbol{x}_m\approx\boldsymbol{x}(t_m)$ - the numerical approximation of this solution on the grid $\Omega:\{t_{m+1}=t_m+h_m,\ h_m>0,\ m=0,1,...\}$.

The global discretization error on a grid Ω is the value [3]:

$$\boldsymbol{\delta}(t_m)=\boldsymbol{x}_m-\boldsymbol{x}(t_m),\ t_m\in\Omega. \tag{2}$$

The value of the global discretization error depends on the type of method used for numerical integration of the system (1) and the value of its local truncation error.

If for an algebraization of the derivative in (1) on the grid Ω an implicit k-step method of the backward differentiation formulas (the BDF method [4]) is used

$$\dot{\boldsymbol{x}}_{m+1}=-\frac{1}{h_{m+1}}\sum_{i=0}^{k}\alpha_i\boldsymbol{x}_{m+1-i} \tag{3}$$

then, as shown in [2], the numerical approximation $\boldsymbol{\delta}_m$ of the global error $\boldsymbol{\delta}(t_m)$ on the grid Ω satisfies the equation

$$\boldsymbol{A}_{m+1}\boldsymbol{\delta}_{m+1}=\boldsymbol{B}_{m+1}\frac{1}{h_{m+1}}\left[\sum_{i=0}^{k}\alpha_i\boldsymbol{\delta}_{m+1-i}+\boldsymbol{e}_{m+1}\right], \tag{4}$$

where

$$\begin{aligned}\boldsymbol{A}_{m+1}&=\left[\frac{\partial F(\boldsymbol{x},\dot{\boldsymbol{x}},t)}{\partial\boldsymbol{x}}+\frac{\alpha_0}{h_{m+1}}\frac{\partial F(\boldsymbol{x},\dot{\boldsymbol{x}},t)}{\partial\dot{\boldsymbol{x}}}\right]_{x=x_{m+1},\dot{x}=\dot{x}_{m+1},t=t_{m+1}},\\ \boldsymbol{B}_{m+1}&=\left[\frac{\partial F(\boldsymbol{x},\dot{\boldsymbol{x}},t)}{\partial\dot{\boldsymbol{x}}}\right]_{x=x_{m+1},\dot{x}=\dot{x}_{m+1},t=t_{m+1}},\end{aligned} \tag{5}$$

$\boldsymbol{e}_{m+1}$ - is the local truncation error of the BDF method at the time moment t_{m+1}, calculated by the Brayton formula [2]:

$$\boldsymbol{e}_{m+1}=\frac{h_{m+1}}{t_{m+1}-t_{m-k}}\left(\boldsymbol{x}_{m+1}-\boldsymbol{x}_{m+1}^0\right), \tag{6}$$

where $\boldsymbol{x}_{m+1}^0$ - is the predicted value for $\boldsymbol{x}_{m+1}$.

The use of global error in modeling programs is justified in the case where the procedure for determining it is cheap enough. If the discrete system model (1)

$$\boldsymbol{F}(\boldsymbol{x}_{m+1}, \dot{\boldsymbol{x}}_{m+1}, t_{m+1}) = 0, \tag{7}$$

where $\dot{\boldsymbol{x}}_{m+1}$ is determined by the expression (3), is solved at each time step by the Newton method, then the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are calculated in the process of numerical integration of the system (1), and the matrix $\boldsymbol{A}$ is usually represented by its LU-expansion. Therefore, the computational cost for finding $\boldsymbol{\delta}_{m+1}$ from equation (4) does not exceed the cost of one Newtonian iteration for equation (7).

Equation (4) is obtained by linearization of the corresponding nonlinear equation, which correlates the true value of the global error $\boldsymbol{\delta}(t_{m+1})$ with the true value of the local error $\boldsymbol{e}(t_{m+1})$ of the numerical method [3]. Therefore, in the case of the linear system (1), equation (4) is valid for any of the values of the global error and the accuracy of the determination $\boldsymbol{\delta}_{m+1}$ is determined only by the error of estimation $\boldsymbol{e}_{m+1}$. In the case of a nonlinear system (1), equation (4) is valid only for small values $\boldsymbol{\delta}_{m+1}$ at the points of the grid Ω, and the domain of its applicability, generally speaking, can be determined only experimentally.

III. Calculation of transient regimes

The numerical experiments have shown that the global error for nonlinear systems by using BDF methods is determined by (4) very reliably, even when its value reached 10-15% of the calculated variable value. Therefore, there are every reason to use a numerical approximation of global error to correct the calculation results in order to increase their accuracy.

If $\boldsymbol{\delta}_m$ is a good approximation for $\boldsymbol{\delta}(t_m)$, then, at is follows from (2),

$$\hat{\boldsymbol{x}}_m = \boldsymbol{x}_m - \boldsymbol{\delta}_m \tag{8}$$

will be a better approximation for $\boldsymbol{x}(t_m)$, than $\boldsymbol{x}_m$ at the grid Ω.

Example 1. Consider the calculation of the transient process in the sequential linear oscillatory *RLC*-circuit, which is acted upon by the harmonic e.m.f. $E(t) = A\sin(\frac{2\pi}{T}t)$. Mathematical model of *RLC*-circuit:

$$\begin{aligned} &C\dot{U}_C - I_L = 0, \\ &L\dot{I}_L + RI_L + U_C - A\sin\left(\frac{2\pi}{T}\right) = 0, \end{aligned} \tag{9}$$

where U_C and I_L - voltage on the capacitor and current through inductance

The circuit parameters are: *R*=10(Ω), *L*=10^{-3} (H), *C*=10^{-7} (F); The e.m.f. parameters are: *A*=10(V), *T*=62,5·10^{-6} (s).

The proximity of the external force frequency (*ω* = 1.0053 10^5 s^{-1}) to the resonance circuit frequency (*ω*$_0$ = 10^5 s^{-1}) increases the effect of numerical errors on the calculation results. The simulation was carried out using the BDF method with various fixed orders of formulas (3) from zero initial conditions. In all experiments, the limit of local error at the step was assumed to be 10^{-4}. Together with the solution, its global error was determined in accordance with (4). Table 1 present the values of the circuit state variables $U_C(t_M)$ and $I_L(t_M)$ at the instant of time $t_M = 4T$: exact (obtained from the analytical expression for the transient process), calculated (obtained by numerical integration of the circuit equations) and improved (taking into account the global error in accordance with the formula (8)).

.

TABLE I. Calculation of transient regim in an oscillatory circuit

The order of the BDF method	Value U_C (V) exact: -71,284		Value I_L (mA) exact: -28,820	
	calculation	improvement	calculation	improvement
1	-66,287	-71,210	-23,245	-28,600
2	-71,021	-71,339	-38,548	-28,953
3	-71,636	-71,302	-30,121	-28,552
4	-71,394	-71,261	-27,274	-28,621

It follows from the above results, in this example, the global error is determined using equation (4) very confidently. Moreover, the calculated values of the global error make not less than 80% of the true values, and their consideration can significantly improve the accuracy of the simulation.

IV. Calculation of steady-state regimes

Let the system (1) be time periodic with the period T> 0, that is:

$$\boldsymbol{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}, t+T) = \boldsymbol{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}, t) . \tag{10}$$

The search for the periodic regime of the system (1) can be reduced to the solution of the equation [5,6]

$$\boldsymbol{y} - \boldsymbol{P}(\boldsymbol{y}) = 0\,, \tag{11}$$

where $\boldsymbol{y} = \boldsymbol{x}(t_0)$ - vector of initial conditions, $\boldsymbol{P}(\boldsymbol{y})$ - mapping of a point along the trajectory $\boldsymbol{x}(t) = \boldsymbol{\varphi}(t, t_0, \boldsymbol{y})$ of system (1) for period *T*.

In the simulation of system (1) we have to deal with not exact mapping $\boldsymbol{P}(\boldsymbol{y})$, but with its numerical approximation $\tilde{\boldsymbol{P}}(\boldsymbol{y})$, which is calculated by means of some formulas of numerical integration on a grid Ω, and instead of (11) solve the equation

$$\boldsymbol{y} - \tilde{\boldsymbol{P}}(\boldsymbol{y}) = 0\,, \tag{12}$$

Assume that in the discrete system (7) there is a periodic mode, that is, the mapping $\tilde{\boldsymbol{P}}$ has a fixed point. Let $\boldsymbol{y}^*$ - a fixed point of the exact mapping $\boldsymbol{P}$, and $\tilde{\boldsymbol{y}}^*$ - a fixed point of the calculated mapping $\tilde{\boldsymbol{P}}$. For these fixed points, the relation:

$$\boldsymbol{y}^* = \boldsymbol{P}(\boldsymbol{y}^*), \quad \tilde{\boldsymbol{y}}^* = \tilde{\boldsymbol{P}}(\tilde{\boldsymbol{y}}^*) + \boldsymbol{d}\,, \tag{13}$$

where $\boldsymbol{d}$ - some vector satisfying the condition $\|\boldsymbol{d}\| \le \varepsilon$, $\varepsilon > 0$ - the accuracy of solving the equation (12).

We introduce the error of the calculation of the mapping $\boldsymbol{P}$ and the fixed point:

$$\boldsymbol{\delta} = \tilde{\boldsymbol{P}}(\boldsymbol{y}) - \boldsymbol{P}(\boldsymbol{y})\,, \tag{14}$$

$$\boldsymbol{\eta} = \tilde{\boldsymbol{y}}^* - \boldsymbol{y}^*\,, \tag{15}$$

The value $\boldsymbol{\delta}$ is simply defined in (2) the global error of a numerical method at a time moment $t_N = t_0 + T$ when integrating the system (1) from the initial conditions $\boldsymbol{x}(t_0) = \boldsymbol{y}$ on the interval $[t_0, t_0 + T]$. From (15), taking into account (13), we have:

$$\begin{aligned}&\boldsymbol{\eta} = \tilde{\boldsymbol{y}}^* - \boldsymbol{y}^* = \tilde{\boldsymbol{P}}(\tilde{\boldsymbol{y}}^*) - \boldsymbol{P}(\boldsymbol{y}^*) + \boldsymbol{d} = \\ &\left(\tilde{\boldsymbol{P}}(\tilde{\boldsymbol{y}}^*) - \tilde{\boldsymbol{P}}(\boldsymbol{y}^*)\right) + \left(\tilde{\boldsymbol{P}}(\boldsymbol{y}^*) - \boldsymbol{P}(\boldsymbol{y}^*)\right) + \boldsymbol{d}\,,\end{aligned} \tag{16}$$

If the BDF method (3) is used for the system (1) discretization, then, under the assumptions made about the $\boldsymbol{F}(\boldsymbol{x}, \dot{\boldsymbol{x}}, t)$ smoothness, it can be shown [5] that the calculated mapping $\tilde{\boldsymbol{P}}$ will be differentiable whit the derivative continuous in the sense of Lipschitz, that is|:

$$\tilde{\boldsymbol{P}}(\boldsymbol{y}) - \tilde{\boldsymbol{P}}(\boldsymbol{z}) = \tilde{\boldsymbol{P}}'(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{z}) + O\left(\|\boldsymbol{y} - \boldsymbol{z}\|^2\right), \quad \forall \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{R}^n\,, \tag{17}$$

under the condition that $\tilde{\boldsymbol{P}}(\boldsymbol{y})$ i $\tilde{\boldsymbol{P}}(\boldsymbol{z})$ are calculated on the same grid Ω and in the same order of *k* formulas of the numerical method (3) in the corresponding steps, and the matrix $\boldsymbol{A}_{m+1}$ defined in (5) is regular for all $m = 0, 1, \ldots N-1$.

Then we obtain from (16) taking into account (14) and (17)

$$\boldsymbol{\eta} \approx \left(\boldsymbol{I} - \tilde{\boldsymbol{P}}'(\boldsymbol{y}^*)\right)^{-1} (\boldsymbol{\delta} + \boldsymbol{d})\,, \tag{18}$$

where $\boldsymbol{I}$ - is a unit matrix, and the sign $\approx$ means “with precision to members $O\left(\|\boldsymbol{\eta}\|^2\right)$”.

If we do not take into account the rounding errors, then in the linear case it is possible to put $\boldsymbol{d} = 0$, and in the nonlinear case, with the iterative process of the equation (12) is convergent (for example, Newton's method), the value $\|\boldsymbol{d}\|$ can be very low.

Consequently, the error $\boldsymbol{\eta}$ for calculating a fixed point $\boldsymbol{y}^*$ is mainly defined by the value $\left\|\left(\boldsymbol{I} - \tilde{\boldsymbol{P}}'(\boldsymbol{y}^*)\right)^{-1}\right\|$ and global error $\boldsymbol{\delta}$ of the mapping calculation.

For weakly damped systems $\left\|\left(\boldsymbol{I} - \tilde{\boldsymbol{P}}'(\boldsymbol{y}^*)\right)^{-1}\right\| >> 1$. So, in the case of linear system (1), it may be shown [6], that $\left\|\left(\boldsymbol{I} - \tilde{\boldsymbol{P}}'(\boldsymbol{y}^*)\right)^{-1}\right\| \sim Q$, where Q - is the maximum figure of merit of the system (1). Therefore, when looking for a periodic regime of weakly damped systems to obtain reliable results, it is necessary to calculate with very high accuracy, which naturally leads to significant computational costs.

For this to be eliminated, in the search process of the steady-state regime the improved value of the calculated mapping $\tilde{\boldsymbol{P}}(\boldsymbol{y})$ may be used as

$$\hat{\boldsymbol{P}}(\boldsymbol{y}) = \tilde{\boldsymbol{P}}(\boldsymbol{y}) - \boldsymbol{\delta}_N\,, \tag{19}$$

where $\boldsymbol{\delta}_N$ - is the numerical approximation of global error $\boldsymbol{\delta}(t_N)$.

Example 2. Consider the problem of periodic states definition of the Duffing equation:

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1^3 - 0,2x_2 + 0,3\cos t\end{aligned} \tag{20}$$

As is known, this equation has three periodic states, two of which are stable and one unstable. In [6] the periodic states of equation (20) are calculated, with high accuracy. Whit the initial conditions $\left(x_1(0) = 0,027;\ x_2(0) = 1,10\right)$ the stable periodic state $\left(x_1^*(0) = 0,6263873;\ x_2^*(0) = 1,03347995\right)$ is obtained, and with $\left(x_1(0) = -0,027;\ x_2(0) = 0,729\right)$ the unstable periodic state an unstable periodic state

$\left(x_1^*(0) = -0,71598261;\ x_2^*(0) = 0,74740203\right)$. Let's take these values as benchmarks.

In our case, the mapping was calculated by numerical integration of the equation (20) on the interval [0.2π] by the BDF method of the first order (implicit Euler method) with a constant step. The derivative of the reflection for the period $\tilde{\boldsymbol{P}}'(\boldsymbol{y})$ was determined by solving the equation in the variations for (20) simultaneously with $\tilde{\boldsymbol{P}}(\boldsymbol{y})$. The equation (12) was solved by Newton's method with $\varepsilon = 10^{-4}$ accuracy. The accuracy of the calculation of the mapping $\tilde{\boldsymbol{P}}(\boldsymbol{y})$ from experiment to experiment was increased by reducing the value of the integration step. Experiment results are given in Table 2

In the first column of Table 2, the number of steps for integrating N in the period of forced fluctuations T in the various experiments is indicated. The last line shows the calculation results for N = 1000, taking into account the calculated value of global sampling error. The first line of each part of Table 3 gives the initial conditions from which the search for periodicity began. The third column of each part of the table specifies the processor time for solving the task (along with the output of the results).

TABLE II. CALCULATION OF PERIODIC STATES OF DUFFING EQUATION

N=T/h	$x_1(0)$ =0,027; $x_2(0)$ =1,1			$x_1(0)$ =-0,027; $x_2(0)$ =0,729		
	$x_1^*(0)$	$x_2^*(0)$	t s	$x_1^*(0)$	$x_2^*(0)$	t s
1000	0,575	1,039	30	-0,689	0,767	26
5000	0,620	1,034	129	-0,711	0,750	100
10000	0,629	1,032	254	-0,715	0,748	198
1000+gl.error	0,627	1,033	32	-0,717	0,747	28

Comparing the obtained values $x_1^*(0),\ x_2^*(0)$ of periodic states with reference ones, it is not difficult to make sure that the use of numerical approximation of global sampling error in the process of solution allowed, at small computational costs, to increase the accuracy of determining the periodic state of the system (20) and significantly accelerate the process of its search.

REFERENCES

[1] H. J. Stetter, “Global Error Estimation in Ordinary Initial Value Problems,” Lect. Notes Math., 1982, vol. 968, pp. 269–279.

[2] “Modern Numerical Methods for Ordinary Differential Equation,” Edited by G. Hall and J. M. Watt, Clarendon Press, Oxford, 1976 (Russian translation, 1979).

[3] R. K. Brayton, F. G. Gustavson, G. D. Hechtel, “A New Efficient Algorithm for Solving Differential-Algebraic Systems Using Implicit Backward Differentiation Formula,” Proc. IEEE, vol. 60, No. 1, January 1972, pp. 98-108.

[4] Ya. Shmygelsky, “On Definition of Global Error in Numerical Calculation of Electronic Circuit Equations,” Teor. Electrotekhn., 1892, vyp. 34, pp.124–129 (in Russian).

[5] S. Scalboe, “Time Stationary Analisis of Nonlinear Electrical Systems,” Proc. IEEE, vol. 70, No. 10, October 1982, pp. 89-111.

[6] T. J. Aprill and T. N. Trick, “Steady State Analisis of Nonlinear Circuits with Periodic Inputs,” Proc. IEEE, vol. 60, No. 1, January 1972, pp. 108-114.