A STOCHASTIC APPROACH TO GLOBAL ERROR ESTIMATION IN THE NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

A.Rios Neto and N.C.Cardenuto
Instituto de Pesquisas Espaciais, CNPq
São José dos Campos, Brasil

ABSTRACT. A stochastic scheme is proposed to estimate the order of magnitude of global errors in the numerical integration of ordinary differential equations initial value problems. The basic idea consists in modelling a Markov stochastic vector sequence with variances in each discretization time that constrain the order of magnitude of the global error vector components to be within estimated limits with 99% probability. This is done by properly discretizing the dynamics given by the system first order variational equations and by properly perturbing this dynamics with white noise derived from local estimates of truncation and round-off errors. The paper presents the procedure to implement the scheme when Runge-Kutta embedded formulas of two consecutive orders are used. Results of preliminary tests are presented.

1. INTRODUCTION

In the numerical treatment of ordinary differential equations initial value problems, theoretical results concerning the behavior of global errors are not so new. In a step by step solution, these results either qualitatively relate the order of magnitude of global and local errors or quantitatively give upper bounds for global errors in terms of parameters dependent upon local errors, step size, problem and method used (Dahlquist and Björck, 1974). It was only recently that one started finding in the literature results which translate the theory into procedures of practical use (Zadunaisky, 1979; Prothero, 1980). A common characteristic of the existing procedures is that they always give a deterministic global error estimate. However, a stochastic approach is necessary if one wants to have procedures which are naturally fitted to

S.Ferraz-Mello and P.E.Nacozy (eds.), The Motion of Planets and Natural and Artificial Satellites, 73-81. Universidade de São Paulo, São Paulo, 1983.

be used as part of the software needed in the guidance and control of dynamic systems. Typical situations are those where one has to deal with state estimates predictions - as is the case in statistical orbit determination - or when one has to calculate the evolution with time of system state from noisy measurements of strap-down instruments.

In the procedure to be proposed in the following sections, global errors are considered in terms of their order of magnitude and estimated through the dispersions of zero mean random variables belonging to stochastic sequences which take in account the influence of both local truncation and round-off errors. Although the procedure is not restricted to only this situation, in this paper it is applied using a one-step method where truncation error is locally evaluated by using Fehlberg's embedded formulas of two consecutive orders (1968). To propagate and accumulate the influence of local errors, the first order variational equations associated to the system being integrated are used. However, first order Euler method can be considered to approximate the covariance matrix necessary in each step to characterize global error in the estimation scheme (Prothero, 1980).

2. PROBLEM STATEMENT

Consider a system of first order ordinary differential equations:

$$\dot{x} = f(x,t) \tag{1}$$

where $f(x,t)$ is a vector function of the $n \times 1$ vector dependent variable x and of the scalar independent variable t, assumed to have the required regularity to guarantee the existence of an unique solution, $x(t)$, once given the initial conditions, $x(t_0)$. Suppose that a discretization method is used to generate a numerical approximation $x(j)$ in the mesh points $t_j (t_0 < t_j \leqslant t_f)$. What is intended is, in some way, to have the capability of estimating the behavior of the global error $\varepsilon(j)$, defined as:

$$\varepsilon(j) = x(j) - x(t_j) \tag{2}$$

where $x(t_j)$ indicates the true solution, and where the numerical approximation $x(j)$ is obtained with the use of a zero-stable method of order p, in order to have equal to one the difference between the exponents in the asymptotic dependence on the step size of the local and global errors (Dahlquist and Björck, 1974).

3. GLOBAL ERROR ESTIMATION

In what follows the situation analysed is that where a one-step method is considered and where truncation error is locally evaluated by the use of embedded formulas of two consecutive orders (Fehlberg, 1968). Since the errors involved are small, it is assumed that linear perturbations give valid approximations allowing the use of the superposition principle after separately treating the contributions of truncation errors and round-off errors.

3.1 - CONTRIBUTION OF TRUNCATION ERRORS

In an intermediate point t_j, as input information for the procedure to be proposed, a numerical approximation $x(j)$ of the true solution $x(t_j)$ of system (1) together with a random estimate $E(j)$ to represent the global error are given. The vector random variable $E(j)$ belongs to a Markov stochastic sequence, with statistics:

$$E\left[E(j)\right] = 0 \ , \quad E\left[E(j)E^T(j)\right] = P(j) \tag{3}$$

where $E[\cdot]$ is the expectation operator. The standard deviations correspondent to the diagonal variances of $P(j)$ are supposed to be related to an approximation of the order of magnitude of the true global error $\varepsilon(j)$ as shown in developments that follow. To start the process in t_0, this same reasoning is used to calculate $P(0)$ as a diagonal matrix.

In a typical step, if $T^{(p-1)}(\cdot)$ is the operator representing an one step numerical integrator of order (p-1), then it results:

$$x(t_{j+1}) = T^{(p-1)}(x(t_j)) + \varepsilon^p(j+1) \tag{4}$$

where $\varepsilon^p(j+1)$ is the error in the $(j+1)^{th}$ step, due to truncation and given by (Fehlberg, 1968):

$$\varepsilon^p(j+1) = T^p(x(t_j)) - T^{(p-1)}(x(t_j)) + e(j+1) \tag{5}$$

where the difference of integrators of consecutive orders gives an approximation to the upper bound of the error in each step. Control of step size is done to keep $\varepsilon^p(j+1)$ inside the tolerance limits of the accuracy asked for the results. Thus $e(j+1)$ is negligible, that is, it can be looked as a numerical zero. Since what is required is to have an error inside the limits of accuracy and since it does not matter where this error is inside these limits, as long as it is inside the limits, then it is reasonable to model e(j+1) as a white vector sequence of independent components, $e^m(j+1)$, as follows:

$$E[e^m(j+1)] = 0 \;, \quad Q_{ii}(j+1) = E[e_i^{m^2}(j+1)] =$$

$$= \frac{1}{100}(T_i^p(x(j)) - T_i^{(p-1)}(x(j))^2 \tag{6}$$

where i=1,2,...,n, n being the dimension of vector $x(t_j)$; and Tchebycheff inequality was used to guarantee that, independently of its distribution, $e_i^m(j+1)$ has realizations with 99% chance of being inside tolerance limits. Following the same reasoning, $x(t_{j+1})$ is modelled by $x^m(t_{j+1})$ given by

$$x^m(t_{j+1}) = T^{(p-1)}(x^m(t_j)) + \varepsilon^{pm}(j+1) \tag{7}$$

where

$$x^m(t_j) = x(j) + E(j) \tag{8}$$

$$\varepsilon^{pm}(j+1) = T^{(p)}(x^m(t_j)) - T^{(p-1)}(x^m(t_j)) + e^m(j+1) \tag{9}$$

resulting

$$x^m(t_{j+1}) = T^p(x^m(t_j)) + e^m(j+1)$$

which after a Taylor's expansion up to first order about x(j), disregarding terms of higher order, leads to:

$$x^m(t_{j+1}) = T^p(x(j)) + T^p_{x(t_j)}(x(j))\; E(j) + e^m(j+1) \tag{10}$$

where $T^p(x(j))$ gives the value of x(j+1) and $T^p_{x(t_j)}(x(j))$ is the result of numerically integrating the state transition matrix equation

$$\dot{\Phi}(t,t_j) = f_x(x,t)\cdot\Phi(t,t_j) \tag{11}$$

in parallel with the system of Equation 1, using the integrator $T^p(\cdot)$

with initial conditions:

$$\Phi(t_j, t_j) = I_n \;, \quad x(j)$$

Combining the definition of Equations 7 and 8 with the previous results, one gets:

$$x^m(t_{j+1}) = x(j+1) + E(j+1) \tag{12}$$

$$E\left[E(j+1)\right] = 0 \;, \quad E\left[E(j+1)E^T(j+1)\right] = P(j+1)$$

$$P(j+1) = \Phi(j+1,j) \;\; \Phi^T(j+1,j) + Q(j+1)$$

where Q(j+1) is as defined in Equation 6.

However, for any given realization of E(j), the term $T^p_{x(tj)}(x(j))\cdot E(j)$ can be viewed as resulting from the numerical integration of the variational equation

$$\dot{E} = f^p_x(x,t)E \tag{13}$$

where $f^p_x(x,t)$ is the matrix of first order partials evaluated in the numerically integrated trajectory. However, if E(j) is such to differ from the truncation error by only one order of magnitude, Equation 13 can be integrated using Euler's method. This leads to the representation of E(j+1) as belonging to the stochastic sequence characterized by:

$$E(j+1) = E(j) + h_j \cdot f_x(x(j),\; t_j) \cdot E(j) + e^m(j+1) \tag{14}$$

where h_j is the step size in t_j; and the associated covariance matrix given by:

$$P(j+1) = (I_n + h_j \cdot f_x(x(j),\; t_j))P(j)(I_n + h_j \cdot f_x(x(j),\; t_j))^T + Q(j+1) \tag{15}$$

3.2 - CONTRIBUTION OF ROUND-OFF ERRORS

In a typical interval of numerical integration, the input variables and the ones resulting from arithmetic operations are rounded off inside the computer. If floating arithmetic is used, it results that for any pair of variables w and v the errors due to round off obey the condition

(Dahlquist and Björck, 1974):

$$/fl(w \text{ op } v) - w \text{ op } v/ \leqslant /w \text{ op } v/ \cdot u \tag{16}$$

where fl(·) indicates floating arithmetic; "op" indicates any of the four elementary operations; and u is the machine unit (round off unit). Thus, for a given problem, the round off error in each step strongly depends upon the particular algorithm used.

In the case considered in this paper, the Runge-Kutta formulas presented by Fehlberg (1968) are used. Assuming round off errors to be at most of the order of magnitude of local truncation errors, it results

$$x(j+1) = fl(x(j) + fl(h_j \cdot fl(\sum_{k=0}^{K} c_k f_k))) = \bar{x}(j+1) + \varepsilon_x(j+1) \tag{17}$$

where the over bar is to indicate numerical values obtained after rounding; the c_k and f_k as in Fehlberg (1968); and for each component $x_i(j+1)$, i = 1,2,...,n, the error can be limited as follows (Dahlquist and Björck, 1974):

$$/\varepsilon_{x_i}(j+1)/ \leqslant \left\{\bar{h}_j \cdot \sum_{k=0}^{K} /(K+2-k) \cdot \bar{c}_k \cdot \bar{f}_{i_k} / + 2\bar{h}_j \cdot / \sum_{k=0}^{K} \bar{c}_k \cdot \bar{f}_{i_k} / + /\bar{x}_i(j)/\right\} \cdot 1.06\, u =$$

$$= r_{ii}(j+1)$$

Taking the approach of also modelling this error as a white vector sequence of independent components, under conditions similar to those taken for the truncation errors, results:

$$E[\varepsilon^m_{x_i}(j+1)] = 0, \quad E[\varepsilon^{m^2}_{x_i}(j+1)] = R_{ii}(j+1) = \frac{1}{100} r^2_{ii}(j+1) \tag{18}$$

and whenever its order of magnitude is significative relative to truncation error, its contribution is superposed to E(j+1) leading to a global error represented by:

$$E^t(j+1) = E(j+1) + \varepsilon^m_x(j+1) \tag{19}$$

and

$$P^t(j+1) = P(j+1) + R(j+1) \tag{20}$$

where the i^{th} diagonal variance of $P^t(j+1)$ gives a measure of the order of magnitude of the accumulated global error associated to $\bar{x}(j+1)$.

4. PRELIMINARY TESTS

The preliminary tests were done using Fehlberg's RK7-8 formulas (Fehlberg, 1968)。 The test case selected was one of those suggested by Krogh (1970), corresponding to a restricted three body problem, that models the motion of a satellite moving under the influence of the Earth and the Moon. The equations of motion in the rotating (synodic) coordinate system, using dimensionless quantities are (Szebehely, 1967):

$$x_1'' = 2x_2' + x_1 - \mu'(x_1 + \mu)/r_1^3 - \mu(x_1 - \mu')/r_2^3 \qquad (21)$$

$$x_2'' = -2x_1' + x_2 - \mu' x_2/r_1^3 - \mu x_2/r_2^3$$

with initial conditions (Krogh, 1970):

$$x_1(0) = 1.2 \ , \quad x_1'(0) = 0 \ , \quad x_2(0) = 0$$

$$x_2'(0) = -1.04935\ 75098\ 30319\ 90726\ \ldots$$

where

$$r_1 = ((x_1+\mu)^2 + x_2^2)^{1/2} = \text{distance from the Earth}$$

$$r_2 = ((x_1-\mu')^2 + x_2^2)^{1/2} = \text{distance from the Moon}$$

$$\mu = 1/82.45 \ , \quad \mu' = 1 - \mu$$

The initial conditions taken give an orbit with period $T = 6.19216\ 93313\ 19639\ 70674\ \ldots$ providing a mean of calculating the true global error at the end of each period.

TABLE 1. ESTIMATED ERROR STANDARD DEVIATIONS/TRUE GLOBAL ERRORS

TIME	$\sigma_1(t)/\varepsilon_1(t)$	$\sigma_1'/\varepsilon_1'(t)$	$\sigma_2(t)/\varepsilon_2(t)$	$\sigma_2'(t)/\varepsilon_2'(t)$
T	.186113855E-08/ .363797881E-09	.436232905E-08/ .241853322E-09	.252317107E-08/ .658361999E-09	.191675823E-08/ .378349796E-09
2T	.190382914E-08/ .436557457E-09	.890784686E-08/ .158197366E-08	.409984179E-08/ .446576109E-11	.239164388E-08/ .378349796E-09
3T	.239409823E-08/ .145519152E-08	.100865133E-07/ .112227905E-08	.453268581E-08/ .288094744E-08	.279354623E-08/ .119325705E-08
4T	.240477181E-08/ .291038305E-10	.106932148E-07/ .190822637E-08	.774237305E-08/ .504425926E-08	.311185527E-08/ .698491931E-09
5T	.284394204E-08/ .727595761E-09	.105259284E-07/ .264338488E-08	.100359916E-07/ .111394019E-08	.355154685E-08/ .436557457E-09
6T	.340670801E-08/ .139698386E-08	.132512009E-07/ .249632562E-08	.107424875E-07/ .809652178E-09	.400246841E-08/ .190630089E-08
7T	.360552234E-08/ .222644303E-08	.187472581E-07/ .616642697E-08	.107717040E-07/ .282498378E-08	.433280095E-08/ .275031198E-08
8T	.382770053E-08/ .116415322E-08	.225626169E-07/ .111531063E-07	.111101079E-07/ .341051678E-08	.456368641E-08/ .324507710E-08
9T	.364301863E-08/ .157160684E-08	.241609986E-07/ .134607525E-07	.137799294E-07/ .203140615E-08	.456978366E-08/ .510772225E-08
10T	.385400056E-08/ .291038305E-10	.241907691E-07/ .140615121E-07	.170231611E-07/ .308867393E-08	.474949470E-08/ .500585884E-08

Results of Table 1 were obtained in a Burroughs 6800 using control of step-size and the exigence of a local accuracy in each component not worse than 1.E-10 (relative error). To propagate the error covariance matrix, the state transition matrix differential equations were also integrated with the RK7-8 formulas. Results tried with Euler's method showed to be too much conservative.

5. CONCLUSIONS

A new scheme to get global error estimates was presented. The objective is to have a valid tool in the evaluation of quality of numerical results in the simulation of dynamic systems. The preliminary tests done are only illustrative and, of course, very dependent upon the problem chosen. Before the stochastic scheme can be considered as properly qualified for general use, an exhaustive testing is necessary. A price to be paid by its users is the need of also having to numerically integrate the first order variational equations associated to the system simulated. Future developments should pay special attention to the possibility of getting valid approximations which lead to a decrease in computer time spent in the numerical treatment of these equations.

6. REFERENCES

Dahlquist, G.; Björck, A.: 1974, *Numerical methods*. Englewood Cliffs, NJ, Prentice-Hall.

Fehlberg, E.: 1968, "Classical fifth-sixth-seventh and eighth order Runge-Kutta formulas with step-size control". NASA TR R-287.

Krog, F.T.: 1970, "On testing a subroutine for the numerical integration of ordinary differential equations". JPL Section 314, Technical Memorandum nº 217.

Prothero, A.: 1980, "Estimating the accuracy of numerical solution to ordinary differential equations".in *Computational techniques for ordinary differential equations*, I. Gladwell and D.K.Sayers, eds.,Academic Press, London p. 103.

Zadunaisky, P.E.: 1979, "On the accuracy in the numerical solution of the N-body problem". Celes. Mech., *20*, 209.

Szebehely, V.: 1967, *Theory of orbits*, Academic Press, New York.