Interval Computations
No 3(5), 1992

ELLIPSOIDAL ERROR ESTIMATES FOR ADAMS METHOD

Alexey F.Filippov

An error that arises in the approximated solution of a system of differential equations due to the errors of procedure and the rounding errors is estimated by the method of ellipsoids. The sources of errors and the results of the method application to the system of two equations on a large time interval are analyzed.

ЭЛЛИПСОИДАЛЬНЫЕ ОЦЕНКИ ОШИБКИ ПРИБЛИЖЕННОГО РЕШЕНИЯ, ПОЛУЧАЕМОГО МЕТОДОМ АДАМСА

А.Ф.Филиппов

Ошибка приближенного решения системы дифференциальных уравнений, возникающая из-за ошибок метода и ошибок округления, оценивается методом эллипсоидов. Анализируются источники ошибок и результаты применения метода к системе двух уравнений на большом интервале времени.

For estimating errors simultaneously with a solution computing, a time-variable ellipsoid is constructed, whose center is a point representing an approximated solution found at a given moment of time. This ellipsoid contains the exact solution of the system. At each step of the approximated method, the variation of the ellipsoid in the motion over the trajectories of the given system and the influence of errors are taken into account. The ellipsoid method was applied in [1] to estimate accessible sets of a control system. It does not lead to the Moore effect that often arises in coordinate-wise estimates.

A detailed exposition of the method is contained in [2]. Here, only the basic formulas are presented, a particular example is considered, and the influence of different errors on the growth rate of the total error estimate is investigated.

Let $z(t_i)$ be approximate values of a solution of the following problem, obtained at points $t_i = t_0 + ih,\ i = 1, 2, \ldots$

$$x' = f(t, x), \quad x(t_0) = x^0 \quad (x \in R^n). \tag{1}$$

An estimate of the difference $x(t_i) - z(t_i) = w(t_i)$, where $x(t)$ is an exact solution of problem (1), is required. Let $x_{i-1}(t)$ be an exact solution of equation (1) with the initial condition $x_{i-1}(t_{i-1}) = z(t_{i-1})$, let $\Delta_i = z(t_i) - x_{i-1}(t_i)$ be a local error due to error of procedure and rounding errors. We have

$$w(t_i) = y(t_i) - \Delta_i \qquad (y(t) \equiv x(t) - x_{i-1}(t)), \tag{2}$$

$$y'(t) = f(t, x_{i-1}(t) + y(t)) - f(t, x_{i-1}(t)) = C(t)y(t) + \psi(t), \tag{3}$$

where $C(t)$ is a matrix of derivatives $\partial f_j / \partial x_k (j, k = 1, \ldots, n)$, taken for $x = x_{i-1}(t)$, and $\psi(t)$ is the residual of the Taylor formula.

Given the estimates $\Delta_i \in E(0, M_i)$, $\psi(t) \in E(0, B(t))$. The ellipsoid $E(a, Q)$, here, as in [1], for $a \in R^n$ and symmetric matrix Q (such that $x \cdot Qx \geqslant 0$ for all $x \in R^n$), is a set of points $x = a + Qy$ such that $y \cdot Qy \leqslant 1$.

Let it is known that $w(t_0) \in E(0, Q_0)$. We show how from the estimate $w(t_{i-1}) \in E(0, Q_{i-1})$ to pass to the estimate $w(t_i) \in E(0, Q_i)$, that is how to find Q_i on the basis of the known Q_{i-1}. By [1], §8, (3) implies

$$y(t) \in E(0, Q(t)) \quad (t_{i-1} \leqslant t \leqslant t_i), \tag{4}$$

$$Q' = CQ + QC^{\top} + qQ + q^{-1}B(t), \quad Q(t_{i-1}) = Q_{i-1}, \tag{5}$$

where $C^{\top}$ is the transpose of C, the number q can be taken to be near $\sqrt{n^{-1}\operatorname{Tr}(Q^{-1}B)}\Big|_{t=t_{i-1}}$. Let, instead of the exact $Q(t_i)$ from (5), an approximated Q^* be found with the error estimate $||Q^* - Q(t_i)|| \leqslant \rho$ (any norm). Then, instead of (4), for $t = t_i$, we have

$$y(t_i) \in E(0, Q_\rho), \qquad Q_\rho = Q^* + \rho I, \tag{6}$$

I is the identity matrix. By (2) and the estimates for Δ_i and $y(t_i)$,

$$w(t_i) \in E(0, Q_\rho) + E(0, M_i) \subset E(0, Q_i);$$

since Δ_i is small with respect to $y(t_i)$, with accordance to [1], §6,

$$Q_i = (1+p)Q_\rho + (p^{-1}+1)M_i + n\mu I, \; p \approx \sqrt{n^{-1}\mathrm{Tr}(Q_\rho^{-1}M_i)}. \qquad (7)$$

The term of the sum $n\mu I$ compensates rounding errors in the first of the formulas of (7) if these errors do not exceed the absolute value of μ in every element of the matrix Q_i. We need not take into account the errors in computing p (any p would be appropriate). Formulas (5), (6) and (7) determine the passage from Q_{i-1} to Q_i.

We shall show a way of constructing the matrix $B(t)$ in (5). In (3), if for each coordinate f_j of the vector-function $f(t,x)$ we have

$$|\partial^2 f_j / \partial x_k \partial x_l| \leqslant m_{jkl}, \; m_j = \left(\sum_{k,l=1}^{n} m_{jkl}^2 \right)^{1/2},$$

then the vector $\psi(t) = (\psi_1(t), \ldots, \psi_n(t))$ is estimated as follows ($|y|^2 = y_1^2 + \ldots + y_n^2$):

$$|\psi_j(t)| \leqslant \frac{1}{2} \sum_{k,l=1}^{n} m_{jkl}|y_k y_l| \leqslant \frac{1}{2} m_j |y|^2.$$

Then $\psi(t) \in E(0, B(t))$, where $B(t)$ is a matrix with elements

$$b_{ij}(t) = 0 \quad (i \neq j), \quad b_{jj}(t) = (n/4)|y(t)|^4 m_j^2. \qquad (8)$$

By (4), $|y(t)|$ is no greater than the greatest semiaxis of the ellipsoid $E(0, Q(t))$; therefore,

$$|y(t)|^2 \leqslant \lambda_{\max}(Q(t)) \leqslant \mathrm{Tr}Q(t).$$

Thus, in (5), $B(t)$ is of order $||Q||^2$, q is of order $||Q||^{1/2}$, and two last terms in (5) causing the growth of the ellipsoid due to the accounting $\psi(t)$ in (3) have the order $||Q||^{3/2}$. Therefore, the relative rate of error growth depending on $Q'/||Q||$, is small for small $||Q||$, and increases when $||Q||$ increases.

Sometimes, it is appropriate that equation (5) would be linearized by replacing $b_{jj}(t)$ in (8) by constant on each interval (t_{i-1}, t_i) or linear functions of entries of the matrix $Q(t)$.

The estimating of a local error Δ_i is proceeded as in [3] or [4], but in contrast to [4], the vectors and the matrix $(\partial f_j/\partial x_k)$ should be evaluated in norm. A rounding error is estimated as in [4], but separately for each coordinate of the vector. The roughness of the estimate in [3] of higher derivatives of a solution can be compensated by the choice of the step h sufficiently small; this will cause the increase of the rounding error, which is not dangerous in the computations with sufficiently large number of binary digits. Another way to estimate derivatives: to partition the domain into parts and, in the parts where the solution goes through, to estimate successively the derivatives $x' = f(t,x)$, $x'' = f'_t + (f'_x)f, \dots$ from above and from below, by interval computations approaches.

Values of an approximated solution at several initial points $t_1, t_2, \dots$ that are necessary for the Adams method, can be obtained by an one-step method with a lesser step h_1 for example, by the second order Adams interpolation method or by the fourth order Euler method:

$$x(\bar{t}) = x(t) + \frac{h_1}{2}(x'(t) + x'(\bar{t})) + \frac{h_1^2}{12}(x''(t) - x''(\bar{t})), \qquad (9)$$

where $\bar{t} = t + h_1$, $x'(t) = f(t, x(t))$, $x''(t) = f'_t + (f'_x)f$, and the error of procedure is equal to $-(h_1^5/720)x_j^{(5)}(t_j^*)$, $t < t_j^* < \bar{t}$ (separately for every coordinate x_j of the vector x).

Since the matrix $C(t)$ is known at mesh points t_i only, system (5) can be resolved, for example, by the Taylor formula

$$Q(t_i) = Q(t_{i-1}) + hQ'(t_{i-1}) + (h^2/2)Q''(t_{i-1})$$

with the error $-h^3Q'''/6$. When estimating $C(t_i - 0)$ via $x_{i-1}(t_i) = z(t_i) - \Delta_i$ with the compensation of the error by the increase of the matrix $B(t)$ in (5), the method of (9) with $h_1 = h$ can be applied to system (5).

The above method was applied to the system

$$x' = x - y, \quad y' = 2x - y^3; \quad x(0) = 0.25, \quad y(0) = 0.$$

The solution is contained in the domain $|x| < 1$, $|y| < 1.22$, $|x-y| < 0.9$. The 8-th order interpolating Adams method with $h = 1/256$ was used, values of a solution for $t \leqslant 6h$ were computed on each interval (t_{i-1}, t_i), the linearization error was covered by the increasing of b_{jj} in (8). This system was solved by the method (9). The computation was performed with 56 binary digits.

The solution tends rapidly to a limiting cycle with period about 7.29. The ellipse that contains an exact solution extends along trajectories, its large axis a, alternately, augments and diminishes. The following table presents the values of a obtained, that is, the upper bound of the error of the found approximated solution, at some intersection points of the trajectory with the x-axis.

t	9.16	23.74	38.31	52.89	67.46	82.04	96.61
$10^{13}a$	5.7	20.0	35.5	52.1	69.8	89.0	110.0

On the segment $0 \leqslant t \leqslant 100, \max a \leqslant 1.2 \cdot 10^{-11}$; it is attained for $t = 96.8$.

The following factors are influencing the growth of the quantity a : the local error Δ in solving system (1), the estimate of the term $\psi(t)$ in (3), the error of procedure of the approximated method for system (5), the deviation of trajectories of system (1). The method is admissible on large time intervals t, provided the quantity a remains small.

References

1. Chernousko, F. L. *Estimating of a phase state of dynamical systems. Ellipsoid method.* Nauka, Moscow, 1988 (in Russian).
2. Filippov, A. F. *Ellipsoidal estimates for a solution of a system of differential equations.* Interval Computations **2** **(4)** (1992), pp. 6–16.
3. Stewart, N. F. *Compatible, guaranteed local error bounds for the Adams method.* Math. Nachr. **60** (1-6) (1974), pp. 145–153.
4. Filippov, A. F. *Computer-obtaining strong estimates for solutions differential equations.* J. Comp. Math. and Math. Phys. **31** (7) (1991), pp. 994–1005 (in Russian).