ERRORS IN NUMERICAL INTEGRATIONS AND CHAOTIC MOTIONS.

A. Milani and A.M. Nobili
Universita' di Pisa
Dipartimento di Matematica
Piazza dei Cavalieri 2
I-56100 Pisa Italy

ABSTRACT. The methods to estimate the integration errors, including the effects of truncation, rounding off, and instability of the solutions, are discussed. Polynomial error accumulation depends upon numerical method, stepsize, orbital period and also eccentricity; it is also machine dependent. Comets correspond to the most difficult case of exponentially diverging orbits; however they can be very close to resonant ordered regions.

Numerical integration is an essential tool in the study of cometary orbits; however the results of the integration are not always reliable. Integration error is a complex phenomenon; it is not a purely numerical effect, but the result of a complex interaction between the approximations introduced in the computation and the physical instability of the real orbit.

In this paper we review the possible causes of integration error, and try to give explicit estimates of their size, both for the local error (i.e. within one integration step) and for the accumulated error. For the sake of this discussion, we shall distinguish among the possible causes of error the truncation of the discretisation formula, the rounding-off of the numbers in the computer arithmetic unit, the errors due to the use of implicit formulae, the physical model errors and the errors in the initial conditions.

Truncation errors are conceptually the best understood; nevertheless practically useful formulae to estimate the order-of - magnitude of both the local and the propagated truncation errors really applicable to the orbits of celestial bodies are not available in the standard literature. In section 1 we try to fill this gap.

A. Carusi and G. B. Valsecchi (eds.), Dynamics of Comets: Their Origin and Evolution, 215–226.

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

Rounding-off is difficult to study, because there is no way to understand the process underlying the accumulation of this kind of errors without a deep understanding of the way the arithmetic unit of the computer works and of the way a high level language code is translated into machine code. In section 2 we review briefly some recent improvements in the understanding of the machine - dependent round - off errors.

Implicit formulae are always used together with a control of their convergence, therefore the contribution to the integration error from their imperfect convergence is always minor and anyway well under the control of the programmer. Sometimes it is difficult to explain why the errors arising from this source are so small, as was discussed in this meeting by E. Everhart: we had the same experience with our implicit Runge-Kutta method.

The physical model errors will be discussed elsewhere in these proceedings. We are left with the error in the initial conditions: strictly speaking it is not an error arising within the process of computing the orbit; however the way it accumulates as time goes by is relevant for our discussion of the integration errors, for the very simple reason that every error results in the displacement of the computed orbit on a nearby orbit whose initial conditions were different. Thus the instability of the real, physical orbit does introduce a numerical instability as well, and the "numerical" error cannot accumulate slower that the rate of growth of the separation of two nearby orbits. This kind of instability is, unfortunately for us, specially relevant for cometary orbits, as it is discussed in section 2.

1. THE TRUNCATION ERROR

The local truncation error is the difference between the actual orbital motion between time t and time t + h and the solution of the discretized problem actually solved in the computational algorithm for the corresponding step; it is often estimated with a formula based only on the product hn (n is the mean motion); however every such formula is valid for circular orbits only. A more general formula valid for nonzero eccentricities is given here.

Let us assume the integration is performed with a Stormer predictor:

$$\underline{x}_{K+1} = 2\underline{x}_K - \underline{x}_{K-1} + h^2 \sum_{j=0}^{m} b_j \nabla^j \ddot{\underline{x}}_K \tag{1}$$

where ∇ is the backward difference operator ∇ f(t)=f(t)-f(t-h). Since formula (1) is obtained by truncation of the summation to order m, the local truncation error is:

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

$$R_{m+1}=b_{m+1}\, h^{2}\, \nabla^{m+1} \ddot{\underline{x}}_{k} + \dots \qquad (2)$$

and by the Lagrange formula:

$$\nabla^{j} f(t)=h^{j} f^{(j)}(t^{*}) \ , \qquad t - jh < t^{*} < t \qquad (3)$$

$$R_{m+1}=b_{m+1}\, h^{m+3}\, \underline{x}^{(m+3)}\,(t\ )+\dots \qquad (4)$$

If the exact solution were a circular orbit with mean motion n, then for m even (m=2y):

$$\underline{x}^{(m+3)}=(-1)^{y}\, n^{m+2}\, \dot{\underline{x}} \qquad (5)$$

and for m odd (m=2y+1):

$$\underline{x}^{(m+3)}=(-1)^{y}\, n^{m+3}\, \underline{x} \qquad (6)$$

and the magnitude of the local error is:

$$\left| R_{m+1} \right| = b_{m+1}\,(hn)^{m+3}\, a \qquad (7)$$

(a the semimajor axis). However we are mainly interested in the propagation of this local error; this phenomenon can be studied in different ways (Kinoshita, 1968; Henrici, 1962) but the most natural way for astronomers is to use Gauss perturbation equations: the local error (4) can be interpreted as the effect of a constant perturbing force F :

$$F = b_{m+1}\, h^{m+1}\, \underline{x}^{(m+3)}\,(t^{*})+\dots \qquad (8)$$

The resulting perturbations on the orbital elements will depend on the direction as well as the magnitude of the force F: the main effect, the one growing quadratically with the time, will be the perturbation in longitude arising from a secular perturbation in the semimajor axis. All the other perturbative effects will result in much smaller accumulated errors; the short term perturbations will produce errors of the order of

$$F/n^{2} \simeq b_{m+1}\,(hn)^{m+1}\, a \qquad (9)$$

The Gauss equation for the semimajor axis is:

$$\dot{a} = 2\langle F,\dot{\underline{x}}\rangle/n\ a + \text{terms of order} \geqslant 1 \text{ in } e. \qquad (10)$$

For m even (m=2y) we will have for a circular orbit:

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

$$\dot{a}(t) = 2b_{m+1}(-1)^{y}(hn)^{m+1} < \dot{\underline{x}}(t),\dot{\underline{x}}(t^{*})>/na +... \qquad (11)$$

where t^{*} is a few steps before t, by (3); the angle between $\dot{\underline{x}}(t)$ and $\dot{\underline{x}}(t^{*})$ is thus of the order of mhn/2 and:

$$< \dot{\underline{x}}(t),\dot{\underline{x}}(t^{*})> = n^{2}a^{2}\ (1 + O(n^{2}h^{2}m^{2}/4)) \qquad (12)$$

We can assume the O(...) term in (12) to be of higher order, and by substituting in (11):

$$\dot{a}/a = 2b_{m+1}(-1)^{y}(hn)^{m+1}n +.... \qquad (13)$$

the corresponding coefficient of the quadratic error accumulation in longitude is then:

$$\dot{n}/2 = -1.5\ b_{m+1}(-1)^{y}(hn)^{m+1}n^{2}+... \qquad (14)$$

It is worth remarking that a different formula holds for m odd (m=2y+1): because of (6) the direction of the "truncation perturbing acceleration" F is more radial than along track, and from a formula analogous to (12):

$$< \dot{\underline{x}}(t),\underline{x}(t^{*})> = na^{2}\ O(\ n\ h\ m/2) \qquad (15)$$

we find that (13), (14) are modified by a factor O(nhm/2): the secular effect appears to be of higher order; however in most integrations nh is not much smaller than 2/m. The results given by formulae (9) and (14) do not change if another integration method is used; only the constant b_{m+1} does change.

It is intuitively obvious that the error estimates given by (9), (13) and (14) give a grossly underestimated error for eccentric orbits; however how fast the error grows with eccentricity is somewhat surprising. There is a way to compute an estimate of the truncation error in the integration of an eccentric orbit just by recomputing formulae (5), (6). For an eccentric orbit the exact solution, in an appropriate coordinate system xyz, can be expanded in a Fourier series in the mean anomaly M with coefficients formed by Bessel functions (Wintner, 1941; Kovalevski, 1963):

$$\begin{aligned} x/a &= -3e/2 + \sum_{p=1}^{+\infty} 1/p\left\{J_{p-1}(pe) - J_{p+1}(pe)\right\}\cos pM \\ y/a &= (1-e\ ) \sum_{p=1}^{+\infty} 1/p\left\{J_{p-1}(pe) + J_{p+1}(pe)\right\}\sin pM \end{aligned} \qquad (16)$$

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

These series have the D'Alembert property, that is the lowest order term of the coefficient of the cosine (or sine) of pM is $O(e^{p-1})$; however, when the derivatives of order m+3 are computed, e^{p-1} is multiplied by an high power of p:

$$R_{m+1} = b_{m+1}\, a\, (hn)^{m+3} \sum_{p=1}^{+\infty} g(e,p) \text{ trig } (pM) \tag{17}$$

with trig (pM) a trygonometric vector function of length 1; the coefficients g is:

$$g(e,p) = \frac{e^{p-1}\, p^{p+m+2}}{2^{p-1}\, p!} + O(e^{p+1}). \tag{18}$$

The truncation error (17) is then the sum of different harmonic components; at the perihelion all the error harmonics are in phase, and the size of the local error can be estimated by the sum of the series

$$S_1 = \sum_{p=1}^{+\infty} g(e,p) \sim Z_1 = \sum_{p=1}^{+\infty} \frac{e^{p-1}\, p^{p+m+2}}{2^{p-1}\, p!} \tag{19}$$

The accumulated along-track error however is not the result of the error at a specific point but rather of the average error in energy; for the purpose of an estimate of the along-track quadratic error one should consider the different error harmonics as independently acting, thus use the root mean square sum of their size:

$$S_2^2 = \sum_{p=1}^{+\infty} g^2(e,p) \sim Z_2^2 = \sum_{p=1}^{+\infty} \left(\frac{e^{p-1}\, p^{p+m+2}}{2^{p-1}\, p!} \right)^2 \tag{20}$$

and substitute the resulting "average" energy error in the same formulae used to obtain (14):

$$\dot{n}/2 = -1.5\, b_{m+1} (-1)^{Y} (hn)^{m+1} Z_2\, n^2 + \ldots \tag{21}$$

In table 1 the prediction given by formula (21) is compared with the results of a test performed by Cohen et al. (1973) with m=12, h=40 days, n = mean motion of Jupiter. The error is an along track acceleration in arcsec/year^2 .

The comparison shows that our formula gives the right value of the along track error for low eccentricities, and a good order - of - magnitude estimate for moderate values of e. Of course the use of Z_2 instead of S_2 results in inaccuracies for large e because the neglected higher order (in e) terms are not much smaller. The Cohen et al. (1973) test were performed as a preparation for a long integration of planetary orbits; the effect on cometary orbits of the same phenomenon is striking. In table 2 we have listed the predicted increase of the error with respect to a circular orbit as a function of e, both for the local error

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

TABLE 1

e	error predicted by (21)	error found in the test integration
0	-0.4×10^{-12}	
.01	-1.8×10^{-12}	-1.6×10^{-12}
.02	-5.3×10^{-12}	-5.7×10^{-12}
.03	-12.7×10^{-12}	-12.1×10^{-12}
.04	-26.5×10^{-12}	-24.3×10^{-12}
.05	$-52.\ \times10^{-12}$	$-43.\ \times10^{-12}$
.06	$-97.\ \times10^{-12}$	$-68.\ \times10^{-12}$

at perihelion and for the accumulated along-track quadratic error. It can be appreciated that the error grows very fast with e, much faster than the cubic growth hypothesized by Cohen et al.; for cometary orbits, this implies that the use of a fixed stepsize algorithm is not recommended and the use of a short stepsize is not a good solution.

TABLE 2

e	increase of local error at perihelion	increase in accumulated error
.05	3×10^{2}	1×10^{2}
.1	5×10^{3}	2×10^{3}
.15	6×10^{4}	2×10^{4}
.2	6×10^{5}	2×10^{5}
.25	6×10^{6}	2×10^{6}
.3	6×10^{7}	2×10^{7}
.35	7×10^{8}	2×10^{8}
.4	9×10^{9}	2×10^{9}

Of course a solution to the problem of the increase of the error with eccentricity is the use of a time element s such that ds/dt is proportional to 1/r (r=distance from the Sun); then the time element is essentially the eccentric anomaly E and formulae like (16) are substituted by $x=a(\cos E-e)$, $y=a(1-e^{2})^{1/2}\sin E$ that do not contain the higher harmonics. With a time element proportional to the eccentric anomaly, the local and the accumulated along track error are still given by (9) and (14) respectively, for every eccentricity e.

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

2. THE ROUNDING-OFF ERROR

The rounding-off error is usually discussed with reference to the treatment by Brouwer (1937). However that celebrated paper was written before the era of the electronic computers, and thus was based on assumptions which are not necessarily applicable to the process of orbit computation as it is usually performed today.

The hypotheses under which Brouwer theorem applies are as follows: A) the error (or at least the significant part of it) is done in summing up the values of the previous second derivatives, or their differences, multiplied by h^2, to the previous step to get the new value, as in (1). The round-off errors done within the computation of the acceleration at each step are of lesser importance, because h^2 is small. B) the local round off error can be modelled as a random variable, uniformly distributed between $-d/2$ and $d/2$, where d is the "machine precision", i.e. the value of the last bit, or 1 in the last recorded digit; in particular its mean, or expected, value is exactly zero. C) the orbit is in itself orbitally stable, e.g. because the perturbations are negligible. D) other errors are uninfluent, e.g. the truncation error is smaller (in the manual variable order computation algorithms used at the time, this was checked at each step anyway).

Under these assumptions, Brouwer proved that the error in the orbital elements will be distributed as a gaussian random variable, with zero mean and root mean square value growing with the number N of integration steps as $N^{3/2}$ for the anomaly (i.e. along track) and as $N^{1/2}$ for the other five (i.e. cross track). The problem is now to assess how valid are the assumptions A, B, C and D: In particular B requires that the rounding off is performed by computing the sum in (1) with all the significant digits conserved, and then rounded: this is not at all the way the arithmetic operations are performed in modern computers, or better: not at all the way in which the compilers instruct the arithmetic unit of the CPU to operate. In reality, numbers are usually truncated, just forgetting the significant digits that are on the right of the maximum allowable mantissa length. As a result, the "rounf-off" error has an average not equal to zero but to $-d/2$; the latter formula being exactly true only for fixed point arithmetic and provided the negative numbers are represented in complement: the sign of the coordinates modifies the expected error if the negative numbers are represented as modulus plus sign. As a result, the same "random walk" argument used by Brouwer gives an expected secular error in semimajor axis; the expected error along track grows as N^2, for the fixed point, modulus plus sign arithmetic (Fabri and Penco, 1984). Different results are obtained for different arithmetics; floating point arithmetics generates a "quantum

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

effect" by which the relative error depends upon the absolute value of the semimajor axis.

Fabri and Penco have proved rigorously (at least for circular orbits) the existence of errors growing faster than Brouwer's rule by questioning only assumption B: the round-off error is still modelled in a statistical way, but with a realistic distribution (of course the computer errors are not random at all, being uniquely determined by the algorithm and by the initial conditions: the statistic refers to integrations with different initial conditions). However the other assumptions seem questionable as well. We will discuss assumption C in section 3. Assumptions A and D more or less mean that when different error sources are present, the local error will propagate according to the propagation trend of the larger local error: e.g. if a rounding-off error propagating as $N^{3/2}$ (because assumption B is valid; e.g. by the use of guard-digits) is larger than the quadratically-propagating truncation error, the latter is "masked" and is not allowed to grow at its own pace. There are no rigorous arguments to support this way of thinking. We conclude that, unless special attention is paid not only to the integration algorithm but even to the computer firmware, the integration error is bound to grow at least quadratically with time.

3. ERRORS IN CHAOS

We have at last to question the assumption upon which most of the discussions of the numerical errors, including the two previous sections, are based: that the errors done in propagating an unperturbed Keplerian orbit are representative of the errors that would result from the integration, with the same method, of a perturbed orbit.

As it is known since the times of Poincare', the perturbed motion does not belong to an integrable system; this means that some orbits will lie on manifolds of quasi-periodic solutions (Arnold,1963), but in between these there will be chaotic regions where homoclinic orbits generate hyperbolic sets (Smale, 1967). The technical definition of an hyperbolic set does not matter here; the essential property of a chaotic region containing an hyperbolic set can be described with the device first used by Henon and Heiles (1964) for their numerical investigations of non-integrable dynamical systems: if two orbits with initial conditions very close together are propagated (e.g. numerically), the distance between points corresponding to the same time will grow exponentially with time, and the logarithm of the ratio between the initial distance and the distance after some fixed time T will be a measure of the instability. Indeed this same logarithm divided by T is related to the Liapounov characteristic exponent (Benettin et al., 1980).

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

What about the errors in the numerical orbit propagation when the real orbit lies in such a chaotic region? A very simple, and often overlooked, result says that no numerical method can be more stable than the exact solution the method is used to compute. Because does not matter how small is the local integration error: since it is anyway nonzero, after the first step the numerical method will really integrate an orbit starting from different initial conditions: if the latter diverges exponentially from the exact solution, so will the numerical solution (if it is "convergent", i.e. unless it really solves an other equation).

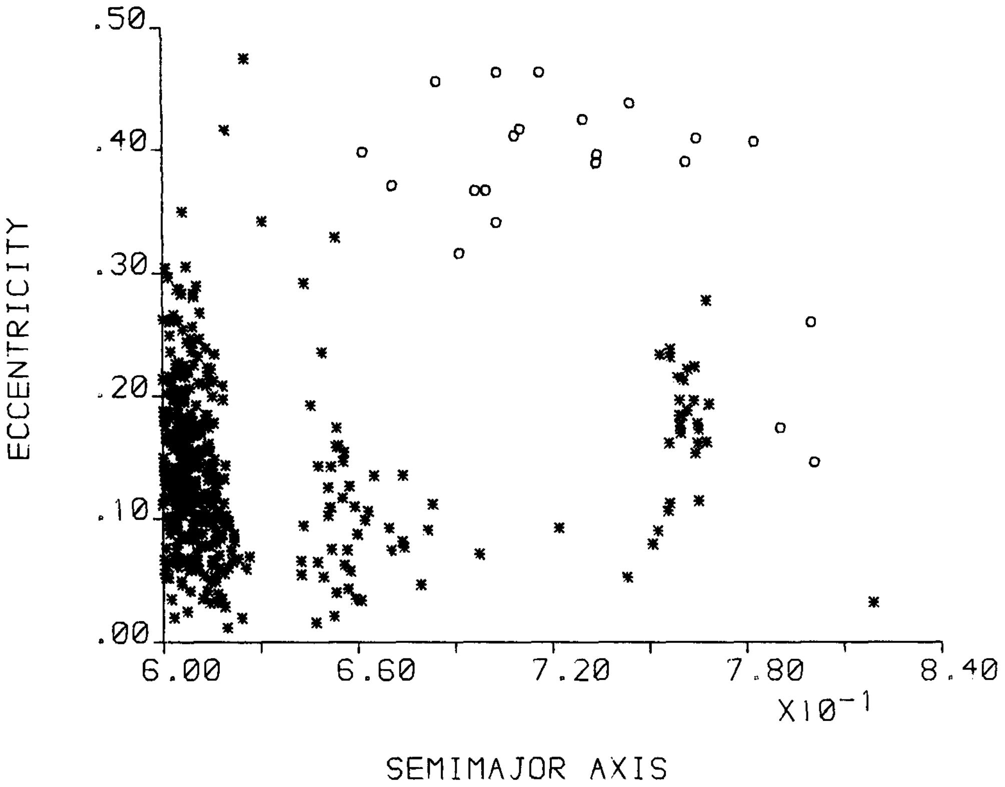

Figure 1: Numbered asteroids (stars) from the TRIAD file and periodic comets (circles) from the catalogue of perihelion passages by Marsden and Roemer, 1982. Semimajor axis (as a fraction of Jupiter's) and eccentricity are plotted for the region 3.2AU<a<4.3AU, 0<e<0.48. The region that appears void in this plot, between the Hildas and the 2/1 gap and for moderate eccentricities is indeed occupied occasionally by comets on temporary "transfer" orbits.

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

Cometary orbits have the special feature of (almost) always lying in chaotic regions. This is in reality an observational selection effect: a comet can exhibit a spectacular coma only provided that its orbit "recently" underwent a change resulting in a large decrease of its perihelion distance; this change results from some strong perturbation, and the strong perturbation in turn generates chaotic behaviour. Because of the very complex structure of the resonances with the major planets, the regions of the phase space where both chaotic behaviour and abrupt orbital elements changes can occur are intermingled with "ordered" regions of dominant quasi-periodic behaviour. This is best illustrated by the boundary region between the outer asteroid belt and the belt of the comets of the Jupiter family (Figure 1).

In between there is a "grey" belt of orbits that are neither obviously cometary nor asteroidal: the best criterion to predict whether a given set of initial conditions will give rise to an ejection from the region plotted in Figure 1 (hence is "cometary" even if in a transient almost quiescent state) is to compute the divergence ratio as described above (Milani and Nobili, 1984b).

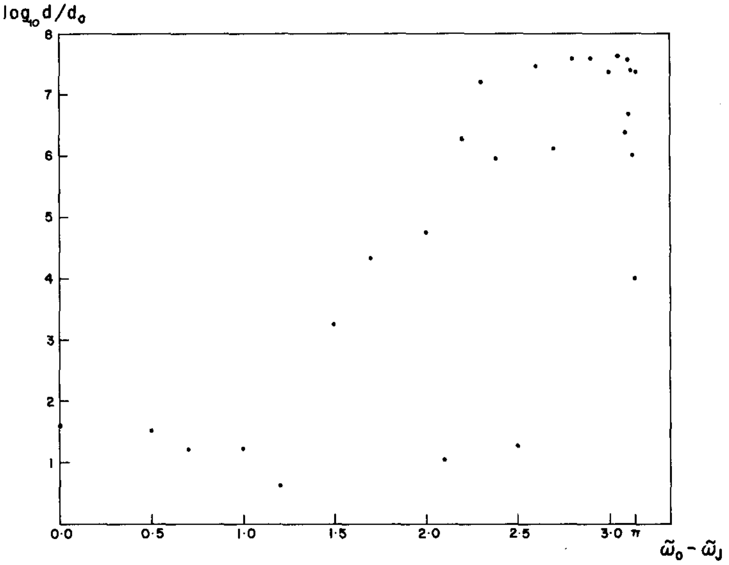

Figure 2: The divergence, i.e. the ratio between the very small initial distance d_0 and the distance d after 50 synodic periods, is plotted for the orbit of a comet with initial conditions a=0.68 a_J , e=0.226 (i.e. in the depleted region of figure 1) as a function of the initial angle between the comet's and Jupiter's perihelion, in radians. Not only the values change dramatically, but no smooth curve can be fitted to the points of this plot: the chaotic behaviour can appear and disappear with very small changes in the initial conditions.

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

But even the chaotic behaviour is not a stable prediction: because in between a chaotic region there are islands of ordered behaviour, where resonant asteroidal orbits can survive the perturbations by the planets; this can be seen by computing the divergence ratio for many nearby orbits (Figure 2) or by plotting the orbital elements and looking for resonant behaviour (Milani and Nobili, 1984a).

We are not stating that the orbits of comets are impossible to compute. For a fixed span of time, by complying with a list of cautions that are suggested by the discussion presented in this paper, it is possible to compute the orbit of a comet to a reasonable accuracy, unless very close approaches to a planet do occur. However the very long-term evolution of the single cometary orbits is not accessible to our computations, and this state of affairs being due to the very nature of the problem it is not likely to change soon; the qualitative behaviour of cometary orbits is on the contrary amenable to study, in a statistical sense, and numerical integrations are an essential tool for this purpose.

REFERENCES

Arnold, V.I. (1963) Usp. Mat. Nauk. 18, 91.

Benettin, G., Galgani, L., Giorgilli, A. and Strelcyn, J.M. (1980) Meccanica, March 1980, 9.

Brouwer, D. (1937) Astr. J. 46, 149.

Fabri, E. and Penco, U. (1984) "Propagation of the Round-off Errors in Numerical Integrations", in preparation.

Heinrici, P. (1962) Discrete Variable Methods in Ordinary Differential Equations, John Wily & Sons, New York-London.

Henon, M. and Helies, C. (1964) Astron. J. 69, 73.

Kinoshita, H. (1968) Pubbl. Astr. Soc. Japan 20, 1.

Kovalevsky,J. (1963) Introduction a la mecanique celeste, Armand Colin, Paris.

Milani, A. and Nobili, A.M. (1984a) Celestial Mechanics, in press.

Milani, A. and Nobili, A.M. (1984b) Astron. Astrophys., in press.

Smale, S. (1967) Bull. A.M.S. 73, 747.

Wintner, A. (1941) The Analytical Foundations of Celestial Mechanics, Princeton Univ. Press.

Cohen, C.J., Hubbard, E.C. and Oesterwinter, C. (1973) Astr. Pap. Am. Ephem. 22, pt.1.

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press

Discussion

Marsden : The inadequacy of the Brouwer error-accumulation model was not a problem with automatic computers until after 1960. One could always program the early computers to act in the same way as mechanical desk calculators with regard to rounding. The problem arose with the introduction of purely binary computers and of high-level computer languages. I recall that around 1961 an assistant at Yale, using one of the new computers, found a rather dramatic decrease in the semimajor axis of an orbit. This was on a friday afternoon. Brouwer then spent the whole weekend integrating the two-body problem over several revolutions, using a desk calculator,but truncating rather than rounding. By monday morning he was convinced that his 1937 paper did not apply in the case of truncation.

Milani : However this was not published.

Valsecchi : What model did you use for your integration of asteroid orbits? What is your expectation about the nature and number of protective mechanisms using more complex models?

Milani : The elliptic restricted planar 3-body model. When the third dimension is taken into account, the protection mechanism based on the inclination can play a role; this has been shown by Froeschlé and Scholl (Astron. Astrophys. 1979). There is at least one asteroid protected in 3 different ways, one based on the inclination: it is 721 Tabora.

Zadumansky : For testing the accuracy of numerical experiments your methods are good. However Lyapounov's theory of stability may give valuable qualitative indications.

Milani : That is true as a matter of principle. However if you happen to find an unpredicted resonance, the theoretical predictions have to be changed.

https://doi.org/10.1017/S0252921100083937 Published online by Cambridge University Press