1937AJ.....46..149B

OCT 7 19

THE

ASTRONOMICAL JOURNAL

FOUNDED BY B. A. GOULD

No. 1072

VOL. XLVI ALBANY, N. Y., 1937, OCTOBER 1 No. 16

ON THE ACCUMULATION OF ERRORS IN NUMERICAL INTEGRATION.

By DIRK BROUWER

1. The problem of the accumulation of errors in the sum of a number of tabular quantities has been investigated by F. Schlesinger*. If tabular values have been computed to an accuracy beyond the last place given, the maximum error of each is ± 0.5 units of the last place, and we may assume that any error between these limits is as likely to occur as any other. The square of the mean error is 1/12 in units of the last tabular place; the mean error of the sum of n tabular quantities is $(n/12)^{\frac{1}{2}} = .289\ n^{\frac{1}{2}}$. Schlesinger proved that for large values of n the frequency-function of the errors approaches the Gaussian law. Therefore, for large values of n the probable errors of the sum equals .6745 times the mean error $= .195\ n^{\frac{1}{2}}$.

2. These results apply immediately to the numerical integration of a definite integral. After n steps of the integration the accumulated error is essentially that of the sum of n tabular quantities. In the case of a double integration a double summation is necessary. Let ϵ_0, ϵ_1, ... ϵ_n be the rounding-off errors in the tabulated quantities; the error in the nth item of the double summation is then

$$n\,\epsilon_0+(n-1)\,\epsilon_1+(n-2)\,\epsilon_2+\ldots\ldots+2\,\epsilon_{n-2}+\epsilon_{n-1}.$$

If, as may be assumed, ϵ_0, ϵ_1, ... are independent, and if the square of the mean error of each is 1/12, the square of the mean error of the double integral after n steps will be, in units of the last tabluar place,

$$\frac{1}{12}\sum_1^n \nu^2 = \frac{n^3}{36} + \frac{n^2}{24} + \frac{n}{72}.$$

For large values of n the mean is therefore $1/6\ n^{\frac{3}{2}}$, and the probable error $.1124\ n^{\frac{3}{2}}$.

Newcomb† considered the limitation due to the accumulation of rounding-off errors in the method of special perturbations in celestial mechanics. He dealt particularly with the application of numerical integration to the method of variation of arbitrary constants. The perturbations are obtained as the result of single integrations for all the elements except the mean longitude. The latter requires a double integration, and it is in this element only that serious accumulation of errors, proportional to $n^{\frac{3}{2}}$, occurs.

If the equations in rectangular coordinates are integrated numerically, either by Encke's or Cowell's method, one has a differential equation of the second order in each coordinate. The accumulation of errors to be expected is of the order of $n^{\frac{3}{2}}$ in each coordinate after n steps of the integration, as stated by Newcomb. This consideration may lead to a preference for obtaining the perturbations in the elements rather than in the coordinates, since in the former the serious accumulation occurs in the mean longitude only. It will be shown in the present paper that his prejudice is not justified, and that the accumulation of errors in the numerical integration by Encke's or Cowell's method is equivalent to an error in the mean longitude proportional to $n^{\frac{3}{2}}$, and to errors in the other elements proportional to $n^{\frac{3}{2}}$.

3. There is an important difference between the numerical evaluation of a definite integral and the numerical integration of a differential equation. In the former case the values of the integrand are given functions of the independent variable that can be computed in advance before the summation is started. In the latter case the integration proceeds step by step, the result of previous summations being used in computing each tabular value. The tabular errors are, therefore, not independent of the accumulation of errors incurred in earlier steps of the integreation. This feature changes the problem completely.

4. In the case of the numerical integration of the equations of planetary motions by Cowell's method, the equations are

* *Astronomical Journal*, **30**, 183, 1917.
† *Astronomische Nachrichten* **148**, 321, 1899.

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1937AJ.....46..149B

$$\frac{d^2x}{dt^2} = f_x, \qquad \frac{d^2y}{dt^2} = f_y, \qquad \frac{d^2z}{dt^2} = f_z, \tag{1}$$

The accelerations f_x, f_y, f_z are, in the case of elliptic motion,

$$f_x = -Kxr^{-3}, \qquad f_y = -Kyr^{-3}, \qquad f_z = -Kzr^{-3},$$

with $r^2 = x^2+y^2+z^2$, K being a positive constant. If the body concerned is a minor planet or a comet, $K = w^2k^2$, w being the interval of integration in days and k the Gaussian gravitational constant.

I take the integration-interval w as unit of time. Then, if μ is the mean motion in radians in this unit of time, and a the semi-major axis in astronomical units, Kepler's third law gives

$$a^3\,\mu^2 = k^2w^2 = K \tag{2}$$

In disturbed motion the attractions due to the disturbing planets are added to f_x, f_y, f_z, but this complication is of no consequence in the following.

Using the conventional notation of $'f$, $''f$ for items in the first and second summation columns and f', f''... for first and second differences of f, and putting

$$'f(t) = \tfrac{1}{2}\{'f(t-\tfrac{1}{2}) + 'f(t+\tfrac{1}{2})\}, \text{ etc.},$$

the coordinates and velocity components of any date t are obtained by the formulae:

$$x = ''f_x(t) + \tfrac{1}{12}\,f_x(t) - \tfrac{1}{240}\,f_x''(t) + \ldots.$$
$$x' = 'f_x(t) - \tfrac{1}{12}\,f_x'(t) + \tfrac{11}{720}\,f_x'''(t) - \ldots.$$

and similar expressions for y, z, y', z'.

In considering the errors I shall omit the corrective terms, and put the errors in x, x', etc., equal to those in $''f_x$, $'f_x$ etc., respectively. A rigorous analysis has been carried sufficiently far to show that the results of the first approximation treated here are not materially affected by this omission.

In computing f_x, f_y, f_z for $t=0$ from x, y, z for that date, the rounding-off errors introduced are

$$\epsilon_{x,0}, \qquad \epsilon_{y,0}, \qquad \epsilon_{z,0}.$$

For $t=1$, the errors are

$$\epsilon_{x,1}, \qquad \epsilon_{y,1}, \qquad \epsilon_{z,1}.$$

We may consider these as independent pure rounding-off errors. They introduce the following errors in the coordinates and velocity components for $t=1$.

$$\begin{aligned} \delta x_1 &= \tfrac{1}{2}\,\epsilon_{x,0}, & \delta x_1' &= \tfrac{1}{2}(\epsilon_{x,0}+\epsilon_{x,1}), \\ \delta y_1 &= \tfrac{1}{2}\,\epsilon_{y,0}, & \delta y_1' &= \tfrac{1}{2}(\epsilon_{y,0}+\epsilon_{y,1}), \\ \delta z_1 &= \tfrac{1}{2}\,\epsilon_{z,0}, & \delta z_1' &= \tfrac{1}{2}(\epsilon_{z,0}+\epsilon_{z,1}). \end{aligned} \tag{3}$$

These are the errors in the values derived by numerical integration for $t=1$ as compared with the exact values corresponding to the constants of integration defined by the coordinates and velocity components for $t=0$. Let these constants be

$$c_{i,0}, \qquad i = 1.....6.$$

The errors in the constants for $t=1$ are

$$c_{i,1} - c_{i,0} = \delta c_{i,1} = \left(\frac{\partial c_i}{\partial x}\right)_1 \delta x_1 + \left(\frac{\partial c_i}{\partial y}\right)_1 \delta y_1 + \left(\frac{\partial c_i}{\partial z}\right)_1 \delta z_1 + \left(\frac{\partial c_i}{\partial x'}\right)_1 \partial x_1' + \left(\frac{\partial c_i}{\partial y'}\right)_1 \delta y_1' + \left(\frac{\partial c_i}{\partial z'}\right)_1 \delta z_1'. \tag{4}$$

In dealing with the step from $t=1$ to $t=2$, I consider the errors referred to the orbit defined by the coordinates and velocity components obtained by the integration for $t=1$, i.e. referred to the orbit with constants of integration $c_{i,1}$. In the values of f_x, f_y, f_z for $t=1$ we may still consider the errors

$$\epsilon_{x,1}, \qquad \epsilon_{y,1}, \qquad \epsilon_{z,1};$$

those for $t=2$ are

$$\epsilon_{x,2}, \qquad \epsilon_{y,2}, \qquad \epsilon_{z,2}.$$

Consequently,

$$\begin{aligned} \delta x_2 &= \tfrac{1}{2}\epsilon_{x,1}, & \delta x_2' &= \tfrac{1}{2}(\epsilon_{x,1}+\epsilon_{x,2}), \\ \delta y_2 &= \tfrac{1}{2}\epsilon_{y,1}, & \delta y_2' &= \tfrac{1}{2}(\epsilon_{y,1}+\epsilon_{y,2}), \\ \delta z_2 &= \tfrac{1}{2}\epsilon_{z,1}, & \delta z_2' &= \tfrac{1}{2}(\epsilon_{z,1}+\epsilon_{z,2}). \end{aligned} \tag{5}$$

From (3), (5) the relations can be written down immediately for any value of t, and the general formula corresponding to (4) is:

$$c_{i,t} - c_{i,t-1} = \delta c_{i,t} = \left(\frac{\partial c_i}{\partial x}\right)_t \delta x_t + \left(\frac{\partial c_i}{\partial y}\right)_t \delta y_t + \left(\frac{\partial c_i}{\partial z}\right)_t \delta z_t + \left(\frac{\partial c_i}{\partial x'}\right)_t \delta x_t' + \left(\frac{\partial c_i}{\partial y'}\right)_t \delta y_t' + \left(\frac{\partial c_i}{\partial z'}\right)_t \delta z_t'. \tag{6}$$

5. I consider in particular the semi-major axis, for which the energy integral

$$\frac{1}{a} = \frac{2}{r} - \frac{x'^2+y'^2+z'^2}{K}$$

gives immediately

$$K\frac{\delta a_1}{2a^2} = \frac{K}{r^3}(x_1\delta x_1 + y_1\delta y_1 + z_1\delta z_1) + (x_1'\delta x_1' + y_1'\delta y_1' + z_1'\delta z_1'), = -(x_1''\delta x_1 + y_1''\,\delta y_1 + z_1''\delta z_1) + (x_1'\delta x_1' + y_1'\delta y_1' + z_1'\delta z_1').$$

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1937AJ.....46..149B

Substituting (3) I obtain:

$$K\frac{\delta a_1}{a^2}=(x'-x'')_1\,\epsilon_{x,0}+(y'-y'')_1\,\epsilon_{y,0}+(z'-z'')_1\,\epsilon_{z,0}+x_1'\epsilon_{x,1}+y_1'\epsilon_{y,1}+z_1'\epsilon_{z,1}\doteq(x'\epsilon_x+y'\epsilon_y+z'\epsilon_z)_0+(x'\epsilon_x+y'\epsilon_y+z'\epsilon_z)_1.$$

Introducing for brevity

$$E_t=(x'\epsilon_x+y'\epsilon_y+z'\epsilon_z)_t,$$

the general formula becomes

$$\delta a_t=\frac{a^2}{K}(E_{t-1}+E_t). \tag{7}$$

The error in a after n steps of the integration is

$$\sum_1^n \delta a_t=\frac{2a^2}{K}\left\{\sum_1^n E_t+\tfrac{1}{2}E_0-\tfrac{1}{2}E_n\right\} \tag{8}$$

Since all the errors ϵ are independent, and have a mean value $\sqrt{1/12}$ in units of the last place, the square of the mean error in a is, ignoring the last two terms of (8),

$$\frac{4a^4}{K^2}\cdot\frac{1}{12}\cdot\sum_1^n V_t^2, \tag{9}$$

where $V_t^2=(x'^2+y'^2+z'^2)_t$, the mean value of which is K/a. Consequently, the mean error in a after n steps of the integration is

$$\frac{1}{\sqrt{3}}a^{\frac{3}{2}}K^{-\frac{1}{2}}n^{\frac{1}{2}}=\frac{1}{\sqrt{3}}\mu^{-1}n^{\frac{1}{2}} \tag{10}$$

It is easily seen that the mean errors in all the elements except the mean longitude are proportional to the square root of the number of steps.

6. Let for the mean longitude the general formula (6) for time t give $(\delta\lambda_t)$. This is, according to the definitions, the excess of the mean longitude corresponding to the coordinates and velocities for time t over the mean longitude for that time computed from the coordinates and velocity components for time $t-1$. Since in the latter orbit the mean motion is μ_{t-1}, the full error incurred in the step from $t-1$ to t is

$$\delta\lambda_t=(\delta\lambda_t)+\mu_{t-1}-\mu_0,$$
$$=(\delta\lambda_t)+\delta\mu_1+\delta\mu_2+\ldots\ldots+\delta\mu_{t-1}.$$

The total error in the mean longitude after n steps of the integration is

$$\sum_1^n\delta\lambda_t=\sum_1^n(\delta\lambda_t)+(n-1)\delta\mu_1+(n-2)\delta\mu_2+\ldots\ldots+\delta\mu_{n-1}. \tag{11}$$

Since, according to (7) and (2),

$$\delta\mu_t=-\frac{3a\mu}{2K}(E_{t-1}+E_t),$$

the right-hand member of (11) becomes

$$\sum_1^n(\delta\lambda_t)-\frac{3a\mu}{2K}[(n-1)E_0+(2n-3)E_1+(2n-5)E_2+\ldots\ldots 3E_{n-2}+E_{n-1}].$$

The first part of this expression has a mean value proportional to $n^{\frac{1}{2}}$ and will be neglected. Evaluation of the second part similar to that of the right hand member of (8) gives for the square of the mean error in the mean longitude

$$\frac{9a^2\mu^2}{4K^2}\cdot\frac{1}{12}\cdot[(n-1)^2V_0^2+(2n-3)^2V_1^2+(2n-5)^2V_2^2+\ldots+3^2V_{n-2}^2+V_{n-1}^2]\ldots \tag{12}$$

Using again K/a for the mean value of V_t^2, and replacing K by $a^3\mu^2$, the result becomes

$$\frac{3}{16a^2}\left[\frac{4}{3}n^3-3n^2+\frac{5}{3}n\right]$$

For large values of n the last two terms become small compared with the first term, and the mean error in the mean longitude after n steps, in units of the last place, is found to be

$$\frac{1}{2a}n^{\frac{3}{2}}. \tag{13}$$

7. The use of the mean value K/a for V_t^2 in evaluating the sums (9), (12) needs additional justification. This is correct in (9) if the summation extends over a whole period of revolution, and in (12) if the summation extends over many periods of revolution. The complete expression for V_t^2 is

$$V_t^2=\frac{K}{a}\left\{1+\Sigma\,C_j\cos j(M_0+\mu t),\right\},$$

Cj being power series in the eccentricity e, and M_0 the mean anomaly for $t=0$. If this is used in (12), the sum will contain terms factored by e, e^2,..... that depend on M_0, even if this sum extends over an integral number of periods. These additional terms become small compared with the contribution due to the mean value K/a if the integration covers a large number of periods.

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1937AJ.....46..149B

In evaluating the sum (12) for a single period it is found that the accumulation of errors is maximum if the integration is started preceding perihelion and minimum if it is started preceding aphelion by an amount that depends on the eccentricity. If the integration extends over a large number of periods it is immaterial whether the integration is started near perihelion or aphelion.

8. It is of some interest to compare the accumulation of errors in the numerical integration of the equations of elliptic motion with that of the equation of harmonic motion

$$\frac{d^2x}{dt^2} + \kappa^2 x = 0. \tag{14}$$

In the latter case the frequency κ is a numerical constant, independent of the amplitude. Consequently, the only cause for accumulation of errors proportional to $n^{\frac{3}{2}}$ is absent. The accumulation of errors is proportional to $n^{\frac{1}{2}}$, although the integration involves a double summation. This result can also be derived by considering directly the effects of the rounding-off errors on the integrated values for x, rather than on the constants of integration.

9. Since the two methods, ENCKE's and COWELL's, are analytically identical, the results obtained apply equally to both methods. They also apply to NUMEROV's Extrapolation method which may be considered to be a modification of COWELL's method.

The main result (13) that the mean error of the mean longitude is proportional to $n^{\frac{3}{2}}$ suggests that the longer the interval used, the higher the accuracy attained. There is, however, an important compensating factor in favor of a shorter interval, namely, that with a shorter interval a higher number of decimal places may be used. COMRIE* suggests as a safe rule that, for purposes of checking and convenient step-by-step extropolation, the sixth difference of the accelerations should not exceed two figures. Since halving the interval reduces the sixth differences by a factor $2^{-8} = 1/256$, one can use two or three more decimal places with the shorter interval. The accumulation of errors in the two cases is therefore in ratio $100 \times 2^{-\frac{3}{2}} = 35$, or $1000 \times 2^{-\frac{3}{2}} = 354$ in favor of the shorter interval with two or three extra decimal places.

The relative advantages of the existing methods depend on so many different factors as to make almost every particular problem a case for special consideration. In some problems the interval that can be used effectively with ENCKE's method is twice that with COWELL's method. In such cases ENCKE's method has the advantage of slower accumulation of errors in one-half the number of steps. NUMEROV has given special attention to using the greatest possible interval in each case.

* Planetary Coordinates for the years 1800–1940, p. xviii.

10. An interesting numerical test-case for the theorem developed in this paper is provided by the orbit of comet Comas Sola (1926f), treated by VINTER HANSEN*.

Starting with the same osculating elements for 1926, November 30.0, obtained by VINTER HANSEN, the two methods of numerical integration were used, MISS VINTER HANSEN using COWELL's method and MR. D. H. SADLER using ENCKE's method. MISS VINTER HANSEN carried the computations to the seventh place "most of the time," MR. SADLER to the eighth place to the middle of 1934, and from then on to the seventh place. For about seven years the ENCKE-integration was performed with twice the interval used in the COWELL-integration. For 1935, August 26.0, after 135 steps with COWELL's method and 82 steps with ENCKE's method the values for x, y, z, x', y', z', are given by VINTER HANSEN. Now, since the only serious difference to be expected is in the mean longitude, we expect the differences COWELL minus ENCKE in the coordinates, Δx, Δy, Δz, to be proportional to the velocity components and those in the velocity components $\Delta x'$, $\Delta y'$, $\Delta z'$, to be proportional to the components of the acceleration. Introducing an unknown u, the following relations will be satisfied:

$$x'u = \Delta x \qquad x''u = \Delta x'$$
$$y'u = \Delta y \qquad y''u = \Delta y'$$
$$z'u = \Delta z \qquad z''u = \Delta z'$$

Solving u, I find in units of the sixth place, $u = -207$, corresponding to a difference in mean longitude $\Delta\lambda = -242 \times 10^{-7}$ radians. The numerical values are as follows:

	Comp. ($u = -207$)	Diff.
$-.903u = \Delta x = +187$	+187	0
$-.093u = \Delta y = +19$	+19	0
$+.177u = \Delta z = -36$	−37	+1
$-.043u = \Delta x' = +8$	+9	−1
$-.263u = \Delta y' = +53$	+54	−1
$-.136u = \Delta z' = +29$	+28	+1

The remaining differences are as small as expected: they should be of the order of C/n times the original differences, C being independent of n.

I also computed the differences (C − E) in the coordinates for seven dates, each separated by 16 steps in the COWELL-integration, and made a solution for $\Delta\lambda$ for each date. The results are given in the table below:

The last column of the table gives the ratio between

* *Publ. og mindre Medd. Kob. Obs.* Nr. 85, 1933, The same orbit is used as an illustration in COMRIE's Planetary coordinates.

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1937AJ.....46..149B

TABLE

J. D.	n	$\Delta\lambda . 10^7$	$n^{\frac{3}{2}}/2a$	$\lvert\Delta\lambda\rvert . 10^7 \div n^{\frac{3}{2}}/2a$
242 5000.5	16	− 5.6	7.7	.73
5160.5	32	− 16.1	21.7	.74
5360.5	48	− 34.1	40.0	.85
5680.5	64	− 58	61	.95
6160.5	80	− 76	86	.88
6800.5	96	−136	113	1.20
7440.5	112	−221	130	1.70

the value for $\Delta\lambda$ actually found, and the mean value predicted by (13), assuming that the integration by COWELL'S method was performed to seven decimal places. The errors in the ENCKE-integration are small compared with those in the COWELL-integration owing to the use in the former of an additional decimal place, and a greater interval for almost the entire period. The difference $\Delta\lambda$ is, therefore, almost entirely due to the accumulation of errors in the COWELL-integration.

On account of the short period covered by the integrations, and the limitations of formula (13) dealt with in previous sections, the ratios obtained serve to show only that the order of magnitude predicted by the formula is correct. For a complete numerical test integrations covering a much longer period would be necessary.

In order to obtain satisfactory solutions for the seven intermediate dates it was necessary to compute the coordinates by ENCKE'S method with a set of elements that differs slightly from that used by VINTER HANSEN for starting the COWELL-integration. Owing to the limited number of significant figures, especially in the velocity components, the starting values correspond to a set of elements that is sufficiently different from the original set to cause periodic differences in the coordinates up to about 6 units in the sixth decimal place. By introducing corrections to the elements, the residuals obtained by eliminating the parts due to $\Delta\lambda$ were reduced from a maximum of 12 to one of 5 units in the sixth place; they could probably be reduced still more by a second approximation. This difficulty with the starting values for the velocity components has a very small effect on the final date, which is within 80 days a whole period of revolution after the starting date.

11. It has been shown in the present paper that the accumulation of rounding-off errors in the numerical integration of the equations of planetary motion in rectangular coordinates, after n steps of the integration, is equivalent to a mean error proportional to $n^{\frac{3}{2}}$ in the mean longitude, and to $n^{\frac{1}{2}}$ in the other orbital elements.

A new problem is now suggested: to develop a method of integration in which use is made of this theorem. One may proceed with the integration in rectangular coordinates to an accuracy sufficient as far as the accumulation of errors proportional to $n^{\frac{1}{2}}$ is concerned, and make a supplementary integration for the longitude to a higher accuracy. The results of the latter are finally used to reduce the effects of the accumulation of errors in the integration in rectangular coordinates.

One of several possibilities is to choose the equations of the variation of arbitrary constants for the supplementary integration. The theoretical possibility of such a procedure is evident, but further investigation is necessary to decide whether an efficient method can be developed.

Yale University Observatory,
1937, August 14.

ORBITAL MOTION OF ADS 9744 (ι *Serpentis*)

BY RAYMOND H. WILSON, JR.

Since its discovery by HUSSEY in 1902 two occultations of this binary have been definitely observed by VAN BIESBROECK and AITKEN. The motion is mainly in distance, which never exceeds 0".25, showing the eccentricity or inclination, or both, of the orbit to be high. Since the components appear identical, it is impossible to distinguish between single and double reversal of quadrant at an occultation; hence the two sets of elements and residuals—one for each assumption—are found to be almost equally satisfactory.

The conflicting Greenwich observations during the occultation of 1908-10 were thought to be safely disregarded because in every case the observer had noted "poor conditions" or "doubtful."

All the elements, except the inclination, were estimated directly from the interpolation curves. In deriving the inclination it was most convenient to use the following relation, which I have not found in the literature:

$$\rho^2 \frac{d\theta}{dt} = 2na^2 \cos i \cos \varphi \qquad (e = \sin \varphi)$$

The Orbit I hypothesis (high eccentricity) is favored

© American Astronomical Society • Provided by the NASA Astrophysics Data System