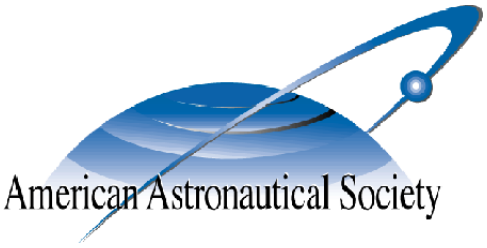
Paper AAS 03-171



COMPARISON OF ACCURACY ASSESSMENT TECHNIQUES FOR NUMERICAL INTEGRATION

Matt Berry
Virginia Tech and Naval Research Laboratory
and
Liam Healy
Naval Research Laboratory

13th AAS/AIAA Space Flight Mechanics Meeting

Ponce, Puerto Rico **9-13 February 2003**

AAS Publications Office, P.O. Box 28130, San Diego, CA 92198

AAS 03-171

COMPARISON OF ACCURACY ASSESSMENT TECHNIQUES FOR NUMERICAL INTEGRATION

Matthew M. Berry* **Liam M. Healy†**

Abstract

Knowledge of accuracy of numerical integration is important for composing an overall numerical error budget; in orbit determination and propagation for space surveillance, there is frequently a computation time-accuracy tradeoff that must be balanced. There are several techniques to assess the accuracy of a numerical integrator. In this paper we compare some of those techniques: comparison with two-body results, with step-size halving, with a higher-order integrator, using a reverse test, and with a nearby exactly integrable solution (Zadunaisky's technique). Selection of different kinds of orbits for testing is important, and an RMS error ratio may be constructed to condense results into a compact form. Our results show that step-size halving and higher-order testing give consistent results, that the reverse test does not, and that Zadunaisky's technique performs well with a single-step integrator, but that more work is needed to implement it with a multi-step integrator.

INTRODUCTION

Any orbit determination or propagation computation is subject to error from a number of sources: observational error, modeling error, numerical integration error, etc. In building an error budget, it is useful to have a numerical estimate of each. For space surveillance, there is frequently a computation time-accuracy tradeoff that must be balanced. Instead of improving accuracy indefinitely, it may be desirable to accept a degraded accuracy to a certain level in order to improve the computation time or to simplify the integrator in an embedded computer. In this paper, we survey the common techniques of assessing integration error and compare their results.

Although assessment of integration accuracy is a long-standing problem, much of the focus of the available literature is on the three-body problem in astronomy, particularly over long periods of time. Because of the chaotic nature of such systems, accuracy is desired to assess characteristics of chaotic regions in time and space. In astrodynamics, we are more concerned with the problem of orbiting a planet with geopotential, atmospheric drag, solar radiation and other perturbations, and often for a relatively modest number of orbits. Accuracy is desired here for precise knowledge of spacecraft position and orbit over a short period of time. Although the problems are similar, the needs are sufficiently different that the integrator, including order and step-size, may be chosen differently. Nevertheless, as astronomers have developed techniques to assess integrators, we may make use of this work to develop a means of assessing integrators for astrodynamics. Furthermore,

*Graduate Assistant, Department of Aerospace and Ocean Engineering, Virginia Tech, Blacksburg, Virginia 24061, and Naval Research Laboratory, Code 8233, Washington, DC 20375-5355, E-mail: maberry2@vt.edu.

†Research Physicist, Naval Research Laboratory, Code 8233, Washington, DC 20375-5355, and Lecturer, Department of Aerospace Engineering, University of Maryland, College Park, MD 20742. E-mail: Liam.Healy@nrl.navy.mil.

we would like a quantitative estimate of error which will assist in choosing a less but still sufficiently accurate method to gain computation speed.

Attempts at characterization of integration error frequently stop with two-body integration because of the ability to determine absolute accuracy. Perturbations have a significant effect on integration accuracy that is not apparent from a two-body study, as is shown below.

An N^{th} order system of ordinary differential equations, with initial conditions given at $t = t_0$, can be written in the general form (Ref. 1)

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(t, \boldsymbol{x}), \qquad \boldsymbol{x}(t_0) = \boldsymbol{x_0}, \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{f}$ are vectors of N functions and $\boldsymbol{x_0}$ is a vector of N constants. In our case $\boldsymbol{x}$ consists of three position functions and three velocity functions. If some numerical integration algorithm is used to solve (1), the algorithm generates an approximate solution $\tilde{\boldsymbol{x}}$. After n steps the accumulated error is

$$\boldsymbol{\xi_n} = \boldsymbol{x}(t_n) - \tilde{\boldsymbol{x}}_{\boldsymbol{n}}, \tag{2}$$

where $\boldsymbol{x}(t)$ is the exact solution, $t_n = t_0 + nh$, and h is the step-size. The error can be written in the form (Ref. 1)

$$\boldsymbol{\xi}_n = [\boldsymbol{x}(t_n) - \boldsymbol{x_n}] + [\boldsymbol{x_n} - \tilde{\boldsymbol{x}}_{\boldsymbol{n}}], \tag{3}$$

where the first difference is the *truncation error*, and the second difference is the *round-off error* (Ref. 2). Truncation error exists because the numerical integration algorithm has been truncated at some (locally correct) order, p. Truncation error is dependent on the step-size h and the order p, and decreases as the step-size is decreased. Round-off error exists because computers only keep track of numbers to a finite number of significant digits, and some error is introduced during every calculation. Round-off error increases as the step-size is decreased, because more computations are performed.

Various techniques exist to estimate the error, $\boldsymbol{\xi}$, of a numerical integrator. Each technique has strengths and weaknesses; our purpose in this paper is to describe what these are and to identify a practical procedure testing integrators. These techniques have common features, so at the outset, we consider the choice of a set of test orbits that provide a variety of realistic conditions and thus adequately balance the integrators. Moreover, since all techniques involve the numerical comparison of two prospective orbits, we describe the computation of the error ratio, which provides a figure of relative merit between two integrated orbits. One of the orbits is produced by the integrator being tested; the other, the reference orbit, covers the same time period and initial conditions and may be produced by either integration or analytic computation.

TEST ORBITS AND INTEGRATORS

If the integrator is expected to perform satisfactorily over a wide variety of orbits, some representative sample of these orbits is needed as a test set of initial conditions. In particular, forces that stress an integrator, such as atmospheric re-entry, should be included to give a sense of the worst case. For the case of space surveillance catalog maintenance, circular orbits near the earth and at geosynchronous altitude represent the bulk of the catalog; the addition of a high-eccentricity elliptical orbit with perigee dipping well into the atmosphere stresses the integrator.

Three test cases are considered, a low earth orbit, an elliptical orbit, and a geosynchronous orbit. The initial conditions of these test cases are shown in Table 1. The velocities have been listed to ten significant figures so that the orbital elements have the desired value. The test cases all have an initial epoch of 2001-10-01 00:00:00 UT, and a ballistic coefficient of $0.01\,{}^{\text{m}^2}/_{\text{kg}}$.

Since our purpose here is to compare accuracy assessment techniques, we need not only test orbits, but test integrators on which to try the various techniques. We have chosen two integrators commonly used in astrodynamics. The first is a fourth-order Runge-Kutta, as described in Ref. 2 (pp 320-321). The Runge-Kutta is a single-step, single-integration integrator. The step-sizes used for the Runge-Kutta integrator are 5 seconds for test cases 1 and 2, and 1 minute for test case 3.

Table 1: Test Case Initial Conditions

test #	$\boldsymbol{r}$ (km)	$\dot{\boldsymbol{r}}$ (km/sec)	Perigee Height (km)	Ecc.	Inc. (°)
1	$6678.137\hat{\boldsymbol{I}}$	$5.918276127\hat{\boldsymbol{J}} + 4.966023315\hat{\boldsymbol{K}}$	300	0.0	40
2	$6578.137\hat{\boldsymbol{I}}$	$7.888427772\hat{\boldsymbol{J}} + 6.619176834\hat{\boldsymbol{K}}$	200	0.75	40
3	$42164.172\hat{\boldsymbol{I}}$	$3.074660237\hat{\boldsymbol{J}}$	35786	0.0	0.0

The second integrator is an eighth-order combined Gauss-Jackson and summed Adams, as described in Ref. 3. The Gauss-Jackson is an multi-step, predictor-corrector, double-integration integrator which computes position directly from the accelerations. To get velocity information, the Gauss-Jackson is combined with an eighth-order summed Adams integrator, which is a single-integration, multi-step, predictor-corrector integrator. The corrector is applied only once at each integration step, giving a predict, evaluate, correct cycle. For the Gauss-Jackson integrator, the step-sizes are 30 seconds for test cases 1 and 2, and 20 minutes for test case 3.

ERROR RATIO

In the tests, a metric for integration accuracy is found by defining an error ratio in terms of the RMS error of the integration (Ref. 4). First define position errors as

$$\Delta r = |r_{\text{computed}} - r_{\text{reference}}|. \tag{4}$$

The RMS position error can be calculated,

$$\Delta r_{\text{RMS}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta r_i)^2}. \tag{5}$$

The RMS position error is normalized by the apogee distance and the number of orbits to find the position error ratio,

$$\rho_r = \frac{\Delta r_{\text{RMS}}}{r_A N_{\text{orbits}}}. \tag{6}$$

In Ref. 4, Merson uses a position error ratio to test integrators. A velocity ratio test is added here because velocity and position are often integrated with different numerical integrators, and it may be useful to estimate the error in orbital elements, which depend directly on the velocity. The velocity error ratio is the RMS velocity error normalized by the number of orbits and the perigee speed.

$$\rho_v = \frac{\Delta v_{\text{RMS}}}{v_P N_{\text{orbits}}}. \tag{7}$$

ACCURACY ASSESSMENT TECHNIQUES

Two-body test

If the force is a simple two-body (Kepler) force, an exact solution is available for comparison. The advantage of this technique is that the error is then known exactly, but the disadvantage is that the force may not be realistic. This test does not necessarily indicate how well the integrator handles perturbations. Orbits typically integrated for space surveillance may cover a time period of several days or more. During that time, the integrated effects of the perturbations cause a substantial deviation from the two-body solution. If an integrator handles the two-body force well, but not one or more of the perturbations, the integration error has a great affect on the computed orbit. Typically, drag and solar radiation pressure are the forces that may cause trouble with integrators,

drag because of its variability and dependence on velocity, and solar radiation pressure because it is a step function in time, which multi-step integrators have difficulty with because of the back-points.

Fox (Ref. 5) performed extensive tests on integrators using the two-body test. Part of the purpose of his study was to assess accuracy in light of the execution time, so he put "dead weight" into the force evaluation — computations that do nothing but soak up processor time — in an attempt to simulate the evaluation-dominated execution time of integrations with realistic force models. Although we have not addressed the issue of execution time here, the use of a full perturbation model for most tests precludes the need to artificially slow the force evaluation, should that be of interest. Montenbruck (Ref. 6) also used the two body force to assess integrators, even for earth-orbiting artificial satellites.

Table 2 shows position and velocity error ratios with the two-body test for both Runge-Kutta and Gauss-Jackson integrators. The position and velocity error ratios are found using (6) and (7), respectively. Ephemeris generated by the numerical integrators over a three day time span is used for r_{computed} and v_{computed}, and $r_{\text{reference}}$ and $v_{\text{reference}}$ are the values given by the exact analytic solution. Though the cases use different step-sizes, the ephemeris is generated at one minute intervals, so that N in (5) is 4321. For case 3 with the Gauss-Jackson integrator, where the step-size is 20 minutes, the intermediate points are found using a 5^{th} order interpolator. Table 3 gives the maximum position error over three days for each test case. Comparing Table 2 to Table 3 demonstrates how an error ratio corresponds to the maximum position error, which may be of interest.

Table 2: Two-Body Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	2.05×10^{-10}	2.05×10^{-10}	7.96×10^{-14}	7.98×10^{-14}
2	2.49×10^{-10}	5.15×10^{-10}	1.03×10^{-11}	2.26×10^{-11}
3	3.27×10^{-11}	3.25×10^{-11}	8.95×10^{-12}	8.57×10^{-11}

Table 3: Two-Body Position Error (mm)

test #	Runge-Kutta	Gauss-Jackson
1	133	0.0494
2	286	14.9
3	7.21	2.60

From Table 2 we see that the Gauss-Jackson integrator is more accurate in every case, except in the velocity for case 3. Gauss-Jackson is roughly three orders of magnitude more accurate than Runge-Kutta for the low earth orbit, roughly one order of magnitude more accurate for the eccentric orbit, and of comparable accuracy for the geosynchronous orbit. Both integrators show the least accuracy in position with the eccentric orbit, though the Gauss-Jackson has the least accuracy in velocity with the geosynchronous orbit.

The two-body test directly applied is a necessary but not sufficient test for assessing integrators, because of the significance to integration of perturbations, but two-body integration used with the other techniques can provide an indicator of uncaptured error in those techniques. The techniques described in the subsequent sections are each tested two ways. In the first test only the two-body force is considered, so that the error measured by the technique can be compared to the known error. In the second test a perturbation model is used. The perturbations forces are 36×36 WGS-84 geopotential, the Jacchia 70 drag model (Ref. 7), and lunar and solar forces. Note that all of these perturbations are continuous forces. Solar radiation pressure, a perturbation that has discontinuities at eclipse boundaries, can cause numerical integration errors, especially when using multi-step integrators, because the discontinuity violates the assumption made in the formulation

of the integrators that the forces are smooth and continuous (Ref. 8). To simplify our study this perturbation is not considered.

Step-size halving

For the step-size halving test, the reference integration is produced with the same integrator but with the step-size cut in half. Because the truncation error is related to the step-size, this technique can give a good estimate of the former, and even an estimate of the order of the integration method, provided the step-size is large enough that truncation error dominates the total error. If the step-size is too small, round-off error, which will increase as the step-size is decreased, will appear very different than truncation error. However, in the step-size region where truncation error gives way to round-off error, there can be confusion as to what is being measured and to quantify that measurement, so that a further decrease in step-size is necessary to confirm the onset of round-off error. As long as the step-sizes used are much larger than this mixed regime, an error ratio may be computed.

Tables 4 and 5 show error ratios using the step-size halving test, for the two-body force and for the full force model, respectively. The error ratios with the two-body force are the same order of magnitude as the true error ratios in Table 2, and even match to one significant digit, except for case 1 with Gauss-Jackson. This shows that step-size halving gives a reasonable measure of integration error.

Table 4: Two-Body Step-Size Halving Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	1.96×10^{-10}	1.96×10^{-10}	2.22×10^{-14}	2.22×10^{-14}
2	2.34×10^{-10}	4.85×10^{-10}	1.03×10^{-11}	2.56×10^{-11}
3	3.07×10^{-11}	3.05×10^{-11}	8.94×10^{-12}	8.62×10^{-11}

Table 5: Perturbed Step-Size Halving Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	1.19×10^{-9}	1.19×10^{-9}	4.63×10^{-9}	4.64×10^{-9}
2	1.16×10^{-9}	2.50×10^{-9}	9.93×10^{-9}	2.11×10^{-8}
3	3.07×10^{-11}	3.05×10^{-11}	8.95×10^{-12}	8.62×10^{-11}

Comparing Table 4 to Table 5, we see that the numerical integrators perform worse in the presence of perturbations for the low-earth and eccentric cases, but nearly the same for the geosynchronous case. This may be because the geosynchronous orbit is not subject to drag, and the presence of drag causes an increase in integration error. This shows that the two-body test described above does not capture all of the error of an integrator.

A more formal error analysis is also possible with step-size halving. In theory the global error is on the order of the step size to the power of the order of the numerical integrator (Ref. 9),

$$x - \tilde{x} = \xi \approx Ch^p, \tag{8}$$

where x is the actual solution, $\tilde{x}$ is the numerical solution, and C is some constant that depends on the numerical integrator. An estimate of ξ can be found by comparing results with half the step-size. Considering the numerical solution, $\tilde{x}$, to be a function of the step-size h, we form the equation

$$[x - \tilde{x}(h/2)] \approx Ch^p/2^p \approx [x - \tilde{x}(h)]/2^p. \tag{9}$$

This can be solved for the error,

$$\begin{aligned} 2^p[x - \tilde{x}(h/2)] &\approx [x - \tilde{x}(h)] \\ (2^p - 1)[x - \tilde{x}(h/2)] &\approx \tilde{x}(h/2) - \tilde{x}(h) \\ \xi(h/2) &\approx \frac{\tilde{x}(h/2) - \tilde{x}(h)}{2^p - 1} \end{aligned} \tag{10}$$

In order for (10) to be valid, (8) must hold true. This condition can be checked by forming the quotient

$$\frac{[x - \tilde{x}(h/2)] - [x - \tilde{x}(h/4)]}{[x - \tilde{x}(h)] - [x - \tilde{x}(h/2)]} = \frac{\tilde{x}(h/4) - \tilde{x}(h/2)}{\tilde{x}(h/2) - \tilde{x}(h)} \approx \frac{Ch^p/2^p - Ch^p/4^p}{Ch^p - Ch^p/2^p} = \frac{1}{2^p}. \tag{11}$$

The quotients should approach 2^{-p} as the step size decreases. But when round-off becomes a factor, the quotients drift away from the theoretical value. As long as the quotients indicate that (8) is valid, round-off error is not a major concern and (10) can be used. Equations (10) and (11) have been written in scalar form; in practice they can be applied to each component of the vector state $\boldsymbol{x}$.

It is possible to improve this process by generalizing the scaling applied to the step size. By using an arbitrary scale factor instead of only one half, more consistent results are possible and it is possible to locate the onset of roundoff error more precisely. Then, the order of the integration method may be computed numerically when one has an exact solution. In fact, one can numerically compute the order of any series expansion this way; the order computation is a special case where the function $g(h)$ is a dependent variable error (like position) after integrating over a fixed time period from fixed initial conditions; it is a function of step-size h only.

If g is an analytic function, it has a Taylor expansion around y_0,

$$g(y) = c_0 + c_1(y - y_0) + \frac{1}{2}c_2(y - y_0)^2 + \dots \tag{12}$$

(In the integration case, $y_0 = 0$, but we will keep it for completeness.) Presuming that the integrator is correct to some order, we may assume the first few coefficients are zero. As above, p is the order (first exponent dropped in the global Taylor expansion), or *error exponent* we wish to solve. As the first nonzero coefficient $c_j = 0$ for $j < p$, and $c_p \neq 0$ it gives the order of the method. The Taylor series may be written starting at order p,

$$|g(y)| = \left| \frac{c_p}{p!}(y - y_0)^p + \dots \right|. \tag{13}$$

By evaluating this function at two different points $y = y_0 + s_1$ and $y = y_0 + s_2$ and taking the ratio, we may determine p,

$$\frac{|g(y_0 + s_1)|}{|g(y_0 + s_2)|} = \left| \frac{\frac{c_p}{p!}s_1^p + \frac{c_{p+1}}{(p+1)!}s_1^{p+1} + \dots}{\frac{c_p}{p!}s_2^p + \frac{c_{p+1}}{(p+1)!}s_2^{p+1} + \dots} \right| \approx \left| \frac{s_1}{s_2} \right|^p, \tag{14}$$

where the approximation presumes that successive terms in both series are insignificant.

The choice of scaling factors s_1 and s_2 are important. In contrast to the "step-size halving" case where the goal is to find the accuracy of the integrator, these will likely not be in the ratio $2:1$. In order to make the results of the determination of e more robust, we can evaluate the ratio (14) at a series of different scaling factors $s_1, s_2, \dots$ This series can be constructed from a *geometric step evaluation* series, σb, $\sigma^2 b$, $\sigma^3 b$, $\dots$, for some scaling $\sigma > 0$. Say $s_1 = \sigma^n b$ and $s_2 = \sigma^{n-1} b$ so that the ratio (14) becomes

$$\frac{|g(y_0 + \sigma^n b)|}{|g(y_0 + \sigma^{n-1} b)|} \approx \sigma^p. \tag{15}$$

Now, to find p, we take the logarithm

$$p \approx \log \frac{|g(y_0 + \sigma^n b)|}{|g(y_0 + \sigma^{n-1} b)|} \Big/ \log \sigma. \tag{16}$$

The accurate determination of p is hampered by the approximation we have made in (15) and the necessity of dealing with a finite-sized floating point number on a computer. The terms after the first ones in (14) become relatively insignificant if s_1 and s_2 are small, i.e., if b is small. But choosing b too small can cause significant digits to be lost due to finite word size.

Traditionally when "eyeballing" a scaling result, one chooses $\sigma = 10$ (or $\sigma = 0.1$) because it is easy for a human to compute p. For a computer-determined value, $\sigma = 2$ might make more sense (and thus we would live up to the section title), however, we have found that better results are obtained with a σ near 1. Specifically, when $\sigma \approx 1.02 - 1.05$, the error exponent p is often very clearly near an integer. The multiple successive terms in the series, i.e., computing (16) for a series of n, serve to confirm the computed value of p and show at what value of s round-off error sets in; in the transition from truncation to round-off error, a noticeable erratic deviation in the computed truncation error exponent p will be observed. This will help identify the lower bound on step size for the integrator accuracy check, below which round off error is significant.

Comparison with high-order integrator

The comparison integration may be a higher-order - high-accuracy integrator. The advantage of this technique is that perturbations can be tested. However, this technique relies on the assumption that the higher-order integrator is correct, or more correct, than the integrator being tested. This is not necessarily true. Tables 6 and 7 show error ratios comparing the two test integrators to a 14^{th}-order Gauss-Jackson, with the two-body force and full perturbation forces, respectively. The step-size used in the 14^{th}-order Gauss-Jackson is 15 seconds for cases 1 and 2, and 1 minute for case 3.

Table 6: Two-Body High-Order Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	2.05×10^{-10}	2.05×10^{-10}	5.34×10^{-14}	5.34×10^{-14}
2	2.49×10^{-10}	5.16×10^{-10}	1.04×10^{-11}	2.30×10^{-11}
3	3.28×10^{-11}	3.25×10^{-11}	9.02×10^{-12}	8.58×10^{-11}

Table 7: Perturbed High-Order Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	4.59×10^{-9}	4.61×10^{-9}	4.62×10^{-9}	4.64×10^{-9}
2	7.19×10^{-9}	1.55×10^{-8}	9.94×10^{-9}	2.11×10^{-8}
3	3.27×10^{-11}	3.25×10^{-11}	9.07×10^{-12}	8.58×10^{-11}

Comparing Table 6 to Table 2, we see that the high-order test results are the same order of magnitude as the true results in every case. For the Runge-Kutta integrator, the test matches the true results to at least two significant figures in each case. For the Gauss-Jackson integrator, the test matches the true results to two significant figures for the eccentric case, and one significant figure for the geosynchronous case. Note that the higher the actual error ratio, the closer the high-order results match the true results. This is because the high-order test assumes that the reference integrator is much better than the integrator being tested.

Comparing Table 7 to Table 5, we see that the step-sizing halving results and the high-order results with perturbations are the same order of magnitude in all cases except the velocity of case 2 with Runge-Kutta. The Gauss-Jackson results match to at least one significant figure in each case, while the only case where the Runge-Kutta results match to one significant figure is case 3. The Gauss-Jackson results match closely between the step-size halving test and the high-order test because in both tests, the reference integrator is a Gauss-Jackson integrator with a low step-size.

Reverse test

In this technique the test and reference orbits come from the same integrator. The test orbit is produced by integrating forward from the initial conditions, and the reference orbit is produced by integrating backward to the original starting time using the final state of the first integration as initial conditions. These two integrations should be identical, and any difference between them is due to integration error. Using this technique to measure integration error is advantageous because it is a relatively simple procedure to perform. A disadvantage with this technique is that it does not measure any reversible integration error, that is, any error that is an odd function of the step-size is canceled off when the sign of the step changes on the reverse integration. It has been used extensively for integration accuracy checks, including recently in the N-body problem (Ref. 10).

The two tables 8 and 9 show the reverse test with a two-body force and full perturbation forces, respectively. Again the tests are over a three day interval with ephemeris generated at one minute increments. Comparing Table 8 to Table 2 shows that the reverse test fails to capture a significant portion of the error; in the case of orbit number 3 for the Runge-Kutta integrator, only about a tenth of the error is captured. This may be an example of the weakness of the reverse test pointed out by Zadunaisky (Ref. 11), who demonstrated for the three-body problem that the reverse test will certify that the a second-order Adams-Moulton multi-step method as perfectly accurate because of the symmetry of the equations used under time reversal. Thus two-body integration has proved its value: it can show when another test is deficient. Comparing Table 9 to Tables 5 and 7, we see that the reverse test is also inconsistent with the other techniques in the presence of perturbations.

Table 8: Two-Body Reverse Test Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	2.27×10^{-10}	2.27×10^{-10}	4.55×10^{-15}	4.54×10^{-15}
2	5.13×10^{-11}	1.08×10^{-10}	2.21×10^{-11}	4.67×10^{-11}
3	3.53×10^{-12}	3.53×10^{-12}	2.11×10^{-11}	2.12×10^{-11}

Table 9: Perturbed Reverse Test Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	2.28×10^{-10}	2.29×10^{-10}	7.79×10^{-10}	7.81×10^{-10}
2	5.18×10^{-11}	1.09×10^{-10}	2.46×10^{-11}	1.11×10^{-10}
3	3.52×10^{-12}	3.52×10^{-12}	1.97×10^{-11}	1.98×10^{-11}

Integral invariants

Another frequently-used technique for integration accuracy assessment is to check the invariance of integral invariants such as energy, which is easy to perform. This technique has been used recently for integrators applied in the N-body problem (Ref. 10). The main drawback is that it does not

capture all errors. For example, energy invariance is blind to in-track errors, because an orbit shifted in time has the same energy. In general, the more forces present that break a particular symmetry, the fewer conserved quantities and thus the fewer quantities that can be checked; the presence of drag means that energy is no longer conserved at all. Huang and Innanen (Ref. 12) show that this accuracy check is not exact and reliable and suggest a revised technique. We have not attempted this technique in the present study, because of its limitations.

Zadunaisky's test

Zadunaisky, in Ref. 13, Ref. 1, Ref. 14, and Ref. 11, suggests a technique for measuring numerical integration error. This technique is based on the construction of an analytical function near the solution to the actual problem, then constructing differential equations for which this function is an exact solution. It has two desirable properties: first, we have an exact solution, because we constructed the problem to match the solution we had, and second, this solution is realistic, because it is close in some sense to the solution of the real problem. It thus is like the two-body test in providing an absolute reference for error computation, but has a behavior that mimics real forces.

First a set of polynomials $\boldsymbol{P}(t)$ are determined which represent the components of $\tilde{\boldsymbol{x}}$, the numerical solution of (1). A *pseudo-system* of equations may be constructed from these polynomials,

$$\dot{\boldsymbol{z}} = \boldsymbol{f}(t, \boldsymbol{z}) + \boldsymbol{D}(t), \tag{17}$$

where

$$\boldsymbol{D}(t) = \dot{\boldsymbol{P}}(t) - \boldsymbol{f}(t, \boldsymbol{P}(t)), \tag{18}$$

and where (17) has the same initial conditions as (1),

$$\boldsymbol{z}(t_0) = \boldsymbol{x_0}. \tag{19}$$

The exact solution of the pseudo-system is known, and is $\boldsymbol{P}(t)$. When the pseudo-system is integrated with the same numerical integrator used to create $\tilde{\boldsymbol{x}}$, the errors of the pseudo-system,

$$\boldsymbol{\xi} = \tilde{\boldsymbol{z}} - \boldsymbol{P}(t), \tag{20}$$

may be a good approximation of the error in the original problem.

In Ref. 13, Zadunaisky gives conditions for $\boldsymbol{P}(t)$ under which this technique gives a good approximation of the original error,

$$\left.\begin{array}{r} ||\tilde{\boldsymbol{x}} - \boldsymbol{P}(t)|| \\ ||\tilde{\boldsymbol{x}}^{p+1} - \boldsymbol{P}^{p+1}(t)|| \end{array}\right\} \leq \delta = O(h^2), \tag{21}$$

where p is the order of the numerical integrator, and $\tilde{\boldsymbol{x}}^{p+1}$ are the numerical approximations for the $(p+1)^{\text{th}}$ derivatives of $\boldsymbol{x}$. These conditions are based on error propagation theory, and are meant to ensure that the asymptotic behavior of the errors accumulated in the numerical integration of both the original system and the pseudo-system are the same. In Ref. 11, Zadunaisky gives a different condition, which is that $\boldsymbol{D}(t)$ must not be larger than either the local truncation error or the local round-off error.

In Ref. 14 and Ref. 11, Zadunaisky finds the polynomial, $\boldsymbol{P}(t)$, needed to form the pseudo-system, using Newton's interpolation formula with backward divided differences. Because an $\tilde{N}^{\text{th}}$ degree polynomial is needed to interpolate $\tilde{N}+1$ points, the original interval is broken into subintervals to avoid the problems involved with interpolating data to a high degree polynomial. After integrating the original problem for $\tilde{N}$ steps, where $\tilde{N} \leq 10$, he applies Newton's formula to obtain polynomials $\boldsymbol{P}(t)$ of $\tilde{N}^{\text{th}}$ degree. Each polynomial $P_i(t)$ interpolates one of the components of position or velocity through the $\tilde{N}+1$ points spanned by the $\tilde{N}$ steps. These polynomials are used to construct the pseudo-system, (17), and the error estimate is obtained. This process is then repeated using the last point of the previous set as the first point on the next set, and using $\tilde{\boldsymbol{x}}_{\tilde{N}+1}$ and $\tilde{\boldsymbol{z}}_{\tilde{N}+1}$ as initial

conditions in the original system and pseudo-system, respectively. Using this method, the solution to the pseudo-system over the entire interval, $\boldsymbol{z}(t)$, is a function of successive $\tilde{N}^{\text{th}}$ degree polynomials that match up the subintervals of $\tilde{N}+1$ points. Therefore $\boldsymbol{z}(t)$ is continuous over the entire interval, but its derivative is discontinuous at the last point of each subinterval. Zadunaisky claims that these discontinuities are irrelevant to the validity of the technique.

To implement Zadunaisky's technique for a given test case, we first generate ephemeris $\tilde{\boldsymbol{x}}(t)$, by numerically integrating the test case. The ephemeris should be generated at time increments equal to the step-size h of the numerical integrator. This ephemeris is then used to find coefficients of the polynomials $\boldsymbol{P}(t)$ at each subinterval. The polynomials are of the form

$$P_i(\tilde{t}) = a_0 + a_1\tilde{t} + a_2\tilde{t}^2 + ... + a_{\tilde{N}}\tilde{t}^{\tilde{N}}, \tag{22}$$

where $\tilde{t}$ is the time since the beginning of the subinterval, and the subscript i refers to the component of the state $\tilde{\boldsymbol{x}} = [r_x r_y r_z v_x v_y v_z]^T$ that the polynomial fits. To make the polynomial exactly match the ephemeris at each of the $\tilde{N}+1$ points on the subinterval, the coefficients $a_0 ... a_{\tilde{N}}$ are found by solving the system

$$\begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 1 & \tilde{t}_1 & \tilde{t}_1^2 & ... & \tilde{t}_1^{\tilde{N}} \\ 1 & \tilde{t}_2 & \tilde{t}_2^2 & ... & \tilde{t}_2^{\tilde{N}} \\ . & & & & \\ . & & & & \\ 1 & \tilde{t}_{\tilde{N}} & \tilde{t}_{\tilde{N}}^2 & ... & \tilde{t}_{\tilde{N}}^{\tilde{N}} \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ . \\ . \\ a_{\tilde{N}} \end{Bmatrix} = \begin{Bmatrix} \tilde{x}_i(t_0) \\ \tilde{x}_i(t_1) \\ \tilde{x}_i(t_2) \\ . \\ . \\ \tilde{x}_i(t_{\tilde{N}}) \end{Bmatrix}, \tag{23}$$

where the subscripts on $\tilde{t}$ and t refer to the points on the subinterval, and $\tilde{t}_0 = 0$. This system can be solved for the coefficients by inverting the first matrix and multiplying it by the vector on the right-hand side. Though this procedure must be repeated for each of the six components and for each subinterval, the matrix inversion only needs to be performed once, because the matrix only depends on the order $\tilde{N}$, and the time step h.

After the coefficients are found, the same numerical integrator is used to generate another set of ephemeris, but with the force model modified so that the pseudo-system (17) is being integrated. Normally, the force model returns an acceleration based on position, velocity, and time, $\ddot{\boldsymbol{r}} = \boldsymbol{f}(t, \boldsymbol{r}, \dot{\boldsymbol{r}})$. Instead, the coefficients for the appropriate subinterval are used to find the values of $\boldsymbol{P}$ and $\dot{\boldsymbol{P}}$, and the force model returns

$$\ddot{\boldsymbol{r}} = \boldsymbol{f}(t, \boldsymbol{r}, \dot{\boldsymbol{r}}) + \dot{\boldsymbol{P}}(t) - \boldsymbol{f}(t, \boldsymbol{P}(t)). \tag{24}$$

Note that $\boldsymbol{P}$ is a six component vector consisting of both position and velocity, though $\ddot{\boldsymbol{r}}$ is only a three component vector. Therefore, the vector $\dot{\boldsymbol{P}}$ used in (24) is a three component vector consisting of the first derivative of the velocity polynomials. This makes (24) have a different form than (17), but this change makes the technique easier to implement.

There are two practical considerations to make when generating and using the coefficients in the modified force model. First, the ephemeris from which the coefficients are generated must be in the same coordinate system in which the integration is performed. Second, conversions may be necessary if the units of the ephemeris, and of the time $\tilde{t}$ used in the polynomial equations, are different from the units used by the integrator.

The ephemeris generated in the second integration, $\tilde{\boldsymbol{z}}$, is compared to the original ephemeris, $\tilde{\boldsymbol{x}}$, to determine an error ratio. For test case 3, with the Runge-Kutta integrator and using only the two-body force, over three days we get a position error ratio of 2.65×10^{-7} with a 9^{th} order polynomial, $\tilde{N} = 9$. From Table 2 we know that the actual error ratio is 3.27×10^{-11}, so the technique is four orders of magnitude too high. If we look at only the first 90 minutes, the technique gives an error ratio that is two orders of magnitude too high. When $\tilde{N} = 6$, the error ratio given by the technique is 1.80×10^{-9}, two orders of magnitude too high. However, the error ratio over 90 minutes is the correct order of magnitude. With $\tilde{N} = 5$, the error ratio is 5.89×10^{-10}, one order of magnitude too

high. Over the 90 minute time span, the technique gives an error ratio that is an order of magnitude too low. These results highlight the difficulty in choosing an appropriate degree polynomial.

Because the reliability of Zadunaisky's technique depends on how well the polynomial fits the ephemeris, we suggest another method of determining the polynomials to improve the results. In addition to matching the values of the ephemeris, the derivatives are also matched at the end-points of each subinterval. So for a subinterval consisting of $\tilde{N}-1$ points, an $\tilde{N}^{\text{th}}$ degree polynomial is found by modifying (23),

$$\begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 1 & \tilde{t}_1 & \tilde{t}_1^2 & ... & \tilde{t}_1^{\tilde{N}} \\ . & & & & \\ . & & & & \\ 1 & \tilde{t}_{\tilde{N}-2} & \tilde{t}_{\tilde{N}-2}^2 & ... & \tilde{t}_{\tilde{N}-2}^{\tilde{N}} \\ 0 & 1 & 0 & ... & 0 \\ 0 & 1 & 2\tilde{t}_{\tilde{N}-2} & ... & \tilde{N}\tilde{t}_{\tilde{N}-2}^{\tilde{N}-1} \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ . \\ . \\ a_{\tilde{N}-2} \\ a_{\tilde{N}-1} \\ a_{\tilde{N}} \end{Bmatrix} = \begin{Bmatrix} \tilde{x}_i(t_0) \\ \tilde{x}_i(t_1) \\ \\ . \\ . \\ \tilde{x}_i(t_{\tilde{N}-2}) \\ \dot{\tilde{x}}_i(t_0) \\ \dot{\tilde{x}}_i(t_{\tilde{N}-2}) \end{Bmatrix}, \tag{25}$$

where $\dot{\tilde{x}}_i(t)$ is the velocity given in the ephemeris when i is a position component, and the acceleration given by the same force model used to generate the ephemeris when i is a velocity component.

Tables 10 and 11 show error ratios given by Zadunaisky's technique with the two-body force and the full force model, respectively. The polynomials used in the method match each point in the subinterval, and the derivatives of the polynomial match the force model at the end points of each subinterval, as described above. For the Runge-Kutta integrator a 5^{th} order polynomial is used, $\tilde{N}=5$, and for the Gauss-Jackson a 3^{rd} order polynomial is used, $\tilde{N}=3$. We find these polynomials give the best results. With higher order polynomials, the technique gives error ratios that are too high, compared to the two-body test, and with lower order polynomials the error ratios are too low. In Ref. 14, Zadunaisky uses $\tilde{N} \geq p$, which we have followed for the Runge-Kutta but not for Gauss-Jackson.

Table 10: Two-Body Zadunaisky Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	3.08×10^{-10}	3.08×10^{-10}	3.33×10^{-14}	3.33×10^{-14}
2	3.39×10^{-9}	7.30×10^{-9}	6.83×10^{-14}	1.50×10^{-13}
3	3.87×10^{-11}	3.87×10^{-11}	1.86×10^{-14}	1.85×10^{-14}

Table 11: Perturbed Zadunaisky Results

test #	Runge-Kutta		Gauss-Jackson	
	pos	vel	pos	vel
1	1.81×10^{-9}	1.82×10^{-9}	8.06×10^{-8}	8.08×10^{-8}
2	2.11×10^{-9}	4.40×10^{-9}	6.55×10^{-8}	1.44×10^{-7}
3	3.82×10^{-11}	3.82×10^{-11}	1.01×10^{-12}	9.79×10^{-13}

Comparing Table 10 to Table 2, we see that the technique matches the true error ratios in order of magnitude for cases 1 and 3 for Runge-Kutta, and case 1 for Gauss-Jackson. For the eccentric orbit with Runge-Kutta the technique gives an error ratio that is an order of magnitude too high. In Ref. 11 Zadunaisky suggests a variant of his technique that gives improved results for eccentric orbits, but we have not implemented it here. For Gauss-Jackson the technique gives an error ratio that is three orders of magnitude too low for cases 2 and 3. Comparing the perturbed results in

Table 11 to the results from the step-size halving and high-order test in Tables 5 and 7, we see that the Runge-Kutta results with Zadunaisky's technique match the results from the other techniques, at least in order of magnitude. However the Gauss-Jackson results are an order of magnitude higher for Zadunaisky's technique in cases 1 and 2, and an order of magnitude lower in case 3.

Note that for the method of choosing polynomials we have described above, the minimum order polynomial is $\tilde{N} = 3$, because that involves a subinterval of two points. The fact that we found the best results with Gauss-Jackson with this lowest order polynomial, and that this order violates Zadunaisky's criteria of $\tilde{N} \geq p$, indicates that this method of determining polynomials may not be appropriate for Gauss-Jackson, and may explain why the error ratios do not match the error ratios from the other tests.

CONCLUSION

In this paper we demonstrate five techniques for assessing the accuracy of numerical integrators. We test these techniques on a Runge-Kutta integrator and a Gauss-Jackson integrator, with three test orbits. The test orbits are chosen to give a representation of most earth orbits, as well as to stress the integrator with a high drag case.

The two-body test gives an exact measure of error when perturbations are not considered. This exact measure of error is useful to evaluate the other techniques, by testing them without perturbations. However, when these other techniques are used with perturbations, they show a significantly larger error than the two-body test when drag is a factor. The step-size halving test, the higher-order test, and Zadunaisky's test with Runge-Kutta give results that are consistent with one another, and match the two-body error well. The reverse test gives results that are inconsistent, and previous authors have shown it to be unreliable. Zadunaisky's technique with Gauss-Jackson also gives inconsistent results, though this may be due to the method we have chosen to determine the polynomial $\boldsymbol{P}(t)$.

ACKNOWLEDGEMENTS

We thank Prof. Fred Lutze and Prof. Lee Johnson of Virginia Tech for help in understanding numerical interpolation, and its use in implementing Zadunaisky's technique. We also thank Prof. Johnson for notes on step-size halving. We thank Dr. Paul Schumacher of Naval Network and Space Operations Command for introducing us to the literature.

REFERENCES

[1] Zadunaisky, P. E., "On the Accuracy in the Numerical Computation of Orbits," In Giacaglia, G. E. O., editor, *Periodic Orbits, Stability and Resonances*, pp. 216–227, Dordrecht, Holland, 1970. D. Reidel Publishing Company.

[2] Burden, R. L. and Faires, J. D., *Numerical Analysis*, Brooks/Cole Publishing Company, New York, 1997.

[3] Berry, M. and Healy, L., "Implementation of Gauss-Jackson Integration for Orbit Propagation," In *Advances in Astronautics*, San Diego, CA, August 2001. American Astronautical Society AAS 01–426; to appear.

[4] Merson, R. H., *Numerical Integration of the Differential Equations of Celestial Mechanics*, Technical Report TR 74184, Royal Aircraft Establishment, Farnborough, Hants, UK, January 1975. Defense Technical Information Center number AD B004645.

[5] Fox, K., "Numerical integration of the equations of motion of celestial mechanics," *Celestial Mechanics*, Vol. 33, No. 2, 1984, pp. 127–142, June 1984.

[6] Montenbruck, O., "Numerical Integration Methods for Orbital Motion," *Celestial Mechanics and Dynamical Astronomy*, Vol. 53, pp. 59–69, 1992.

[7] Jacchia, L. G., *New Static Models of the Thermosphere and Exosphere with Empirical Temperature Models*, Technical Report 313, Smithsonian Astrophysical Observatory, 1970.

[8] Woodburn, J., "Mitigation of the Effects of Eclipse Boundary Crossings on the Numerical Integration of Orbit Trajectories Using an Encke Type Correction Algorithm," In *AAS/AIAA Space Flight Mechanics Meeting, Santa Barbara, CA, 11–14 February 2001*, AAS Publications Office, P. O. Box 28130, San Diego, CA 92198, 2001. AAS/AIAA Paper AAS 01-223.

[9] Johnson, L. W. and Riess, R. D., *Numerical Analysis*, Addison-Wesley, Reading, Mass., 1982.

[10] Hadjifotinou, K. G. and Gousidou-Koutita, M., "Comparison of Numerical Methods for the Integration of Natural Satellite Systems," *Celestial Mechanics and Dynamical Astronomy*, Vol. 70, No. 2, 1998, pp. 99–113, 1998.

[11] Zadunaisky, P. E., "On the Accuracy in the Numerical Solution of the N-Body Problem," *Celestial Mechanics*, Vol. 20, pp. 209–230, 1979.

[12] Huang, T.-Y. and Innanen, K. A., "The accuracy check in numerical integration of dynamical systems," *Astonomical Journal*, Vol. 88, No. 6, 1983, pp. 870–876, June 1983.

[13] Zadunaisky, P. E., "A Method for the Estimation of Errors Propagated in the Numerical Solution of a System of Ordinary Differential Equations," In Contopoulos, G., editor, *The Theory of Orbits in the Solar System and in Stellar Systems*, pp. 281–287, New York, 1966. International Astronomical Union, Academic Press.

[14] Zadunaisky, P. E., "On the Estimation of Errors Propagated in the Numerical Integration of Ordinary Differential Equations," *Numerische Mathematik*, Vol. 27, No. 1, 1976, pp. 21–39, 1976.