Achieving Brouwer's law with implicit Runge–Kutta methods

E. Hairer*, R.I. McLachlan†, and A. Razakarivony*

July 3, 2007

Abstract

In high accuracy long-time integration of differential equations, round-off errors may dominate truncation errors. This article studies the influence of round-off on the conservation of first integrals such as the total energy in Hamiltonian systems. For implicit Runge–Kutta methods, a standard implementation shows an unexpected propagation. We propose a modification that reduces the effect of round-off and shows a qualitative and quantitative improvement for an accurate integration over long times.

Keywords: probabilistic error propagation, implicit Runge–Kutta methods, long-time integration, efficient implementation.

1 Introduction

The long-time integration of differential equations with high accuracy is common in astronomy (e.g., the numerical computation of the solar system and the long-term solution for the insolation quantities of the Earth [6]; it is expected to be used for age calibrations of paleo-climatic data over 40 to 50 Myr). As soon as the local truncation error is close to (or below) the round-off unit, the main contribution to the local error of the numerical solution is due to the finite precision arithmetic on the computer. We are interested in better understanding the influence of this source of error to a numerical integration over long times.

In the present article we are mainly concerned with Hamiltonian systems, although much of the discussion can be extended straight-forwardly to the conservation of first integrals in arbitrary differential equations or to the propagation of the global error in integrable systems. Let

$$\dot{p} = -\nabla_q H(p,q), \qquad \dot{q} = \nabla_p H(p,q), \tag{1}$$

*Section de mathématiques, University of Geneva, Switzerland.

†Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand.

where $H(p,q)$ is a smooth function, called the energy of the system. This energy is a first integral of (1), which means that $H\big(p(t),q(t)\big) = Const$ along solutions of the system. The numerical energy $H(p_n,q_n)$, where (p_n,q_n) approximates the solution at time $t_n = nh$, is not constant. With exact arithmetic, the error grows linearly with time in general, but remains bounded and small without any secular drift for symplectic integration methods; see [3, Chap. IX]. Assuming that the round-off error of one step is a random variable with mean zero and variance proportional to the square of the round-off unit *eps*, the error contribution due to round-off will grow like (Brownian motion) the square-root of time. This is often called Brouwer's law [1] in the literature [2]. This model of round-off error was exploited in a detailed study by Henrici [4, 5]. Much attention has been paid to the propagation of round-off with linear multistep methods [8, 2] and composition methods [7]; we know of no such studies for Runge–Kutta methods.

This article is organized as follows. Section 2 presents numerical experiments of standard implementations of various integration methods, where the step size is chosen small enough to guarantee that the truncation error is below round-off. The rather surprising observation is that for composition methods the round-off error grows as expected like square-root of time, but that for Runge–Kutta methods shows a linear error growth. The reasons for this phenomenon are discussed in Section 3. We also propose modifications of a standard implementation of Runge–Kutta methods that allows us to recover the optimal (square-root of time) growth of round-off errors. A probabilistic explanation of the growth of round-off errors for the different implementations is given in Section 4. Finally, in Section 5 we discuss the statistical behaviour of round-off errors when our new constant step size implementation of the Gauss–Runge–Kutta methods is applied to the Hénon–Heiles problem and to the outer solar system.

2 Observed propagation of round-off

An efficient computation of very accurate numerical approximations for ordinary differential equations requires the use of integrators of high order. One can use high order multistep, Runge–Kutta, or composition and splitting methods. We do not discuss multistep methods in the present article. Our limited experiments have shown that they have remarkably good round-off error propagation.

Composition based on Störmer–Verlet. For a basic numerical scheme $\Phi_h(y)$ (usually symmetric and of order two) the symmetric composition

$$\Psi_h = \Phi_{\gamma_s h} \circ \Phi_{\gamma_{s-1} h} \circ \ldots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h}, \tag{2}$$

where $\gamma_{s+1-i} = \gamma_i$ for all i, allows one to get high order for suitable choices of the parameters γ_i. For the numerical experiments of the present arti-

cle we have chosen the coefficients of Suzuki & Umeno (order 8 and $s = 15$) as given in [3, p. 157]. As basic numerical scheme we consider the Störmer–Verlet discretization which, for Hamiltonian systems with separable $H(p,q) = T(p) + U(q)$ reads as

$$\begin{aligned} q_{n+1/2} &= q_n + \frac{h}{2}\nabla_p T(p_n), \\ p_{n+1} &= p_n - h\nabla_q U(q_{n+1/2}), \\ q_{n+1} &= q_{n+1/2} + \frac{h}{2}\nabla_p T(p_{n+1}). \end{aligned} \tag{3}$$

This method is explicit and the implementation of the corresponding composition method is straight-forward.

Implicit Runge–Kutta methods. For general first order differential equations $\dot{y} = f(y)$, Runge–Kutta methods are defined by

$$Y_i = y_n + h\sum_{j=1}^{s} a_{ij} f(Y_j) \tag{4}$$

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i f(Y_i), \tag{5}$$

where the integer s and the coefficients a_{ij}, b_i determine the method. We exclusively consider Gauss methods, which have highest possible order $r = 2s$, and are symplectic and symmetric. Their implementation is not straightforward, because a nonlinear system has to be solved for the internal stages $Y_1, \ldots, Y_s$. If the problem is non-stiff, it is common to apply fixed-point iteration. In our naïve implementation we iterate until the increment of two successive approximations satisfies

$$\Delta^{(k)} := \max_{i=1,\ldots,s} \left\| Y_i^{(k)} - Y_i^{(k-1)} \right\|_\infty \le \delta \tag{6}$$

where $\delta \approx 2 \cdot 10^{-16}$ (problem dependent) is chosen as the smallest positive number such that this criterion is satisfied before the increments start to oscillate due to round-off. In the update formula (5) the vector field $f(y)$ is evaluated at the most recent approximation $Y_i^{(k)}$ to the internal stages.

Numerical experiment. We consider the Hénon–Heiles model which is Hamiltonian with

$$H(p,q) = \tfrac{1}{2}\left(p_1^2 + p_2^2\right) + \tfrac{1}{2}\left(q_1^2 + q_2^2\right) + q_1^2 q_2 - \tfrac{1}{3}\, q_2^3,$$

and we choose initial values $q_1(0) = 0$, $q_2(0) = 0.3$, $p_2(0) = 0.2$, and the positive value $p_1(0)$ such that the Hamiltonian takes the value $H_0 = 1/8$ (the solution is chaotic; see [3, Section I.3]). On an interval of length $2\pi \cdot 10^6$ we apply the integrators mentioned above; a composition method of order 8,

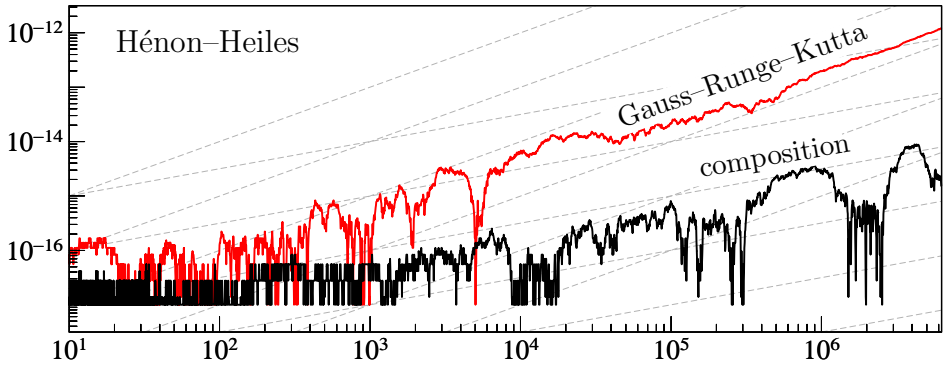

Figure 1: Propagation of round-off in the numerical Hamiltonian for the standard implementation of an implicit Runge–Kutta method of order 8 (step size $h = 2\pi/140$), and for a composition method of order 8 (basic step size $h = 2\pi/240$). The dotted grey lines have slopes 1 and 1/2, respectively.

which is explicit, and the Gauss–Runge–Kutta method of order 8, where the nonlinear system is solved by fixed-point iteration. For both integrators we apply compensated summation (see [3, Section VIII.5]). This can be seen as performing the addition in $y_{n+1} = y_n + h\beta_n$ in higher precision, so that the round-off error is reduced by a factor of h. We use step sizes such that in a computation with quadruple precision the maximal error in the Hamiltonian is approximately 10^{-18}, i.e., below the round-off unit. Notice that both integrators are symplectic so that there is no drift in the Hamiltonian due to the discretization error.

Figure 1 shows the absolute value of the error in the Hamiltonian as a function of time (in double logarithmic scale). Since the truncation error is very small, the curves represent the contribution of round-off. For the composition method it increases, as expected for a random walk, like the square root of time (this corresponds to lines with slope 1/2). More surprisingly, the round-off error of the implicit Runge–Kutta method is a superposition of a statistical error which grows like square root of time and is dominant until about $t = 10^4$, and of a deterministic error which grows linearly with time. This error is about 7.5×10^{-21} per step, or 3×10^{-4} *ulp* per step. Here 1 *ulp* (= units in the last place) is 2^{-55}, for machine $eps = 2^{-52}$ and $H_0 = 1/8$. So the linear drift is very small and not simply due to a naive accumulation of a few *ulp* per step, but rather due to a tiny non-zero bias in the pattern of positive and negative rounding errors.

3 Reducing the influence of round-off

The objective of this paper is to find the reasons of the linear growth of round-off errors in a standard implementation. We propose modifications that allow us to recover the expected square root of time behaviour.

Sources of the unexpected growth of round-off. After many numerical experiments with various methods and problems we came to the conclusion that there are essentially two sources of non-statistical errors that lead to the linear error growth of round-off.

- *Iterative solution of the nonlinear Runge–Kutta equations.* Fixed-point iteration acts like the power method and the error tends to the eigenvector of the dominating eigenvalue of the linearized equation.

- *Inexact Runge–Kutta coefficients.* In general, the coefficients a_{ij} and b_i are not machine numbers, and the computations are done with rounded coefficients $\widehat{a}_{ij}$ and $\widehat{b}_i$ which do not exactly satisfy the order conditions of the Runge–Kutta method. This is a systematic error, because the same (rounded) coefficients are used throughout the integration.[1]

Notice that for composition methods based on the Störmer–Verlet scheme none of these error sources is present. These methods are explicit and no iterative solution of nonlinear equations is involved. They are symplectic even for inexact coefficients γ_i. Thus the use of inexact coefficients does not contribute a term that grows linearly in time to the Hamiltonian, merely one that is bounded in time and of the order of roundoff. This explains the good long-time behaviour of the composition method in Fig. 1.

Remedies. To avoid these systematic errors in the implementation of implicit Runge–Kutta methods we have done many numerical computations over very long time intervals, and we came to the conclusion that the following modifications are the most efficient.

- *Iteration until convergence.* Instead of using the stopping criterion (6), we propose to iterate until either $\Delta^{(k)} = 0$ or $\Delta^{(k)} \geq \Delta^{(k-1)}$ which indicates that the increments of the iteration start to oscillate due to round-off. This stopping criterion has the advantage of not requiring a problem- and method-dependent δ. For the up-date formula (5) we use the values $f(Y_i^{(k-1)})$.

- *Simulating exact Runge–Kutta coefficients.* Our first idea was to use coefficients in quadruple precision. This can, however, be avoided by a trick inspired by compensated summation. We split the coefficients into

$$b_i = b_i^* + \widetilde{b}_i, \quad a_{ij} = a_{ij}^* + \widetilde{a}_{ij}, \tag{7}$$

[1] The use of inexact coefficients in Taylor series methods (multiplication by 1/3 instead of division by 3) leads to the same numerical phenomenon; c.f. the talk by Carlos Simó at the Castellón Conference on Geometric Integration.

where b_i^* and a_{ij}^* are exact machine numbers, e.g., rational approximations to b_i, a_{ij} with denominator 2^{10}, and we compute the internal stages as

$$Y_i = y_n + h\Big(\sum_{j=1}^{s} a_{ij}^* f(Y_j)\Big) + h\Big(\sum_{j=1}^{s} \widetilde{a}_{ij} f(Y_j)\Big),$$

and the up-date formula in a similar way. Since the coefficients $\widetilde{b}_i$ and $\widetilde{a}_{ij}$ are small, this procedure permits one to recover the missing few digits in the Runge–Kutta coefficients.

With regard to iteration until convergence, we note that in the experiments on Hénon–Heiles, a final value of $\Delta = 0$ was obtained in about 99.6% of time steps. In the other 0.4% of time steps, the mean of the final Δ values was 2 *ulp*, with a maximum of 4 *ulp* or 1.1×10^{-16}, the latter occurring only 55 times in 960 000 steps. The experiments confirm that it is not the size of the final errors that is significant, but the lack of systematic bias.

Numerical confirmation. We consider the Hénon–Heiles problem with the same data as in Section 2. Besides a quadruple precision implementation which shows the size of the truncation error, we consider the following implementations of implicit Runge–Kutta methods:

grk-co The stopping criterion is changed to "iteration until convergence" as described above.

grk-ex The Runge–Kutta coefficients are split according to (7). This produces results as if coefficients with higher precision were used.

grk-co-ex Both modifications are applied; the "iteration until convergence" procedure as well as the simulation of Runge–Kutta coefficients with higher precision.

In all implementations, compensated summation is employed to reduce the influence of round-off in the up-date formula.

From Fig. 2, where again the error in the Hamiltonian is drawn as a function of time, we can draw the following conclusions. The errors for the implementation "grk-co" are not much different from those for the standard implementation of Fig. 1. The idea of using (or simulating) Runge–Kutta coefficients with higher precision is much more important and improves considerably the propagation of round-off. However, on very long time intervals both implementations, "grk-co" and "grk-ex" show an undesired linear error growth of round-off in the Hamiltonian. Only the implementation "grk-co-ex" which combines both improvements, shows an optimal square root of time growth of the round-off error. It behaves very similar to the composition method in Fig. 1. With these modification of the implementation we could achieve Brouwer's law also for implicit Runge–Kutta methods.

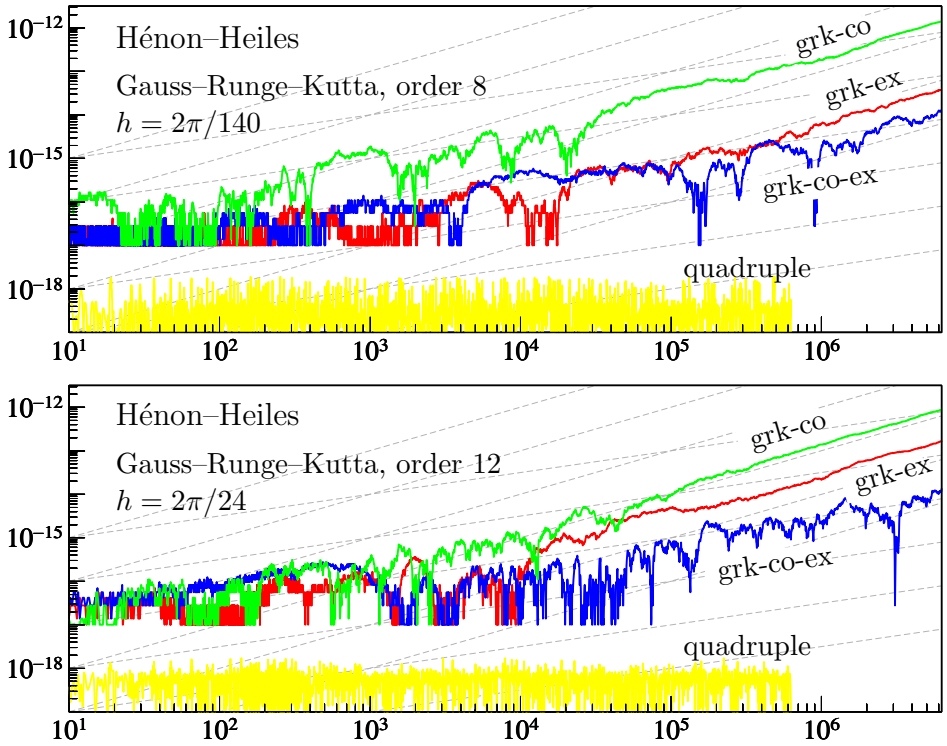

Figure 2: Error in the Hamiltonian of various implementations of implicit Runge–Kutta methods.

4 Probabilistic explanation of the error growth

To understand the long-time behaviour of round-off errors (experiments of Section 3) we make use of probability theory, an approach that has been developed in the classical book of Henrici [5].

Effect of rounded Runge–Kutta coefficients. In the equations (4)-(5), with b_i, a_{ij} replaced by their rounded machine numbers $\widehat{b}_i, \widehat{a}_{ij}$, we consider the internal stages and the numerical approximation at the grid points as random variables with expected values $\overline{Y}_i$ and $\overline{y}_n$, respectively. Assuming that the evaluation of the vector field is not biased and that our new stopping criterion does not give rise to systematic errors in the solution of the nonlinear system, the computed approximations satisfy the Runge–Kutta equations where a random vector is added, each component of which is independent with mean zero. The expected values of the internal stages and the numerical approximation then satisfy

$$\overline{Y}_i = \overline{y}_n + h \sum_{j=1}^{s} \widehat{a}_{ij} f(\overline{Y}_j), \qquad \overline{y}_{n+1} = \overline{y}_n + h \sum_{i=1}^{s} \widehat{b}_i f(\overline{Y}_i).$$

If we denote by $y_{n+1} = \Phi_h(y_n)$ the discrete flow of the Runge–Kutta method

(in exact arithmetic), the difference $\Phi_h(\overline{y}_n) - \overline{y}_{n+1}$ can be expanded into a Taylor series around $h = 0$ and yields the familiar formula

$$\begin{aligned} \Phi_h(\overline{y}_n) - \overline{y}_{n+1} &= h \Big(\sum_{i=1}^{s} b_i - \sum_{i=1}^{s} \widehat{b}_i \Big) f(\overline{y}_n) \\ &+ \frac{h^2}{2} \Big(\sum_{i,j=1}^{s} b_i\, a_{ij} - \sum_{i,j=1}^{s} \widehat{b}_i\, \widehat{a}_{ij} \Big) f'(\overline{y}_n) f(\overline{y}_n) + \dots \ . \end{aligned} \tag{8}$$

It precisely shows the systematic (local) error due to round-off in (implicit and explicit) Runge–Kutta methods. This systematic error is responsible for the linear growth of round-off errors as observed in Fig. 1. Depending on how well the rounded coefficients satisfy the order conditions, the error growth will be more or less pronounced.

Error growth of round-off in the energy. We consider sufficiently small step sizes so that the local truncation error is close to or below round-off. Considering a few terms of the modified Hamiltonian

$$\widetilde{H}(y) = H(y) + h^p H_{p+1}(y) + h^{p+1} H_{p+2}(y) + \dots \tag{9}$$

in the sense of backward error analysis [3, Section IX.3]), we can safely assume that $\widetilde{H}(\Phi_h(y)) = \widetilde{H}(y)$ for the numerical flow with exact Runge–Kutta coefficients. In this case the error contribution over one step in the modified Hamiltonian,

$$\widetilde{H}(\,\overline{y}_{n+1}) - \widetilde{H}(\,\overline{y}_n) = \varepsilon_n,$$

can be considered as a sequence of independent random variables. Their expected value is proportional to the expression in (8) and is negligible if the actually used Runge–Kutta coefficients $\widehat{b}_i$ and $\widehat{a}_{ij}$ are sufficiently close to b_i and a_{ij}. Their standard deviation is proportional to the round-off unit *eps*. The use of compensated summation now ensures that the expected absolute round-off error in $\widetilde{H}$ (or in y_n) per step is proportional to $eps\,h$. Brouwer's argument now gives $E[|\widetilde{H}(y_n) - \widetilde{H}(y_0)|] = C\,eps\,h^{1/2}\,t^{1/2}$ for $t = nh$ for some constant C. Since the perturbation in (9) is close to round-off and remains bounded, it does not affect the long-time behaviour of the error in the Hamiltonian. In contrast, if inexact Runge–Kutta coefficients are used that do not define a symplectic integrator, there will be no modified Hamiltonian and a linear growth of energy errors will result.

The same considerations apply to any first integral of any differential equation as long as there exists a modified first integral for the modified differential equation of the numerical integrator. This is the case for the angular momentum in N-body problems solved with a symplectic integrator, but it is not the case for the Runge–Lenz–Pauli vector in the Kepler problem. The error in this first integral increases linearly with time even with exact

Runge–Kutta coefficients and in exact arithmetics and we cannot hope for doing better with our implementation.

For a composition method we directly consider the modified Hamiltonian corresponding to the method with rounded coefficients (which is also symplectic). The same analysis then shows that round-off errors in the energy always verify Brouwer's law.

5 Statistical confirmation

A single experiment, as that of Fig. 2, could lead to wrong conclusions due to the statistical nature of round-off errors. We first consider the Hénon–Heiles equation, and we repeat the same calculation many times with randomly perturbed initial values all with the same initial value of the Hamiltonian.

Figure 3 illustrates the random walk nature of the energy error. The mean energy error is zero to within sampling error, and the standard deviation is proportional to $\sqrt{n}$. The standard deviation of the energy error is about $8 \times 10^{-18} h\, n^{1/2}$, or $0.3\, h\, n^{1/2}\, ulp$. This is consistent with the above model of round-off, for in this case the standard deviation in the round-off error in energy in one step is about $0.6\, ulp$. Figure 4 shows the histogram of the energy error at the endpoint of integration. We see that it follows a

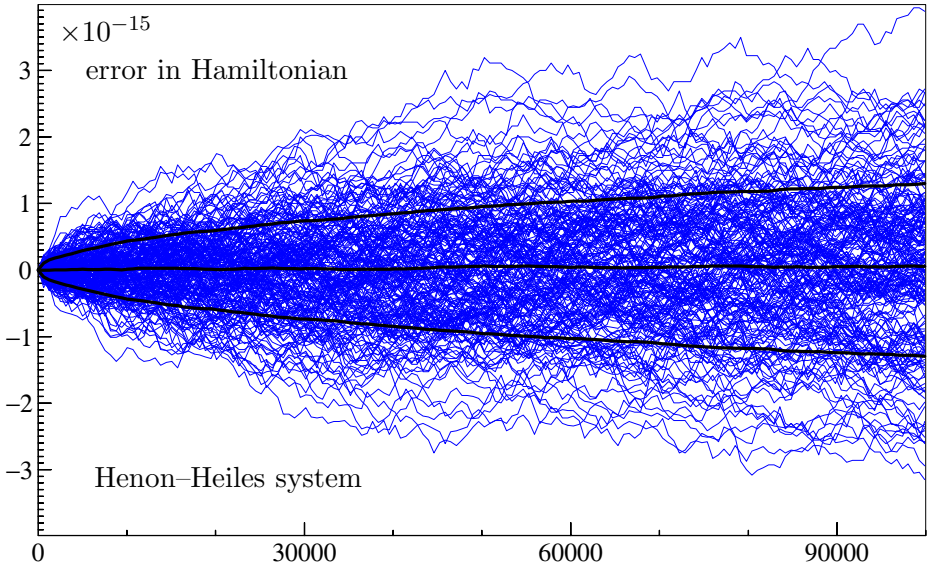

Figure 3: Energy error for Hénon–Heiles with $h = 0.25$, $H_0 = 1/8$, and 1000 initial conditions randomly chosen close to the one of Section 2. The implementation is "grk-co-ex" and the order is 12. The plot shows the error as function of time for 200 initial values. The average as a function of time ($\mu = 0.05 \times 10^{-15}$ at $t = 100\,000$) and the standard deviation ($\sigma = 1.3 \times 10^{-15}$ at $t = 100\,000$) over all 1000 trajectories are included as bold curves.

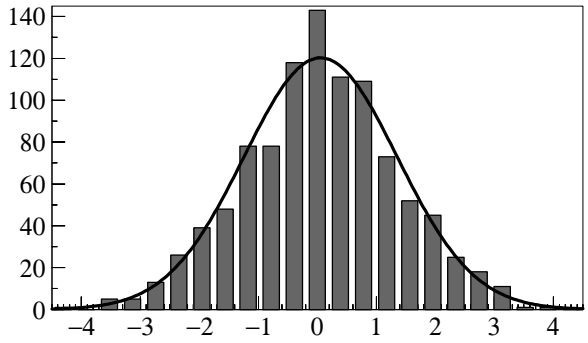

Figure 4: Histogram of energy errors at $t = 100\,000$ over 1000 samples, shown against a normal distribution with the same mean and standard deviation. The horizontal axis is in units of 10^{-15} according to Fig. 3.

normal distribution.

As a more realistic example, we consider the outer solar system (sun, the four outer planets, and Pluto). We take the data and initial values from [3, Sect. I.2.4] and modify the velocities to get zero linear momentum. Figure 5 shows the energy errors for many different initial values (we add random perturbations of size $\mathcal{O}(10^{-12})$ to the positions and keep the velocity

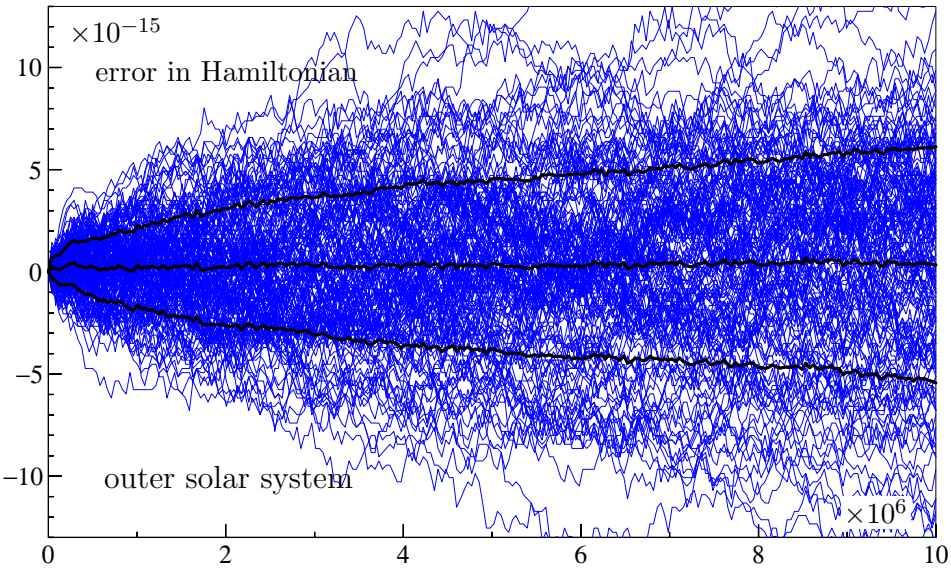

Figure 5: Energy error for the outer solar system with step size $h = 500/3$ days and 500 initial values. The implementation is "grk-co-ex" with order 12. The error as function of time is shown for 166 initial values. The average ($\mu = 0.34\times10^{-15}$ at $t = 10\,000\,000$ days) and the standard deviation ($\sigma = 5.78\times10^{-15}$ at $t = 10\,000\,000$ days) over all 500 trajectories are included as bold curves.

unchanged). Due to the larger complexity of the differential equation, the error is slightly larger than in the previous experiment for the Hénon–Heiles equation. However, the qualitative behaviour (Brouwer's law) is exactly the same. The same error growth of round-off can also be observed for the angular momentum.

Notice that for the initial values of [3, Sect. I.2.4] the linear momentum is non-zero, so that the positions and hence also the round-off error in the evaluation of the vector field increase linearly with time. In this case, the round-off error in the Hamiltonian is expected to grow like $t^{3/2}$. Brouwer's law can be satisfied only if the numerical solution remains in a compact set.

6 Conclusions

Implicit Runge–Kutta methods (based on Gauss quadrature) have a large potential for an accurate computation in geometric integration:

- Methods of arbitrarily high order are available; for efficiency reason it is important to use high order methods (order 8 and higher) for computations close to machine accuracy. For quadruple precision a much higher order of the methods is recommended.
- For expensive vector field evaluations, all s stages in the Runge–Kutta formulas can be evaluated in parallel.
- Implicit Runge–Kutta methods can be applied to general differential equations. In the case of Hamiltonian systems, the Hamiltonian does not need to be separable.

The present article shows that care has to be taken in the implementation of implicit Runge–Kutta methods. A standard straight-forward implementation will produce an undesired linear growth of round-off errors in first integrals such as the total energy. We have presented an implementation that leads to a minimal growth of round-off errors. This is not only important for computations when the local truncation error is smaller than round-off (all experiments of this paper are of this type to emphasize the effect of round-off), but also when the local truncation error is larger but close to the machine epsilon. Since for symplectic methods the energy error in exact arithmetic remains essentially bounded, it will eventually be dominated by round-off. The implicit Runge–Kutta code "grk-co-ex" can be downloaded from the Internet at the homepage <http://www.unige.ch/∼hairer/preprints.html>.

Acknowledgement. This work was partially supported by the Fonds National Suisse, project No. 200020-109158. Numerical experiments have first been presented at the Castellón Conference on Geometric Integration in

September 2006. The new algorithms and the explanations have been elaborated while two of the authors have participated at the HOP program (spring 2007) at the Isaac Newton Institute, Cambridge.

References

[1] D. Brouwer. On the accumulation of errors in numerical integration. *Astronomical Journal*, 46:149–153, 1937.

[2] K. R. Grazier, W. I. Newman, J. M. Hyman, P. W. Sharp, and D. J. Goldstein. Achieving Brouwer's law with high-order Störmer multistep methods. *ANZIAM J.*, 46:C786–C804, 2004/05.

[3] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics 31. Springer-Verlag, Berlin, 2nd edition, 2006.

[4] P. Henrici. The propagation of round-off error in the numerical solution of initial value problems involving ordinary differential equations of the second order. In *Symposium on the numerical treatment of ordinary differential equations, integral and integro-differential equations (Rome, 1960)*, pages 275–291. Birkhäuser, Basel, 1960.

[5] P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley & Sons Inc., New York, 1962.

[6] J. Laskar, P. Robutel, F. Joutel, M. Gastineau, A. C. M. Correia, and B. Levrard. A long-term numerical solution for the insolation quantities of the earth. *Astron. Astrophys.*, 428:261–285, 2004.

[7] J.-M. Petit. Symplectic integrators: rotations and roundoff errors. *Celestial Mech. Dynam. Astronom.*, 70(1):1–21, 1998.

[8] G. D. Quinlan. Round-off error in long-term orbital integrations using multistep methods. *Celestial Mech. Dynam. Astronom.*, 58(4):339–351, 1994.