A quantitative probabilistic investigation into the accumulation of rounding errors in numerical ODE solution

Sebastian Mosbach [a,1], Amanda G. Turner [b,2]

[a] *Department of Chemical Engineering, University of Cambridge, Pembroke Street, Cambridge CB2 3RA, United Kingdom*

[b] *Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, United Kingdom*

Abstract

We examine numerical rounding errors of some deterministic solvers for systems of ordinary differential equations (ODEs) from a probabilistic viewpoint. We show that the accumulation of rounding errors results in a solution which is inherently random and we obtain the theoretical distribution of the trajectory as a function of time, the step size and the numerical precision of the computer. We consider, in particular, systems which amplify the effect of the rounding errors so that over long time periods the solutions exhibit divergent behaviour. By performing multiple repetitions with different values of the time step size, we observe numerically the random distributions predicted theoretically. We mainly focus on the explicit Euler and fourth order Runge-Kutta methods but also briefly consider more complex algorithms such as the implicit solvers VODE and RADAU5 in order to demonstrate that the observed effects are not specific to a particular method.

Key words: Rounding errors, Markov jump processes, numerical ODE solution, limit theorem, saddle fixed point

1 Introduction

Consider ordinary differential equations (ODEs) of the form

$$\dot{x}_t = b(x_t).$$

[1] Corresponding author, *Email address:* `sm453@cam.ac.uk`

[2] *Email address:* `a.g.turner@lancaster.ac.uk`

Preprint submitted to Elsevier *13 March 2009*

These can be solved numerically using iteration methods of the type

$$x_{t+h} = x_t + \beta(h, x_t),$$

where $\beta(h, x)/h \to b(x)$ as $h \to 0$.

The simplest example is the Euler method, where $\beta(h, x) = hb(x)$. This method is generally not used in practice as it is relatively inaccurate and unstable compared to other methods. However, more useful methods, such as the fourth order Runge-Kutta formula (RK4), also fall into this scheme.

When solving an ordinary differential equation numerically, each time an iteration is performed an error ϵ is incurred due to rounding i.e.

$$X^h_{t+h} = X^h_t + \beta(h, X^h_t) + \epsilon \tag{1}$$

(discussed in more detail in section 2).

Rounding errors in numerical computations are an inevitable consequence of finite precision arithmetic. The first work thoroughly analyzing the effects of rounding errors on numerical algorithms is the classical textbook by Wilkinson [1]. A recent comprehensive treatment of the behaviour of numerical algorithms in finite precision, including an extensive list of references, can be found in Higham [2]. Although rounding errors are not random in the sense that the exact error incurred in any given calculation is fully determined (see Higham [2] or Forsythe [3]), in many situations probabilistic models have been shown to adequately describe their behaviour. In fact, statistical analysis of rounding errors can be traced back to one of the first works on rounding error analysis by Goldstine and von Neumann [4].

Henrici [5–7] proposes a probabilistic model for individual rounding errors whereby they are assumed to be independent and uniform, the exact distribution depending on the specific finite precision arithmetic being used. Using the central limit theorem, he shows that the theoretical distribution of the error accumulated after a fixed number of steps in the numerical solution of an ODE is asymptotically normal with variance proportional to h^{-1}. By varying the initial conditions, he obtains numerical distributions for the accumulated errors with good agreement. Hull and Swenson [8] test the validity of the above model by adding a randomly generated error with the same distribution at each stage of the calculation, and comparing the distribution of the accumulated errors with those obtained purely by rounding. They observe that, although rounding is neither a random process nor are successive errors independent, probabilistic models appear to provide a good description of what actually happens.

We shall concentrate on floating point arithmetic, as used by modern computers. However, our methods can be used equally well for any finite precision arithmetic. We use the model, discussed and tested by the authors cited above, whereby under generic conditions the errors in (1) can be viewed as independent, zero mean, uniform random variables,

$$\epsilon_i \sim U[-|X^h_{t,i}|2^{-p}, |X^h_{t,i}|2^{-p}],$$

p being a constant determined by the precision of the computer.

The purpose of this paper is to analyze the cumulative effect of these rounding errors as the step size h tends to 0. Where previous authors have considered the accumulated error at a particular point, we derive a theoretical model for the entire trajectory. In order to do so it is necessary to consider long-time behaviour which has been previously largely unexplored due to difficulties with rigorous analysis. We show for a particular system that on these time scales, the trajectories exhibit genuine randomness. We obtain the distribution of the trajectories analytically and verify our results in numerical experiments.

In general, using a smaller step size h reduces truncation errors. At the same time this necessitates a larger number of steps in order to solve the ODE numerically on a given compact time interval, thereby increasing the accumulation of round-off errors. This gives rise to a central limit theorem as shown in [5–7]. However, randomness can be seen for time scales much longer than would be expected purely from this theory. In order to observe the occurrence of randomness on large time scales it is necessary to consider ODEs whose solutions cover a finite distance in infinite time. This restricts us to systems containing a fixed point with either a periodic orbit or a stable manifold. Cases with periodic orbits have been studied for example in [9], [10], and [11]. Fixed points with only stable manifolds are of limited relevance in this respect as errors are damped and so have little effect on the qualitative behaviour of the system. We therefore investigate the class of ODEs with a saddle fixed point and initial condition on the stable manifold. Even though, as in [11], the initial condition is chosen on a set of measure zero, the solutions are of interest as they appear to exhibit strong statistical properties.

We show for an ODE in $\mathbb{R}^2$ with a saddle fixed point at the origin that the structure of the system amplifies the effect of the rounding errors and causes the numerical solution to diverge from the actual solution. More precisely, there exists a constant c, determined by the ODE system, such that for times much smaller than $-c \log h$ the numerical solution converges to the actual solution; for times close to $-c \log h$ the solution undergoes a transition, determined by a Gaussian random variable whose distribution is obtained; for times much larger than $-c \log h$ the numerical solution diverges from the actual solution.

In the first half of the paper, we outline how rounding errors can be modelled as random variables with specified distributions. We then show that the accumulation of the rounding errors results in a random trajectory. By calculating its theoretical distribution as an explicit function of time, the step size h, and the precision of the computer, we explain the qualitative behaviour described above.

In the second half of the paper, we carry out numerical simulations which illustrate this behaviour. By performing multiple repetitions with different values of the time step size, the random distributions predicted theoretically are observed. Where previous authors have obtained their numerical distributions by varying the initial conditions, we do so by introducing small variations in the step size h. During the transition period described in the previous paragraph, the numerical solution intersects straight lines through the origin and we compare the theoretical and numerical distributions for the points at which these intersections occur. Both the mean and the standard deviation of these distributions are of the form ah^{γ}, where $\gamma \in (0, 1/2]$ is a constant determined by the ODE system, and a can be found explicitly in terms of the precision of the computer, i.e. the number of bits used internally by the computer to represent floating point numbers. We mainly focus on the explicit Euler and RK4 methods, but show that the same behaviour is also observable for more complex algorithms such as the adaptive solvers VODE [12] and RADAU5 [13].

2 Theoretical background

In the paper by Turner [14], limiting results are established for sequences of Markov processes that approximate solutions of ordinary differential equations with saddle fixed points. We shall outline these results and then show that by modelling the rounding errors as random variables, the solutions obtained when performing numerical schemes for solving ordinary differential equations can be viewed as a special case of this. This enables us to quantify how the rounding errors accumulate. The resulting numerical solutions exhibit random behaviour, the exact distribution of which is obtained.

In Section 2.1 we summarize the results of Turner [14]. In Section 2.2 we describe how rounding errors can be modelled as random variables with specified distributions. The results of [14] are applied to obtain a qualitative description of the accumulation of the rounding errors. The distribution is calculated explicitly in Section 2.3.

2.1 *Behaviour of stochastic jump processes*

We are interested in ordinary differential equations of the form

$$\dot{x}_t = b(x_t). \tag{2}$$

We focus on $\mathbb{R}^2$ in the case where the origin is a saddle fixed point of the system i.e. $b(x_t) = Bx_t + \tau(x_t)$, where B is a matrix with eigenvalues $\lambda, -\mu$, with $\lambda, \mu > 0$ and $\tau(x) = O(|x|^2)$ is twice continuously differentiable. This case is of particular interest as the structure of the system amplifies the effect of the rounding errors and causes the numerical solution to diverge from the actual solution over large times. Similar behaviour can be observed in higher dimensions where the matrix B has at least one positive and one negative eigenvalue, although the corresponding quantitative analysis is much harder and we do not go into it here.

The phase portrait of (2) in the neighbourhood of the origin is shown in Figure 1. In particular, there exists some $x_0 \neq 0$ such that $\phi_t(x_0) \to 0$ as $t \to \infty$, where ϕ is the flow associated with the ordinary differential equation (2). The set of such x_0 is the stable manifold. There also exists some x_∞ such that $\phi_t^{-1}(x_\infty) \to 0$ as $t \to \infty$. The set of such x_∞ is the unstable manifold.

[Fig. 1 about here.]

Fix an x_0 in the stable manifold and consider sequences X_t^N of Markov processes starting from x_0, which converge to the solution of (2) over compact time intervals. The processes are indexed so that the variance of the fluctuations of X_t^N is inversely proportional to N. If we allow the value of t to grow with N as a constant times $\log N$, X_t^N deviates from the stable solution to a limit which is inherently random, before converging to an unstable solution (see Figure 2).

[Fig. 2 about here.]

More precisely, we observe three different types of behaviour depending on the time scale:

A. On compact time intervals, X_t^N converges to the stable solution of (2), the fluctuations around this limit being of order $N^{-\frac{1}{2}}$. The exact distribution of the fluctuations is asymptotically $N^{-\frac{1}{2}}\gamma_t$ where γ_t is the solution to a linear stochastic differential equation, described in [14].

B. Let v_1 and v_2 be the unit eigenvectors of B corresponding to $-\mu$ and λ respectively. There exists some $\overline{x}_0 \neq 0$, depending only on x_0, and a Gaussian random variable Z_∞, such that if t lies in the interval $[R, \frac{1}{2\lambda}\log N -$

$R]$, then

$$X_t^N = \overline{x}_0 e^{-\mu t}(v_1 + \epsilon_1) + N^{-\frac{1}{2}} Z_\infty e^{\lambda t}(v_2 + \epsilon_2)$$

where $\epsilon_i(t, N) \to 0$ uniformly in t in probability as $R, N \to \infty$. In other words, X_t^N can be approximated by the solution to the linear ordinary differential equation

$$\dot{y}_t = B y_t \tag{3}$$

starting from the random point $\overline{x}_0 v_1 + N^{-\frac{1}{2}} Z_\infty v_2$.

C. Provided $Z_\infty \neq 0$, on time intervals of a fixed length around $\frac{1}{2\lambda}\log N$, X_t^N converges to one of the two unstable solutions of (2), each with probability 1/2, depending on the sign of Z_∞.

2.2 Accumulation of rounding errors

We can apply the above results to describe quantitatively how rounding errors accumulate when solving ordinary differential equations of the form (2) numerically. In particular we consider using iteration methods of the type

$$x_{t+h} = x_t + \beta(h, x_t) \tag{4}$$

where $\beta(h, x)/h \to b(x)$ as $h \to 0$.

Each time an iteration is performed, an error $\epsilon = \epsilon(h, t)$ is incurred due to rounding, so we obtain a process $(X_t^h)_{t \in h\mathbb{N}}$ iteratively by

$$X_{t+h}^h = X_t^h + \beta(h, X_t^h) + \epsilon. \tag{5}$$

Modern computers store real numbers by expressing them in binary as $x = m2^n$ for some $1 \leqslant |m| < 2$ and $n \in \mathbb{Z}$. They allocate a fixed number of bits to store the mantissa m and a (different) fixed number of bits to store the exponent n [15]. When adding to x a number of smaller order, the size of the rounding error incurred is between 0 and $2^{n-p} = 2^{\lfloor \log_2 |x| \rfloor - p}$, where p is the number of bits allocated to store the mantissa. Although it is possible to carry out the calculations below using the exact value of $2^{\lfloor \log_2 |x| \rfloor - p}$, the calculations are greatly simplified by approximating it by $|x|2^{-p}$. This results in the 'effective' value of p differing from the actual value of p by some number between 0 and 1. Provided $\beta(h, X_t^h)$ is sufficiently small compared with X_t^h, the errors ϵ can therefore be viewed as independent, mean zero, uniform random variables with approximate distribution

$$\epsilon_i \sim U[-|X_{t,i}^h|2^{-p}, |X_{t,i}^h|2^{-p}]$$

(see Henrici [5–7]). The assumption that the ϵ_i are independent is in general not true. In fact, in certain pathological cases, for example where there is a lot of symmetry in the components, the ϵ_i can be strongly correlated. Nevertheless, under generic conditions one would expect any correlations to be weak and so this is a reasonable assumption to make. We shall see by the agreement of our numerical and theoretical results that the effect of making this assumption is indeed small.

Although the above iterations are carried out at discrete time intervals, it is convenient to embed the processes in continuous time by performing the iterations at times of a Poisson process with rate h^{-1}. As $\beta(h,x)$ does not depend on t, this does not affect the shape of the resulting trajectories. In this way Markov processes X_t^h are obtained that approximate the stable solution of (2) for small values of h. If, in addition, the assumption is made that

$$h^{-\frac{1}{2}}\left(\frac{\beta(h,x)}{h} - b(x)\right) \to 0$$

as $h \to 0$ (note that both the Euler and Runge-Kutta methods satisfy this condition), then under the correspondence $N \sim h^{-1}$, the conditions needed to apply the results in [14] are satisfied.

Our numerical solution therefore exhibits the following random behaviour:

A. For times of order much smaller than $-\log h$, X_t^h approximates the stable solution of (2), the fluctuations around this limit being of order $h^{\frac{1}{2}}$.

B. There exists some $\overline{x}_0 \neq 0$, depending only on x_0, and a Gaussian random variable Z_∞, such that if t lies in the interval $[-c\log h, -\frac{1}{2\lambda}\log h + c\log h]$ for some $c > 0$, then X_t^h is asymptotic to

$$\overline{x}_0 e^{-\mu t} v_1 + h^{\frac{1}{2}} Z_\infty e^{\lambda t} v_2, \tag{6}$$

the solution to the linear ordinary differential equation (3) starting from the random point $\overline{x}_0 v_1 + h^{\frac{1}{2}} Z_\infty v_2$.

C. Provided $Z_\infty \neq 0$, in time intervals around $-\frac{1}{2\lambda}\log h$ whose length is of much smaller order than $-\log h$, X_t^h approximates one of the two unstable solutions of (2), each with probability $\frac{1}{2}$, depending on the sign of Z_∞.

The random behaviour resulting from the accumulation of rounding errors is most noticeable on time intervals of fixed lengths around $-\frac{1}{2(\lambda+\mu)}\log h$, as for these values of t the two terms $\overline{x}_0 e^{-\mu t}$ and $h^{\frac{1}{2}} Z_\infty e^{\lambda t}$ in (6) are of the same order. During these time interval, the numerical solution undergoes a transition from converging to the actual solution to diverging from it. During this transition, for each value of $\theta \in (0, \pi/2)$, X_t^h crosses one of the straight lines passing through 0 in the direction $v_1 \cos\theta \pm v_2 \sin\theta$. These intersections are important as they indicate the onset of divergent behaviour. The distribution of the point at which X_t^h intersects one of the lines in the direction $v_1 \cos\theta \pm v_2 \sin\theta$

is asymptotic to

$$h^{\frac{\mu}{2(\lambda+\mu)}}|Z_\infty|^{\frac{\mu}{\lambda+\mu}}|\overline{x}_0|^{\frac{\lambda}{\lambda+\mu}}|\tan\theta|^{\frac{\mu}{\lambda+\mu}}(v_1\cos\theta \pm v_2\sin\theta). \qquad (7)$$

In Section 2.3 we show how to evaluate the variance of Z_∞, doing so explicitly in the linear case and obtaining bounds in the non-linear case. In Section 3 these results are verified by numerically obtaining the predicted distribution for hitting a line through the origin.

2.3 Explicit calculation of the variance

Consider a numerical scheme that satisfies the above conditions, applied to obtain a solution to the ordinary differential equation (2), starting from x_0 for some x_0 in the stable manifold. In the non-linear case we require that x_0 is sufficiently close to the origin such that $\tau(x_0)$ is small. In general, for simplicity, we assume that $|x_0| \leqslant 1$.

We define the flow ϕ associated with this system by

$$\dot{\phi}_t(x) = b(\phi_t(x)), \quad \phi_0(x) = 0$$

and let $x_t = \phi_t(x_0)$.

Suppose that v_1, $v_2 \in \mathbb{R}^2$ are the unit right-eigenvectors of B corresponding to $-\mu$, λ respectively, and that $v_1', v_2' \in (\mathbb{R}^2)^*$ are the corresponding left-eigenvectors (i.e. $v_i'v_j = \delta_{ij}$).

Define

$$\overline{x}_0 = \lim_{t\to\infty} e^{\mu t} v_1' \phi_t(x_0)$$

and

$$D_s = \lim_{t\to\infty} e^{-\lambda t} v_2' \nabla\phi_t(x_s).$$

It is shown in [14] that these limits exist and that $|\overline{x}_0| \leqslant 2|x_0| \leqslant 2$ and $|D_s| \leqslant 2$.

Finally, let

$$a(x) = \frac{1}{3}2^{-2p}\begin{pmatrix} x_1^2 & 0 \\ 0 & x_2^2 \end{pmatrix}$$

be the covariance matrix of the multivariate uniform random variable ϵ, defined in equation (5), when $X_t^h = x$. Then $Z_\infty \sim N(0, \sigma_\infty^2)$, where it is shown in [14] that

$$\sigma_\infty^2 = \int_0^\infty e^{-2\lambda s} D_s a(x_s) D_s^* ds.$$

Note that $\sigma_\infty^2 \leqslant \frac{2}{3\lambda}2^{-2p}$.

In the general non-linear case, evaluating σ_∞^2 explicitly is not possible as it involves solving (2). It is possible to obtain a better approximation than that above, although the important observation is that σ_∞^2 is proportional to 2^{-2p}.

In the linear case, $\phi_t(x) = e^{Bt}x$ and $x_0 = |x_0|v_1$. Hence $x_t = |x_0|e^{-\mu t}v_1$, $\overline{x}_0 = |x_0|$, and $D_s = v_2'$, and so

$$\sigma_\infty^2 = \frac{1}{3(\lambda+\mu)}2^{-2p}|x_0|^2(v_{1,1}v_{2,1}')^2.$$

Note that the directions of v_1 and v_2', relative to the standard basis, are critical. For example, if either v_1 or v_2' is parallel to one of the standard basis vectors, then $\sigma_\infty^2 = 0$.

3 Numerical experiments

In this section we solve ODEs numerically using deterministic solvers and observe the predicted random distributions arising as a consequence of the accumulation of rounding errors. For simplicity, and in order to observe the desired effects as clearly as possible, we mainly focus on the most elementary of all numerical ODE solution methods, the standard explicit Euler algorithm with constant time step size. However, we observe similar behaviour for RK4 and also briefly mention results obtained with more complex solvers, such as VODE [12] and RADAU5 [13].

3.1 The system

For $x : [0,\infty) \to \mathbb{R}^2$, consider the linear ODE

$$\dot{x}(t) = Bx(t),$$

where

$$B = \begin{pmatrix} -\mu & 0 \\ 0 & \lambda \end{pmatrix}$$

for fixed $\lambda, \mu > 0$. Introduce new coordinates

$$\bar{x}(t) = R(\varphi)x(t)$$

by rotating about the origin by a fixed angle $\varphi \in [0, \pi/2)$, i.e.

$$R(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

We arrive at the transformed system

$$\dot{\bar{x}}(t) = \bar{B}(\varphi)\bar{x}(t) \tag{8}$$

with

$$\bar{B}(\varphi) = R(\varphi)BR(\varphi)^\top,$$

which will be the system under consideration in the following. Throughout, the initial value

$$\bar{x}(0) = R(\varphi)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} \tag{9}$$

is used. The phase space evolution is sketched in Figure 3.

[Fig. 3 about here.]

3.2 Theoretical hitting distribution

As discussed in Section 2.2, the numerical solution to the above ODE system undergoes a transition from converging to the actual solution to diverging from it. During this transition, the numerical trajectory crosses one of the straight lines passing through 0 at an angle $\phi \pm \theta$ for each value of $\theta \in (0, \pi/2)$. These intersections are important as they indicate the onset of divergent behaviour. The hitting distributions also provide a means of measuring the random variable Z_∞, which determines the random variations in our solutions, and hence of verifying the theoretical results.

Equation (7) gives the asymptotic distribution of the magnitude of the point at which the numerical solution hits the line through the origin at angle $\varphi \pm \frac{\pi}{4}$ as $|Z|^{\frac{\mu}{\lambda+\mu}}$ where Z is a Gaussian random variable with mean 0 and variance

$$\sigma^2 = h\sigma_\infty^2 = \frac{1}{3(\lambda+\mu)} h 2^{-2p} (\cos\varphi \sin\varphi)^2 \tag{10}$$

i.e. $Z \sim \mathcal{N}(0, \sigma^2)$. We obtain an explicit formula for the asymptotic distribution by starting from the $\mathcal{N}(0, \sigma^2)$ distribution

$$p(x)\mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}x^2\Big)\mathrm{d}x$$

and performing a change of variable given by $y = |x|^{\frac{\mu}{\lambda+\mu}}$. The result is

$$p(y)\mathrm{d}y = \frac{2(\lambda+\mu)}{\sqrt{2\pi}\sigma\mu} y^{\frac{\lambda}{\mu}} \exp\Big(- \frac{1}{2\sigma^2} y^{\frac{2(\lambda+\mu)}{\mu}} \Big)\mathrm{d}y.$$

In the case $\lambda = \mu = 1$, which is considered below, setting $a = \frac{4}{\sqrt{2\pi}\sigma}$ produces the family of distributions

$$f(y)\mathrm{d}y = ay \exp\Big(- \frac{\pi}{16} a^2 y^4 \Big)\mathrm{d}y, \quad y \in (0, \infty), \tag{11}$$

which will be fitted to the numerical data to confirm the theoretical value of a.

3.3 Choice of parameters

Rounding errors are deterministic in the sense that any given number of iterations of a particular numerical scheme will generate the same solution. In order to obtain a distribution from the numerical solutions to (3), for each repetition it is necessary to vary at least one parameter by a small amount. In this section we discuss this issue as well as the choice of the fixed parameters of the system such as the eigenvalues.

The possible parameters that can be varied are the initial value x_0, and the time step size h. As x_0 is constrained to be on the stable manifold, any variation is required to be in the direction of the eigenvector corresponding to eigenvalue $-\mu$. We have found that varying the initial value in a direction orthogonal to the stable manifold does not yield any interesting results as the chosen distribution of initial values is reproduced exactly in the hitting distribution. Varying it within the stable manifold yields identical results to varying the time step size, however in terms of the system we feel it is preferable to vary the step size as this parameter is internal to the algorithm, whereas the initial value is a physical parameter of the system. We varied the time step size as follows. Given a user-supplied value of h, define the step size h_i for the i^{th} repetition by

$$h_i = h + \Delta h(i - 1 - k), \quad i = 1, \ldots, L,$$

where the number of repetitions $L = 2k+1$ and $0 < \Delta h \ll h$ are user-supplied. For all simulations, we set $k = 10^4$.

Reasonable choices of h and Δh are limited by several factors. The hitting distribution predicted theoretically in Section 3.2 is asymptotic as $h \to 0$ and hence, if h is too large (in the considered case, if $h > 10^{-1}$ for both single and double precision), the observed hitting distribution differs substantially from the theoretical one. The onset of such effects can be seen for large values of h in Figure 6. Lower bounds on h are imposed by computational cost and by the numerical precision of the computer. In practice, computational expense becomes prohibitive for values of h much larger than the smallest values permitted by numerical accuracy. Our particular choice of step size distribution requires that $k\Delta h$ should be (much) smaller than h. The lower limit for Δh is determined solely by the numerical precision, i.e. $\Delta h/h$ must not be smaller than the numerical precision.

We did not investigate in detail the dependence of our observations on the distribution of step sizes. However, preliminary experiments with varying Δh and even with non-uniform step size distributions suggest that this dependence is very weak for a wide range of conditions. Figure 4 shows that the shape of the distribution exhibits no discernible systematic dependence on Δh over at least nine orders of magnitude. The deviations seen for values of Δh smaller than about 10^{-19} are due to the fact that $\Delta h/h$ approaches the limits of numerical precision.

[Fig. 4 about here.]

The remaining parameters that we need to choose are the eigenvalues $\lambda, -\mu$ and the rotation angle φ. Since the limit distribution is given by $|Z|^{\frac{\mu}{\lambda+\mu}}$, for some Gaussian random variable Z, if the values of λ and μ differ significantly then the distribution is hard to observe in a numerical experiment. This suggests choosing λ and μ of the same order of magnitude, and we therefore take $\lambda = \mu = 1$ for all simulations.

There is some subtlety in the choice of the rotation angle φ. For certain values, trivial trajectories or symmetry effects can occur which conceal the desired accumulation of rounding errors. For instance, for $\varphi = 0$ the second component $\bar{x}_2$ of the solution is always zero, and therefore the trajectory stays on the line $\bar{x}_2 = 0$ (or equivalently $x_2 = 0$) with no fluctuations. Note that this is in agreement with $\sigma^2 = 0$ in equation (10). For $\varphi = \pi/4$, any rounding error that appears in one component also appears in the other one, which implies that, again, the trajectory always stays on the line $\bar{x}_2 = 0$ (or equivalently $x_1 = x_2$). This case is pathological as it consistently violates our assumption that the rounding errors for the different components are independent. For these reasons, we chose $\varphi = \pi/5$ throughout.

3.4 Results and observations for explicit methods

Using the values of the parameters discussed above, we carried out multiple repetitions of Euler's algorithm and RK4. In each run we noted the point at which the trajectory given by the numerical solution intersected one of the lines $\bar{x}_1 = \pm\bar{x}_2$ (the dashed lines in Figure 3). Histograms were then produced by partitioning the interval $[0, 1]$ into a given fixed number of subintervals of equal length and counting how many times y fell into each subinterval, where y denotes the distance of the point of intersection from the origin. The empirical distributions shown in Figure 5 were obtained. The theoretical distribution (11) was fitted to the empirical distributions with very good agreement.

[Fig. 5 about here.]

For each value of h, we obtained a value for the parameter a by fitting a distribution of the form (11) to our numerical data. In Figure 6 the parameter a is plotted as a function of the time step size h, both for single (Figure 6(a)) and double (Figure 6(b)) precision (4 and 8 bytes internal representation of floating point numbers respectively). Error bars due to the fit are only about 1% and hence insignificant. In both cases, the dependence between a and h is well described by $a \propto \sqrt{h}$.

[Fig. 6 about here.]

Equation (10) predicts the value of $ah^{-\frac{1}{2}}$ to be

$$ah^{-\frac{1}{2}} = \frac{4\sqrt{3}}{\sqrt{\pi}\cos\frac{\pi}{5}\sin\frac{\pi}{5}} \times 2^p = 8.220 \times 2^p.$$

For Euler's method, the above data give $ah^{-\frac{1}{2}} = 9.411 \times 10^7$ for single precision and $ah^{-\frac{1}{2}} = 4.956 \times 10^{16}$ for double precision. For the 4th order Runge-Kutta method, the values are $ah^{-\frac{1}{2}} = 9.27 \times 10^7$ (with a relatively large error of $\pm 0.12 \times 10^7$) for single precision and $ah^{-\frac{1}{2}} = 4.746 \times 10^{16}$ for double precision. Using the approximation discussed in Section 2.2, the actual value of p is between 23 and 24, when working in single precision, and between 52 and 53 when working in double precision. The particular value depends on the exact number being computed. Our theoretical results therefore predict $ah^{-\frac{1}{2}}$ lies between 6.895×10^7 and 1.379×10^8 for single precision and between 3.702×10^{16} and 7.404×10^{16} for double precision.

There are three possible sources of error in our calculations. The first is the error in fitting the numerical data to the theoretical model, the second is that our theoretical models are based on asymptotic results as $h \to 0$, whereas we are applying them to values of h which are necessarily larger than the precision of the computer. The third source of error arises from the assumption that at each stage the rounding error can be viewed as an independent uniform random variable, depending on a fixed value of p. The above results show that these errors are all small and that our theoretical model provides a very good fit.

3.5 Adaptive solvers

Our theoretical results cover ODE solvers which use algorithms of the form (4). In practice, more sophisticated adaptive solvers are used, such as VODE [12] and RADAU5 [13]. For these solvers, the user inputs the error tolerances `RTOL` (relative) and `ATOL` (absolute) and the global time step h_{g} (the time interval after which the user requests solution output from the solver). However, the user has no immediate control over the size of the actual steps taken. These are determined algorithmically as a function of the error tolerance parameters `RTOL` and `ATOL`, generally by trial-and-error methods using heuristics, rather than by an explicit formula.

Although it is not possible to analyze such adaptive solvers in the way that we have analyzed explicit solvers above, it is still of interest to see whether they exhibit the same qualitative random behaviour. We performed numerical experiments similar to those discussed above and obtained the distributions shown in Figure 7 in the case where `RTOL`=0.

[Fig. 7 about here.]

Experiments do not readily suggest a simple relationship between the parameter a in equation (11) and any of the parameters `ATOL`, `RTOL`, and h_{g}. This is possibly not surprising given the lack of direct control over the time step size. However, the fact that the results are qualitatively similar supports the assertion that the observed phenomena are not specific to a particular algorithm, but rather are general effects.

4 Conclusion

We analyzed the cumulative effect of rounding errors incurred by deterministic ODE solvers as the step size $h \to 0$. We considered in particular the interesting case where the ordinary differential equation has a saddle fixed point and showed that the numerical solution is inherently random and also obtained its theoretical distribution in terms of the time, step size and numerical precision. We showed that as the step size $h \to 0$, the numerical solution exhibits three types of behaviour, depending on the time: initially it converges to the actual solution, it then undergoes a transition stage, finally it diverges from the actual solution.

By performing multiple repetitions with different values of the time step size, we observed the random distributions predicted theoretically. We demonstrated that during the transition period described above the numerical solution intersects all the straight lines through the origin. The theoretical and numerical distributions for the points at which these intersections occur showed very good agreement. Both the mean and the standard deviation of these distributions were found to be of the form ah^γ, where $\gamma \in (0, 1/2]$ is a constant determined by the ODE system, and a was found explicitly in terms of the precision of the computer. We mainly focused on the explicit Euler and RK4 methods with constant step size, but also briefly considered the implicit solvers VODE and RADAU5 with automatic step adaption in order to demonstrate that the observed effects are not specific to a particular numerical method.

Acknowledgments

This work has been partially funded by the EPSRC (grant number GR/R85662/01) under the title "Mathematical and Numerical Analysis of Coagulation-Diffusion Processes in Chemical Engineering". The authors thank James R. Norris and Markus Kraft for suggesting the project and the collaboration, and helpful discussions.

References

[1] J. H. Wilkinson. *Rounding Errors in Algebraic Processes*, volume 32 of *Notes on Applied Science*. Her Majesty's Stationery Office, London, 1963. Also published by Prentice-Hall, NJ, USA. Reprinted by Dover, New York, 1994.

[2] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, 1996.

[3] G. E. Forsythe. Reprint of a note on rounding-off errors. *SIAM Rev.*, 1(1):66-67, 1959.

[4] H. H. Goldstine and J. von Neumann. Numerical inverting of matrices of high order II. *Proc. Amer. Math. Soc.*, 2:188-202, 1951.

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

[6] P. Henrici. *Error Propagation for Difference Methods.* John Wiley & Sons, New York, 1963.

[7] P. Henrici. *Elements of Numerical Analysis.* John Wiley & Sons, New York, 1964.

[8] T. E. Hull and J. R. Swenson. Test of probabilistic models for the propagation of roundoff errors. *Comm. ACM*, 9(2):108-113, 1966.

[9] M. Blank. Pathologies generated by round-off in dynamical systems. *Physica D*, 78:93-114, 1994.

[10] J. H. Lowenstein and F. Vivaldi. Anomalous transport in a model of Hamiltonian round-off. *Nonlinearity*, 11:1321-1350, 1998.

[11] F. Vivaldi and I. Vladimirov. Pseudo-randomness of round-off errors in discretized linear maps on the plane. *Int. J. of Bifurcations and Chaos*, 13:3373-3393, 2003.

[12] P. N. Brown, G. D. Byrne, and A. C. Hindmarsh. VODE, a variable-coefficient ODE solver. *SIAM Journal on Scientific and Statistical Computing*, 10(5):1038-1051, 1989.

[13] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer Verlag, Berlin Heidelberg New York, second revised edition, 1996.

[14] A. G. Turner. Convergence of Markov processes near saddle fixed points. *Annals of Probability*, 35(3):1141-1171, 2007.

[15] IEEE standard for binary floating-point arithmetic, ANSI/IEEE Standard 754-1985. Institute of Electrical and Electronics Engineers, 1985. Reprinted in SIGPLAN Notices, 22(2):9-25, 1987.

List of Figures

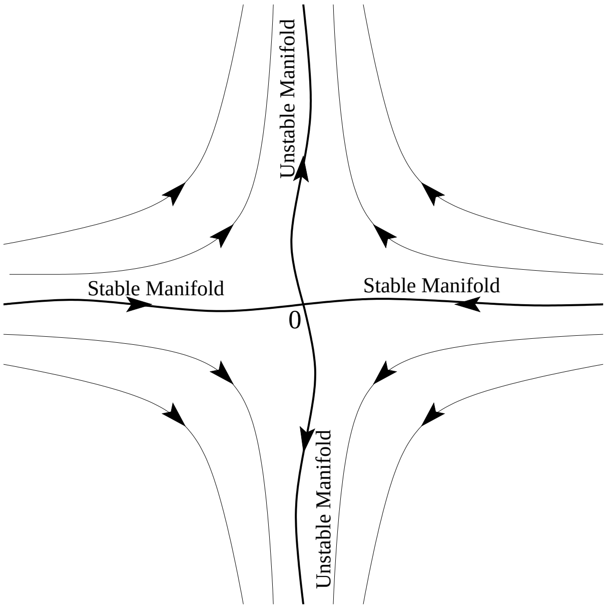


Fig. 1. *The phase portrait of an ordinary differential equation having a saddle fixed point at the origin (taken from [14]).*

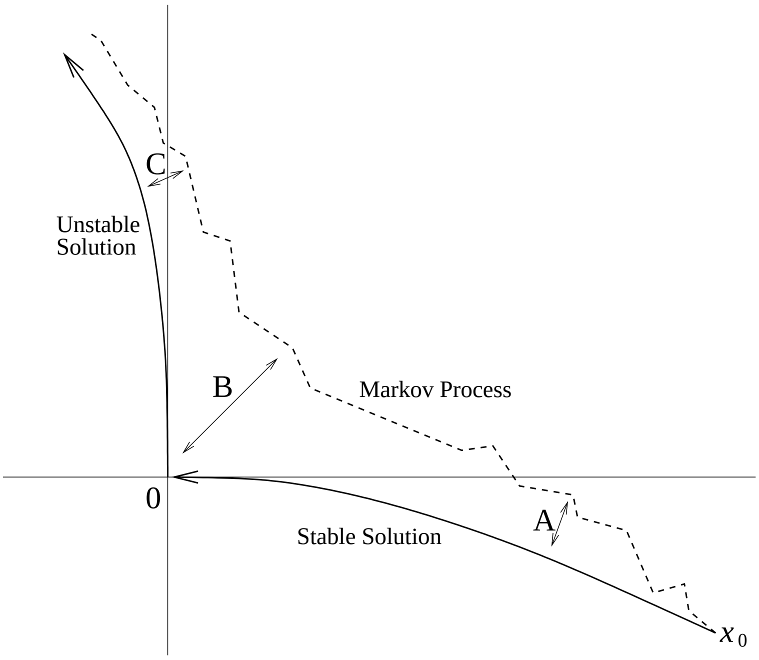


Fig. 2. *Diagram showing how the Markov process X_t^N deviates from the stable solution $\phi_t(x_0)$ for large values of t (taken from [14]).*

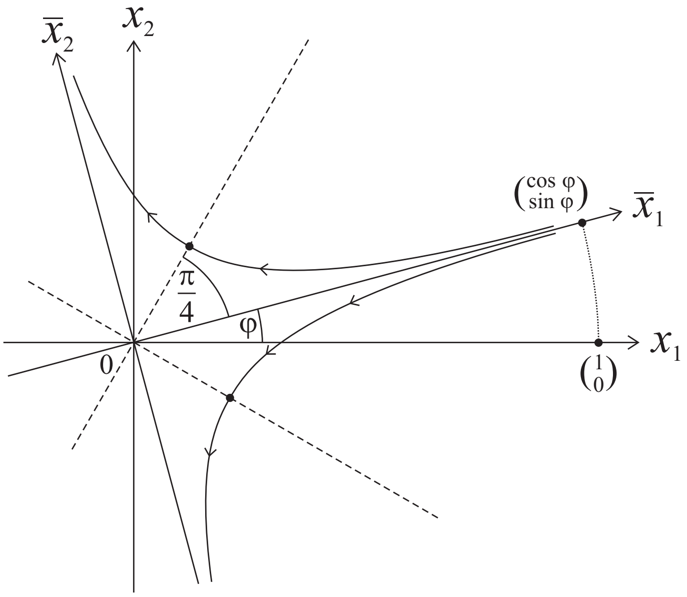


Fig. 3. *Phase space for the saddlepoint ODE system* (8) *with sample trajectories and lines where hitting distributions are recorded (dashed lines).*

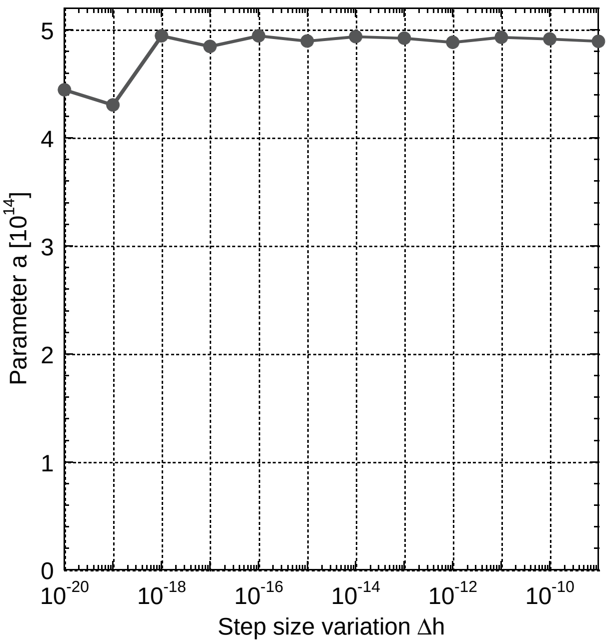


Fig. 4. *Step size variation for Euler's algorithm (double precision, step size* $h = 10^{-4}$*,* $L = 20001$ *repetitions each).*

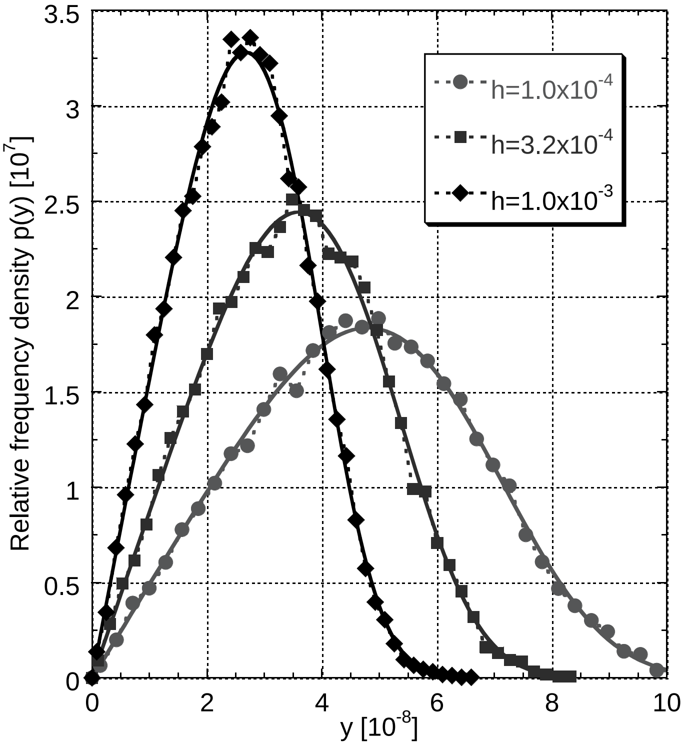
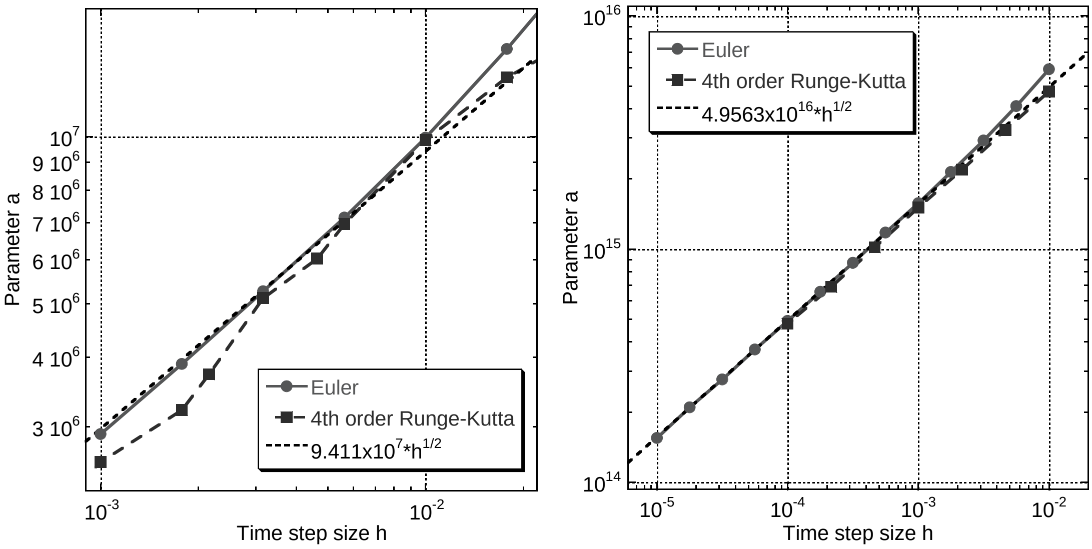


Fig. 5. *Observed hitting distributions (symbols with dotted lines) with theoretical fits (solid lines) for Euler's algorithm ($\Delta h = 10^{-10}$, $L = 20001$ repetitions each).*



(a) Single precision ($\Delta h = 10^{-8}$).

(b) Double precision ($\Delta h = 10^{-10}$).

Fig. 6. *Parameter a in equation* (11) *as function of the time step size h for simple explicit methods (Euler and 4^{th} order Runge-Kutta).*

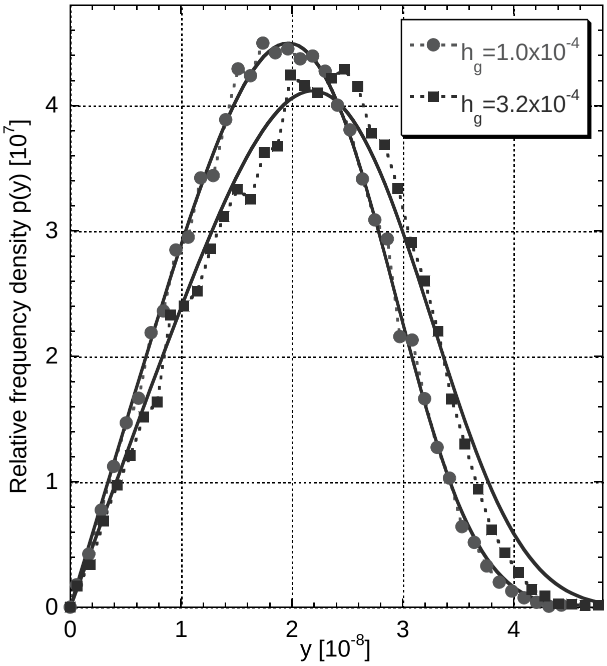


Fig. 7. *Hitting distributions for VODE.*