ELSEVIER

Applied Numerical Mathematics 29 (1999) 3–18

APPLIED NUMERICAL MATHEMATICS

Symplectic integration with floating-point arithmetic and other approximations ☆

Robert D. Skeel

Department of Computer Science and Beckman Institute, University of Illinois at Urbana-Champaign, 1304 West Springfield Avenue, Urbana, IL 61801-2987, USA

Abstract

There is significant theoretical support for the use of symplectic integrators for the numerical solution of Hamiltonian systems. However, the theory does not apply to practical computations because of the failure to take into account the effects of roundoff errors, and other approximations such as the use of fast N-body solvers and the use of "not fully converged iterations" in implicit or semi-implicit integrators. Very often these effects grow exponentially with time and completely overwhelm the numerical results well before the integration is complete. By means of a simple and inexpensive modification of the integrator, we show that it is possible to maintain a symplectic integration with floating-point arithmetic and other approximations. © 1999 Elsevier Science B.V. and IMACS. All rights reserved.

Keywords: Symplectic integration; Floating-point arithmetic; Lattice maps; Leapfrog method

1. Introduction

There is significant theoretical support for the use of symplectic integrators for the long-time numerical solution of Hamiltonian systems [16]. However, the theory does not apply to practical computations because of the failure to take into account the effects of roundoff errors and other approximations such as the use of fast N-body solvers and the use of "not fully converged iterations" in implicit or semi-implicit integrators. Very often these effects grow exponentially with time and completely overwhelm the numerical trajectories well before the integration is complete. By means of a simple and inexpensive modification of the integrator, we show that it is possible to maintain a symplectic integration with floating-point arithmetic and other approximations. This is an extension of work by Scovel [20], and of others cited there, on the use of integer lattices for symplectic integration.

A Hamiltonian system has the form

$$\frac{\mathrm{d}q}{\mathrm{d}t} = H_p(q, p), \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -H_q(q, p),$$

☆ This work was supported in part by NSF Grant DMS-9600088, NSF Grant BIR-9318159, and NIH Grant P41RR05969.

PII: S0168-9274(98)00033-6

where $q(0)$ and $p(0)$ are given. Here, $q = [q_1, q_2, \ldots, q_N]^T$ are position variables and $p = [p_1, p_2, \ldots, p_N]^T$ are momentum variables. In particular, biological molecules and their environments, which are too large for quantum mechanical simulation, are modeled by a classical mechanical Hamiltonian

$$H(q, p) = \tfrac{1}{2} p^T M^{-1} p + V(q),$$

where $V(q)$ is an empirical potential energy function. Letting $F(q) = -V_q(q)$, the system can be written

$$\frac{dq}{dt} = M^{-1} p, \qquad \frac{dp}{dt} = F(q).$$

For molecular dynamics, perturbations in initial conditions grow exponentially in time, see [1, Fig. 3.1] and [8], and even with double precision arithmetic the effects of roundoff error overwhelm the trajectories early in the simulation.

With floating-point arithmetic, phase space is a discrete set of points, each point a $2N$-tuple of machine numbers. More specifically, phase space is the union of lattices

$$2^{-1074}L \cup 2^{-1073}L \cup \cdots \cup 2^{971}L,$$

where L is the set of all $2N$-tuples of the integers $1 - 2^{53}, 2 - 2^{53}, \ldots, 2^{53} - 1$, assuming IEEE standard double precision normalized and denormalized floating-point numbers [21, p. 40]. In a finite phase space, the computed solution must eventually either repeat or go out of range. An orbit that does not go out of range consists of a nonrepeated transient followed by a closed repeated orbit. This could be a limit cycle or an equilibrium point (i.e., a limit cycle of one point). For a Hamiltonian system with no integrals other than energy, a limit cycle represents a computer version of an energy surface.

Building on work of others, it is shown in [20] that if the exact numerical solution of the symplectic Euler method is rounded to a *uniformly* spaced lattice in phase space after each stage of the method, then the mapping is one-to-one and there are no transients. Every lattice point is part of a cycle. (The symplectic Euler method is equivalent to the leapfrog/Störmer/Verlet method [24].) Hence, for a problem such as the simple pendulum, if we use a symplectic lattice method, the computer model will swing back and forth in perpetuity just like the analytical mathematical model even in the presence of roundoff error! This may seem to be only a curiosity, possibly useful for computer graphics or computer games; however, it might also matter for large-scale scientific computing.

More interesting yet, it is shown in [20] that the lattice symplectic Euler method is equivalent to applying the infinite-precision symplectic Euler method to a slightly different Hamiltonian system. What makes such a result possible is the freedom to define the perturbed Hamiltonian $\widehat{H}$ at off-lattice points however we wish.

It is natural to wonder whether the use of a floating-point lattice might be an adequate approximation to a "fixed-point lattice". After all, a limit cycle will be reached with a floating-point lattice after the transient disappears. However, it is an open question whether or not the floating-point limit cycle is time-reversible, and nothing is known about the "reduced" floating-point lattice constructed from only those points that are part of limit cycles. One investigation [5] indicates that lattice maps are superior to floating-point maps for long-time studies of symplectic maps. More specifically, experiments were performed showing that transients can be very long and wander all over phase space; examples in single precision include a transient of 12 534 871 steps followed by a limit cycle of length 10 458 863 and a transient of 6 300 097 steps followed by a limit cycle of length 54 264. Moreover, it is shown that the limit cycle can differ markedly from the initial transient. As described in [6], "roundoff errors cause artificial drifting across invariant curves in Hamiltonian maps and can even lead to confusion between

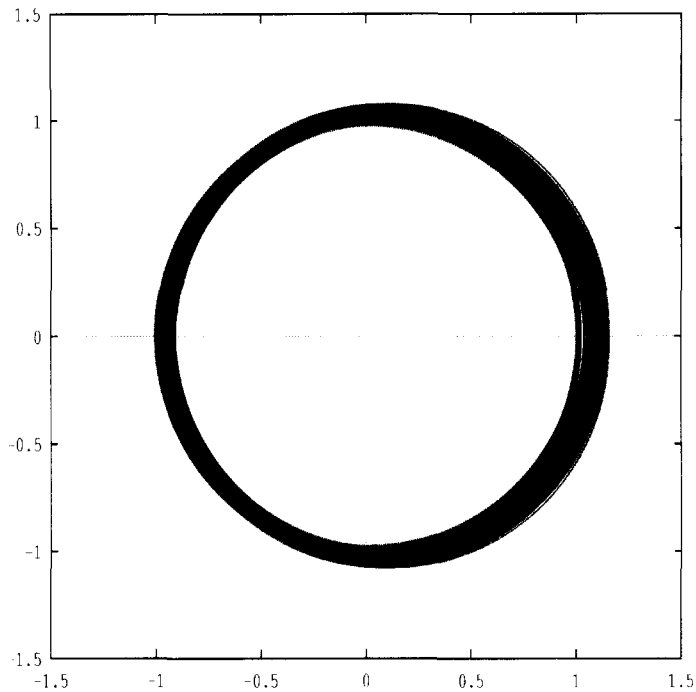

Fig. 1. Projection onto position coordinates of infinite circular Kepler orbit computed by lattice leapfrog method.

regularity and chaos". It is not obvious that this is compelling evidence against the use of floating-point lattices; however, the existence of a simple and highly efficient fixed-point lattice integration algorithm would strengthen the case for the use of fixed-point lattices.

In Section 2, we define a one-parameter family of symplectic integrators [22] that includes leapfrog/Störmer/Verlet, Cowell/Numerov, implicit midpoint/trapezoid, and LIM2 [25]. In Section 3, we devise a very efficient floating-point implementation of such a symplectic integrator, which requires only a simple change or two to the usual floating-point version of the algorithms. In particular, it avoids the need to scale to an integer lattice, to convert between integer and floating-point representation, and to use validated arbitrarily high precision for intermediate results. As an example, we apply the lattice leapfrog method with lattice spacing $\varepsilon = 2^{-16}$ to the Kepler problem in Cartesian coordinates, $M = I$ and $V(q_1, q_2) = -(q_1^2 + q_2^2)^{-1/2}$, and we show in Fig. 1 the computed positions q_1 and q_2 for an infinite number of steps! Initial conditions are $q_1(0) = 1$, $p_1(0) = 0$, $q_2(0) = 0$, $p_2(0) = 1$, and the stepsize is $h = 0.01$.

Also in Section 3 is given an explicit construction of the perturbation to the Hamiltonian due to finite precision and an extension of previous error analyses, in which explicit bounds are obtained for the modification to the Hamiltonian, its gradient, and its Hessian. It is noted that the perturbation term will have large higher derivatives, which might seriously weaken the backward error analysis interpretation of the symplectic integrator.

Section 4 gives results of an experimental study of the effect of finite precision on infinite-time behavior. The thickness of the computational invariant manifolds is an indication of resolution, and limited evidence suggests that this thickness remains more or less constant as the precision increases.

More serious errors than arise from roundoff are those that result from (i) the inexact solution of the equations that arise if one should use an implicit method, (ii) the use of cutoffs to neglect forces when they drop below a specified threshold, and (iii) the approximation of forces by fast methods such as the fast multipole method. These are more serious errors because they are larger and because they are systematic rather than random. It is somewhat remarkable then that the same theory also provides a way to excuse these other types of errors. As discussed in Section 5, symplectic integration is still possible, but so is numerical instability!

2. Symplectic integration

We consider time discretization via a one-parameter family of methods that generalizes the endpoint form of leapfrog. The nth step of integration starts with $q^n \approx q(t^n)$, where $t^n := nh$, $p^n \approx p(t^n)$, and $F^n = F(q^n + \alpha h^2 F^n)$ computed from the previous step. Then

$$
\begin{aligned}
&p^{n+1/2} = p^n + \tfrac{1}{2} h F^n,\\
&q^{n+1} = q^n + h M^{-1} p^{n+1/2},\\
&\text{solve} \quad F^{n+1} = F\bigl(q^{n+1} + \alpha h^2 M^{-1} F^{n+1}\bigr) \quad \text{for } F^{n+1},\\
&p^{n+1} = p^{n+1/2} + \tfrac{1}{2} h F^{n+1}.
\end{aligned}
$$

This is equivalent to applying the formula

$$\frac{1}{h^2}(q^{n+1} - 2q^n + q^{n-1}) = F(\alpha q^{n+1} + (1 - 2\alpha) q^n + \alpha q^{n-1})$$

to compute q^{n+1}. For $\alpha \neq 0$, we must solve implicit equations for a modified force. This extra work might be worthwhile for additional accuracy or stability. With $\alpha = \frac{1}{12}$, we have the Cowell–Numerov method, which is effectively fourth order accurate. With $\alpha = \frac{1}{4}$, we have the implicit midpoint method (equivalent to the trapezoid method), which has unrestricted linear stability. And with $\alpha = \frac{1}{2}$, we have the LIM2 method of [25], which has unrestricted nonlinear stability at equilibria [19].

Let us recall what it means for an integrator to be symplectic. If we collect the dependent variables into a vector

$$y := [q_1, q_2, \ldots, q_N, p_1, p_2, \ldots, p_N]^{\mathrm{T}},$$

the Hamiltonian system assumes the simple form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = J H_y(y), \quad \text{where } J := \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}.$$

A transformation from phase space to phase space $\overline{y} = \chi(y)$ is symplectic if its Jacobian matrix is a symplectic matrix:

$$\chi_y^{\mathrm{T}} J \chi_y = J.$$

A symplectic transformation χ preserves volume in phase space (the Liouville property) meaning that if a set of points Ω is mapped to the set $\chi(\Omega)$ then the two sets have equal volume. The h-flow is the mapping $\Phi_h(y(t)) = y(t + h)$ where $y(t)$ is any solution of the given Hamiltonian system. Φ_h is

symplectic. Numerical methods approximate Φ_h by a mapping Ψ_h. A method is *symplectic* if Ψ_h is symplectic.

Following are some reasons for using symplectic integrators for long-time dynamics:

(1) Backward/forward error analysis: the numerical solution is nearly the exact solution of a nearby *Hamiltonian* system on a time interval of length $O(1/h)$. See [2,9,10,15,18,23]. Although $1/h$ is theoretically a long time, in practice, it may be orders of magnitude shorter than the integration interval. Nonetheless, there is no evidence that the "lifespan" of the backward error analysis cannot be extended to much longer times.

(2) Energy conservation as an error indicator: Ge and Marsden [7] show for a Hamiltonian system without any integrals other than energy that the energy cannot be conserved by a symplectic integrator unless the numerical trajectory is equal to the analytical trajectory except for a possible reparameterization of time.

(3) Enhanced nonlinear stability: symplectic integrators may possess enhanced nonlinear stability as a consequence of KAM theory [17].

Symplectic integrators are also important for computing short-time trajectories in the context of hybrid Monte Carlo methods:

(4) Detailed balance: symplecticness implies volume preservation in phase space, and volume preservation and reversibility imply detailed balance [14].

Backward error analysis seems appropriate for an application like molecular dynamics in which the forms of the Hamiltonians are ad hoc and their parameters are fitted to quantum mechanics calculations and experimental observables. It is, of course, also necessary that the desired computables be insensitive to small changes in the energy function.

Generally, the modified Hamiltonian exists only as a nonconvergent asymptotic expansion. As an example, the leapfrog solution for $H(q, p) = T(p) + V(q)$ satisfies $q^n = \widetilde{q}(t^n) + O(h^{2m+2})$ and $p^n = \widetilde{p}(t^n) + O(h^{2m+2})$, where

$$\widetilde{H}(q, p) = H(q, p) + h^2\left(\tfrac{1}{12}T_p^{\mathrm{T}}V_{qq}T_p - \tfrac{1}{24}V_q^{\mathrm{T}}T_{pp}V_q\right) + \cdots + h^{2m}(\cdots). \tag{1}$$

Hence, the theory shows that the leapfrog solution shadows, within $O(h^{2m+2})$, the solution of a nearby ($O(h^2)$) Hamiltonian system. As indicated earlier, the discrepancy $O(h^{2m+2})$ is *not* uniform in time but rather is proved only for an interval of time proportional to $1/h$. The *infinite-time* behavior is generally unknown. Numerical experiments may not be able to answer this question because the infinite-time behavior in finite precision may not, in the infinite-precision limit, be the same as the infinite-time behavior of an infinite-precision calculation. Finite precision could cause a qualitative change in infinite-time behavior. This is, however, an academic question since in practice we must compute with numbers of finite precision. Hence, an open question of great interest is whether a *finite-precision* numerical solution is nearly the exact solution of a nearby Hamiltonian for infinite time.

3. Effect of roundoff errors

The effect of roundoff error was studied in an idealized setting by Scovel [20]. All phase space variables are scaled by a huge number so that they can be expressed as integers without loss of precision, and results are rounded to integer lattice points. A serious practical limitation in both [20] and [5] is the assumption that every updating of a phase space variable is obtained by adding a rounded value of the

exact increment. To achieve this, it would be necessary to use validated arbitrary precision calculation of the increment, either by rigorous significance arithmetic or by interval arithmetic. That this can be very expensive is a consequence of the discontinuous and hence ill-posed nature of a rounding operation—an arbitrary amount of computation is needed in general to resolve whether to round up or down. This has been called the "tablemaker's dilemma", see [13] or [21, p. 8]. Here, we propose a much more practical algorithm for symplectic integration on an equally spaced lattice.

Recall that (floating-point) machine numbers are quantities representable as $(b_1.b_2\ldots b_\nu)_2 \times 2^{\text{integer}}$ where typically the precision $\nu = 24$ or 53. A machine operation $\hat{\circ}$ on two machine numbers is defined to be the result of applying the exact operation $\circ$ to the two numbers and rounding it to nearest machine number.

Assume we know *a priori* that $|q_i| \leqslant \overline{q}$ and $|p_i| \leqslant \overline{p}$, where $\overline{q}$ and $\overline{p}$ are integer powers of 2. Suppose μ is the target precision, where $1 \leqslant \mu \leqslant \nu$, for our fixed-point lattice. In other words, suppose we want to get an equally spaced lattice with spacing $2^{-\mu}\overline{q}$ or $2^{-\mu}\overline{p}$. If $\mu \leqslant \nu - 2$, define for any machine number a,

$$\text{round}_q(a) = \left(a \mathbin{\hat{+}} (0.75 \times 2^{\nu-\mu}\overline{q})\right) \mathbin{\hat{-}} (0.75 \times 2^{\nu-\mu}\overline{q}).$$

If $\mu = \nu - 1$, define

$$\text{round}_q(a) = \left(a \mathbin{\hat{+}} \text{sign}(a)\overline{q}\right) \mathbin{\hat{-}} \text{sign}(a)\overline{q},$$

and, if $\mu = \nu$, define

$$\text{round}_q(a) = \left(a \mathbin{\hat{-}} \text{sign}(a)\overline{q}\right) \mathbin{\hat{+}} \text{sign}(a)\overline{q}.$$

If $|a| > \overline{q}$, an error flag is raised. The round function has the effect of rounding to $\mu - \log_2 \overline{q}$ binary *places* after the binary point (rather than μ significant binary digits), thus creating a uniform lattice in q-space with spacing $\varepsilon := 2^{-\mu}\overline{q}$. The function round_p is defined similarly.

The lattice leapfrog method begins with

$$q^0 = \text{round}_q(q(0)), \qquad p^0 = \text{round}_p\left(M \mathbin{\hat{\times}} \frac{\mathrm{d}}{\mathrm{d}t} q(0)\right),$$

and advances step by step using

$$\begin{aligned} p^{n+1/2} &= p^n + \text{round}_p\left(\tfrac{1}{2}h \mathbin{\hat{\times}} \hat{F}^n\right), \\ q^{n+1} &= q^n + \text{round}_q\left(h \mathbin{\hat{\times}} M \mathbin{\hat{\backslash}} p^{n+1/2}\right), \\ \hat{F}^{n+1} &\approx F\left(q^{n+1} + \alpha h^2 \hat{F}^{n+1}\right), \\ p^{n+1} &= p^{n+1/2} + \text{round}_p\left(\tfrac{1}{2}h \mathbin{\hat{\times}} \hat{F}^{n+1}\right), \end{aligned} \tag{2}$$

where $\hat{\backslash}$ denotes matrix "pre-division". We assume that h is small enough so that arguments of the round function are always in range. We also assume that $\hat{F}^n$ depends on q^n only, allowing us to write $\hat{F}^n = \hat{F}(q^n)$, where $\hat{F}(q)$ denotes the floating-point result for $F(q)$. The method is reversible and hence one-to-one. All points are either part of a limit cycle or go to infinity; there are no transients. If we omitted the special fixed-point rounding and relied on the floating-point rounding that occurs in the addition of the increment to p, then the effective value of the increment would depend on p, and we would no longer have a shear, and hence no longer a symplectic mapping. With the special rounding performed first, the addition of the increment to p is performed without error.

Assume that $\|\widehat{F}(q) - F(q)\|_2 \leqslant c\sqrt{N}\varepsilon$. This is realistic because it is an absolute rather than a relative error bound and hence is tolerant of large relative errors which might result from cancellation in the computation of $\widehat{F}$.

Theorem 1. *Under the assumptions given with algorithm* (2) *above, the following is true*:

1. *The lattice integrator* (2) *is extendible to a symplectic mapping on the continuum in the sense that we can define a symplectic mapping from the continuum to the continuum whose restriction to the lattice is the lattice mapping.*
2. *The solution obtained by the lattice leapfrog method* ($\alpha = 0$) *is the exact numerical solution obtained by the (infinite precision) leapfrog method applied to some different Hamiltonian system with Hamiltonian* $\widehat{T}(p) + \widehat{V}(q)$, *where*

$$|\widehat{V}(q) - V(q)| \leqslant \frac{2}{27}\left(c + \frac{2}{h}\overline{p}\right)\sqrt{N}\overline{q}\varepsilon^2, \tag{3}$$

$$\|\widehat{V}_q(q) - V_q(q)\|_2 \leqslant \left(c + \frac{2}{h}\overline{p}\right)\sqrt{N}\varepsilon, \tag{4}$$

$$\|\widehat{V}_{qq}(q) - V_{qq}(q)\|_2 \leqslant 8\left(c + \frac{2}{h}\overline{p}\right)\sqrt{N}\,\overline{q}^{-1}, \tag{5}$$

$$\left|\widehat{T}(p) - \frac{1}{2}p^{\mathrm{T}}M^{-1}p\right| \leqslant \frac{2}{27}\left(\|M^{-1}\|_2\overline{p} + \frac{1}{h}\overline{q}\right)\sqrt{N}\varepsilon^2. \tag{6}$$

Proof. We analyze the roundoff error by expressing its effects as perturbations to the force and to the velocity. By assumption,

$$\widehat{F}(q) = F(q) + \delta_{\mathrm{F}}, \quad \|\delta_{\mathrm{F}}\|_2 \leqslant c\sqrt{N}\varepsilon.$$

Also

$$M \mathbin{\widehat{\backslash}} p = M^{-1}p + \delta_{\mathrm{G}}, \quad \|\delta_{\mathrm{G}}\|_2 \leqslant 2^{-\nu}\|M^{-1}p\|_2 \leqslant \|M^{-1}\|_2\sqrt{N}\varepsilon\overline{p}.$$

We are assuming

$$-\overline{p} \leqslant \mathrm{round}_p(0.5h \mathbin{\widehat{\times}} \widehat{F}(q^n)) \leqslant \overline{p},$$

from which it follows that

$$\mathrm{round}_p(0.5h \mathbin{\widehat{\times}} \widehat{F}(q^n)) = 0.5h \times \widehat{F}(q^n) + \delta,$$

and

$$\|\delta\|_2 \leqslant (\tfrac{1}{2}\cdot 2^{-\nu}\overline{p} + \tfrac{1}{2}\varepsilon\overline{p})\sqrt{N} \leqslant \varepsilon\overline{p}\sqrt{N}.$$

Hence,

$$p^{n+1/2} = p^n + 0.5h \times (F(q^n) + \delta_{\mathrm{F}}) + \delta = p^n + 0.5h \times (F(q^n) + \Delta_{\mathrm{F}}(q^n)),$$

where $\Delta_{\mathrm{F}}(q^n) := \delta_{\mathrm{F}} + (2/h)\delta$ satisfies

$$\|\Delta_{\mathrm{F}}(q^n)\|_2 \leqslant \left(c + \frac{2}{h}\overline{p}\right)\sqrt{N}\varepsilon.$$

Similarly,

$$q^{n+1} = q^n + h\bigl(M^{-1}p^{n+1/2} + \Delta_{\mathrm{G}}(p^{n+1/2})\bigr),$$

where

$$\|\Delta_{\mathrm{G}}(p^{n+1/2})\|_2 \leqslant \left(\|M^{-1}\|_2\overline{p} + \frac{1}{h}\overline{q}\right)\sqrt{N}\varepsilon.$$

But this is not enough: we must show that these perturbations can be expressed as a gradient. Let

$$\phi(x) = \begin{cases} (1 - \|x\|_2)^2, & \|x\|_2 \leqslant 1, \\ 0, & \|x\|_2 \geqslant 1. \end{cases}$$

For future reference, note that

$$\phi_x(x) = 2x - 2\|x\|_2^{-1}x, \tag{7}$$

$$\phi_{xx}(x) = 2(1 - \|x\|_2^{-1})I + 2\|x\|_2^{-3}xx^{\mathrm{T}}, \tag{8}$$

for $0 < \|x\|_2 < 1$. Let

$$\phi_i(q) = \phi\left(\frac{2}{\varepsilon\overline{q}}(q - Q_i)\right),$$

where the Q_i are machine points in q-space, and define

$$\widehat{V}(q) = V(q) + \sum_i \phi_i(q)(q - Q_i)^{\mathrm{T}}\Delta_{\mathrm{F}}(Q_i). \tag{9}$$

At any given point q, only one term in the sum (9) can be nonzero, say the ith term. For convenience, write $\Delta_{\mathrm{F}}(Q_i)$ as Δ and $2(\varepsilon\overline{q})^{-1}(q - Q_i)$ as x, then

$$\widehat{V} - V = \frac{\varepsilon\overline{q}}{2}\phi(x)x^{\mathrm{T}}\Delta.$$

For future reference, note that

$$\widehat{V}_q - V_q = \phi_x x^{\mathrm{T}}\Delta + \phi\Delta,$$

and

$$\widehat{V}_{qq} - V_{qq} = \frac{2}{\varepsilon\overline{q}}(\phi_{xx}x^{\mathrm{T}}\Delta + \phi_x\Delta^{\mathrm{T}} + \Delta\phi_x^{\mathrm{T}}).$$

Because $\|x\|_2 \leqslant 1$, we have

$$|\phi(x)x^{\mathrm{T}}\Delta| \leqslant (1 - \|x\|_2)^2\|x\|_2\|\Delta\|_2 \leqslant \tfrac{4}{27}\|\Delta\|_2,$$

proving (3). Eq. (6) is proved similarly. Note that

$$\phi_i(Q_j) = \begin{cases} 1, & j = i, \\ 0, & j \neq i, \end{cases}$$

and $\nabla_q\phi_i(Q_j) = 0$, $j \neq i$. If we evaluate the gradient $\widehat{V}_q$ at a lattice point Q_i, it yields the desired result, viz.,

$$\widehat{V}(Q_j) = V(Q_j) + \Delta_{\mathrm{F}}(Q_j).$$

Eq. (4) follows from

$$\begin{aligned}\|\widehat{V}_q - V_q\|_2 &\leqslant (\|\phi_x\|_2\|x\|_2 + |\phi|)\|\Delta\|_2\\ &= (2(1-\|x\|_2)\|x\|_2 + (1-\|x\|_2)^2)\|\Delta\|_2\\ &= (1-\|x\|_2^2)\|\Delta\|_2 \leqslant \|\Delta\|_2.\end{aligned}$$

It remains to verify Eq. (5). Letting $u = \|x\|_2^{-1}x$ and $\xi = \|x\|_2$,

$$\widehat{V}_{qq} - V_{qq} = \frac{2}{\varepsilon\overline{q}}(2(\xi-1)u^{\mathrm{T}}\Delta I + 2u^{\mathrm{T}}\Delta u u^{\mathrm{T}} + 2(\xi-1)(u\Delta^{\mathrm{T}} + \Delta u^{\mathrm{T}})).$$

With $\Delta = \|\Delta\|_2 v$ and $u^{\mathrm{T}}v = \gamma$, we have

$$\begin{aligned}\|\widehat{V}_{qq} - V_{qq}\| &\leqslant \frac{4}{\varepsilon\overline{q}}\|(1-\xi)(\gamma I - \gamma uu^{\mathrm{T}} + uv^{\mathrm{T}} + vu^{\mathrm{T}}) - \xi\gamma uu^{\mathrm{T}}\|_2\|\Delta\|_2\\ &\leqslant \frac{4}{\varepsilon\overline{q}}((1-\xi)\|A\|_2 + \xi|\gamma|)\|\Delta\|_2,\end{aligned}$$

where $A = \gamma I - \gamma uu^{\mathrm{T}} + uv^{\mathrm{T}} + vu^{\mathrm{T}}$. Let $x = \alpha u + \beta v + w$, where $w \perp u$, $w \perp v$. We have

$$Ax = \alpha(\gamma u + v) + \beta((1-\gamma^2)u + 2\gamma v) + \gamma w,$$

and

$$\begin{aligned}\|Ax\|_2^2 &= \alpha^2(1+3\gamma^2) + 8\alpha\beta\gamma + \beta^2(1+6\gamma^2-3\gamma^4) + \gamma^2 w^{\mathrm{T}}w\\ &\leqslant 4(\alpha^2 + 2\alpha\beta\gamma + \beta^2 + w^{\mathrm{T}}w) = 4\|x\|_2^2.\end{aligned}$$

Hence $\|A\|_2 \leqslant 2$ and

$$(1-\xi)\|A\|_2 + \xi|\gamma| \leqslant (1-\xi)2 + \xi|\gamma| \leqslant 2. \qquad \square$$

In N-body simulations, it is common to store velocities rather than momenta. This can be accommodated by imagining that we have a momentum lattice which is a scaling by M of the velocity lattice.

Scovel [20] gives an alternative algorithm for the implicit midpoint method ($\alpha = \frac{1}{4}$) but does not prove the existence of a fixed point for his iteration on the lattice. There seems to be a possibility of cycling. Scovel also gives results for higher order explicit methods showing that the map is symplectic. However, the proof does not construct a common $\widehat{V}$ that is the same from one stage of the method to the next.

Suppose we apply the backward error analysis described in the previous section to $\widehat{T}(p) + \widehat{V}(q)$. For a linear problem, the perturbed problem will be nonlinear, so there is no exact nearby Hamiltonian. More generally, because $\widehat{V}_{qq} = V_{qq} + \mathrm{O}(h^{-1})$, the nearby Hamiltonian (1) is made unrecognizable by roundoff error. Third and higher derivatives of our constructed Hamiltonian are proportional to negative powers of ε. Hence, Scovel [20] makes the paradoxical suggestion that reducing precision might improve the quality of the dynamics.

4. Numerical experiments

Computer experiments on infinite-time lattice dynamics is feasible for values of the relative lattice spacing ε which are not too small.

Problem 1. A compressed vibrating beam can be modeled with $H(q, p) = \frac{1}{2}p^2 + \frac{1}{2}q^2(q^2 - 1)$. The compression makes the beam prefer to bend one way or the other. Orbits for $\varepsilon = 2^{-13}, 2^{-14}, 2^{-15}, 2^{-16}$ are shown in Figs. 2–5. The stepsize is $h = 0.2$. We observe that increasing precision does not reduce the thickness of the orbit.

Problem 2. The Hénon–Heiles [11] Hamiltonian

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

exhibits nonchaotic behavior for energies lower than $\frac{1}{8}$, and it is believed to possess a second integral. Similar to [4], we choose initial values $(q_1, q_2, p_1, p_2) = (0.12, 0.12, 0.12, 0.12)$, giving an energy of 0.029952. The solution is computed with time step $\frac{1}{6}$. Shown in Figs. 6–9 for $\varepsilon = 2^{-8}, 2^{-9}, 2^{-10}, 2^{-11}$ are points of intersection of the orbit with the plane $q_1 = 0$ that satisfy $p_1 > 0$, projected onto the (q_2, p_2)-plane (Poincaré sections). Whenever $q_1^n < 0$ and $q_1^{n+1} \geqslant 0$, we plot (q_2^{n+1}, p_2^{n+1}). We observe that the apparent structure of the orbit is not resolved until a precision of $\varepsilon = 2^{-11}$ is reached.

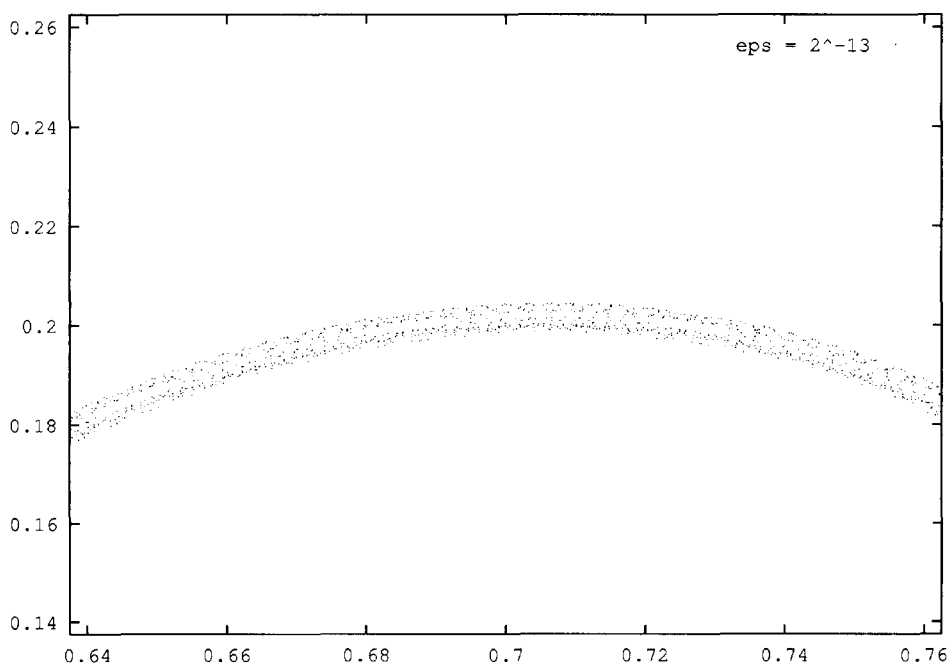

Fig. 2. Piece of infinite orbit for beam displacement with $\varepsilon = 2^{-13}$.

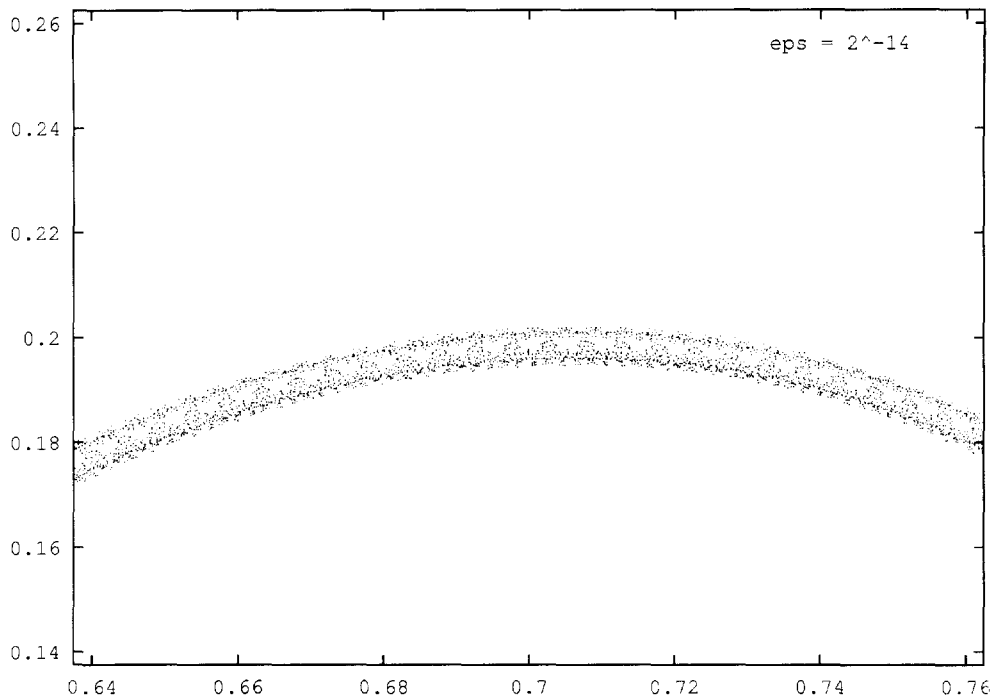

Fig. 3. Piece of infinite orbit for beam displacement with $\varepsilon = 2^{-14}$.

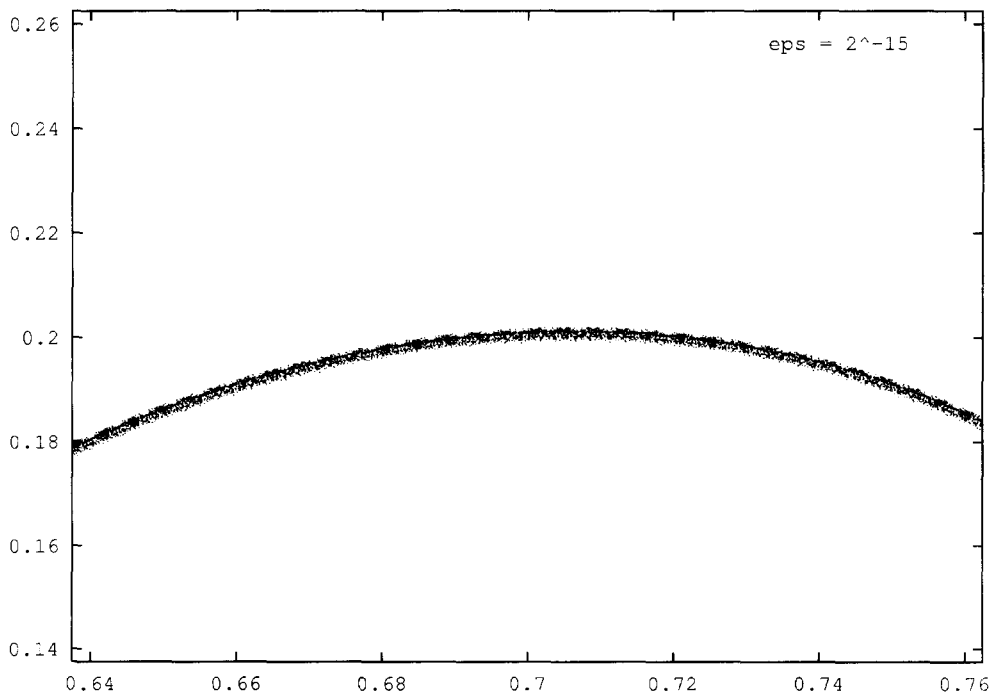

Fig. 4. Piece of infinite orbit for beam displacement with $\varepsilon = 2^{-15}$.

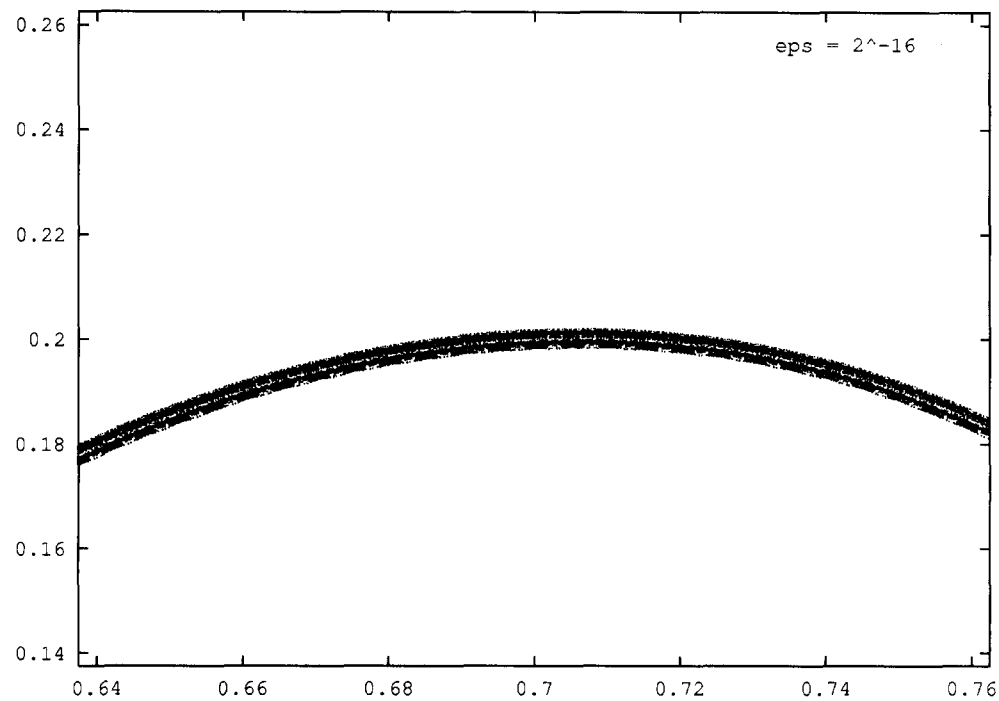

Fig. 5. Piece of infinite orbit for beam displacement with $\varepsilon = 2^{-16}$.

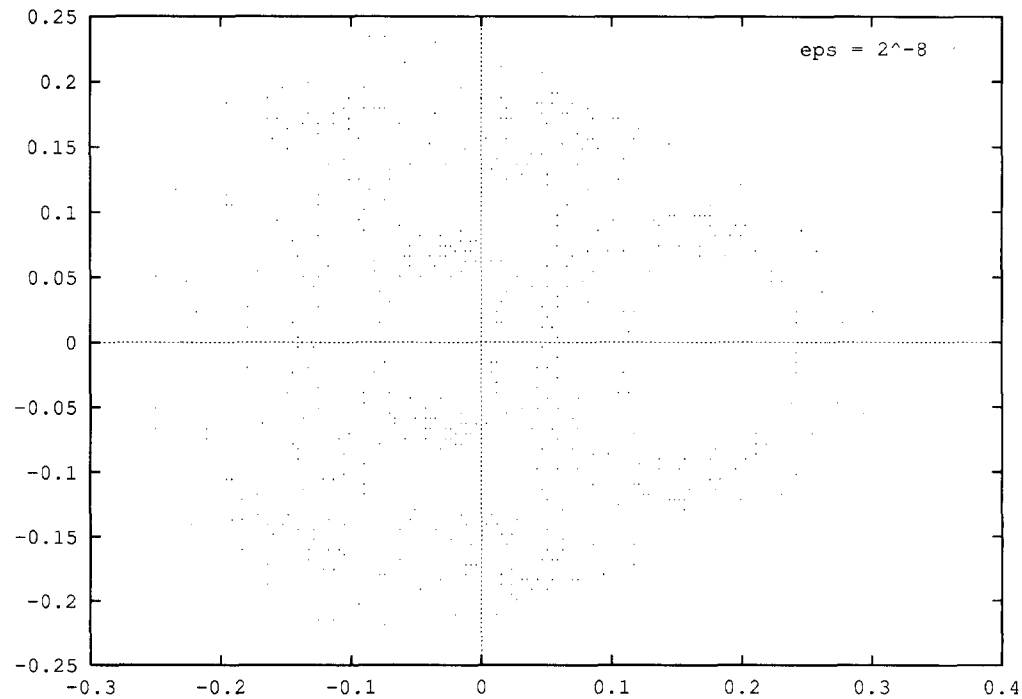

Fig. 6. Poincaré section of infinite orbit for Hénon–Heiles Hamiltonian with $\varepsilon = 2^{-8}$.

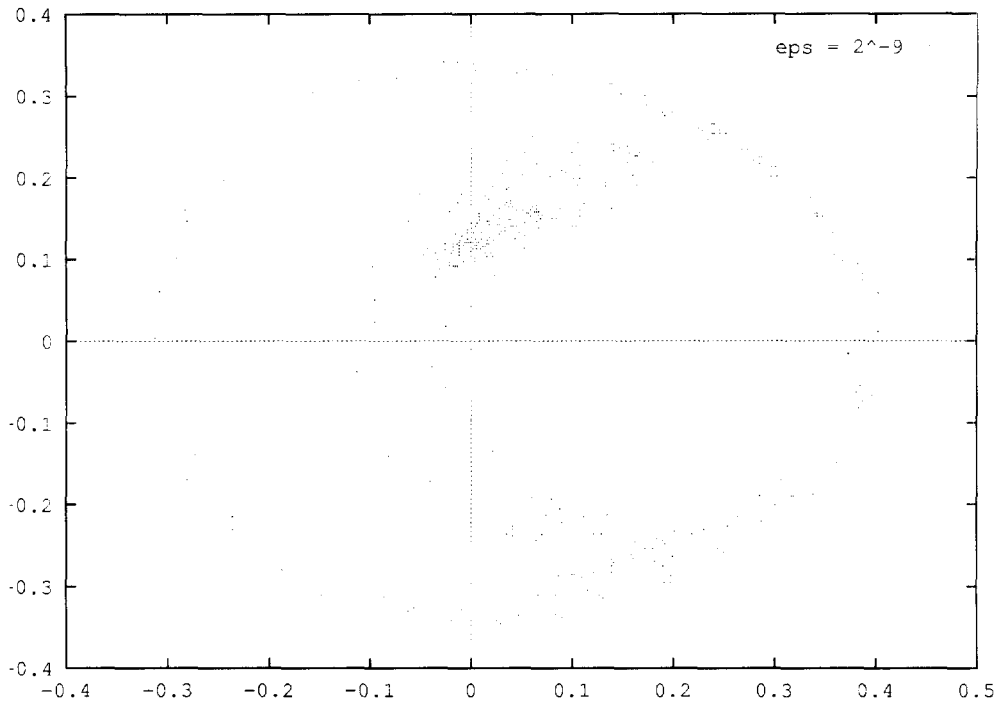

Fig. 7. Poincaré section of infinite orbit for Hénon–Heiles Hamiltonian with $\varepsilon = 2^{-9}$.

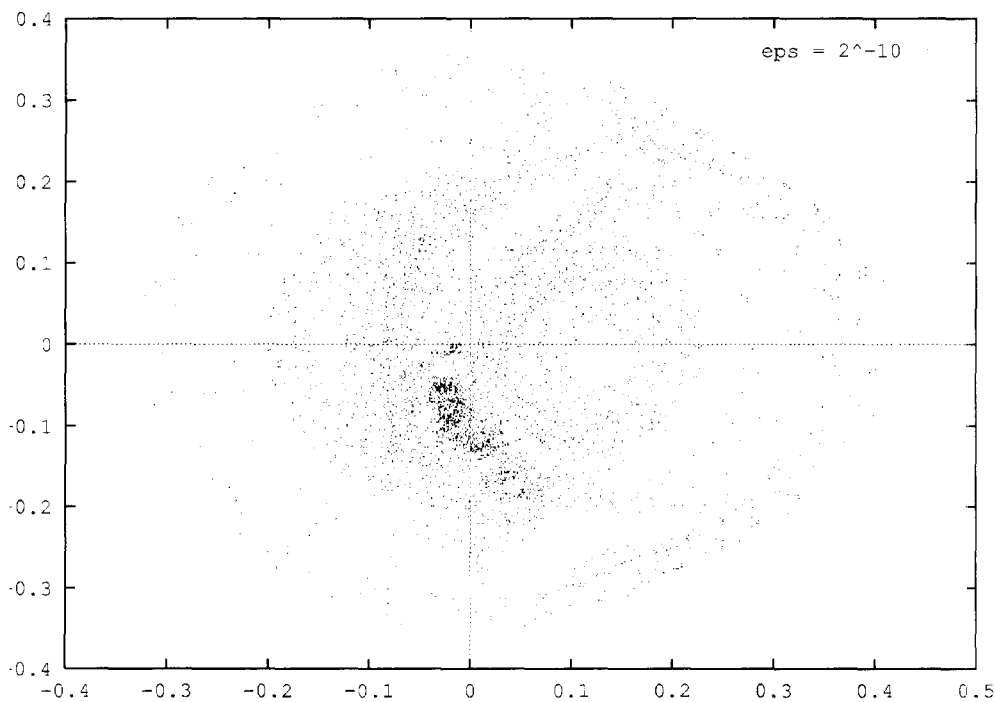

Fig. 8. Poincaré section of infinite orbit for Hénon–Heiles Hamiltonian with $\varepsilon = 2^{-10}$.

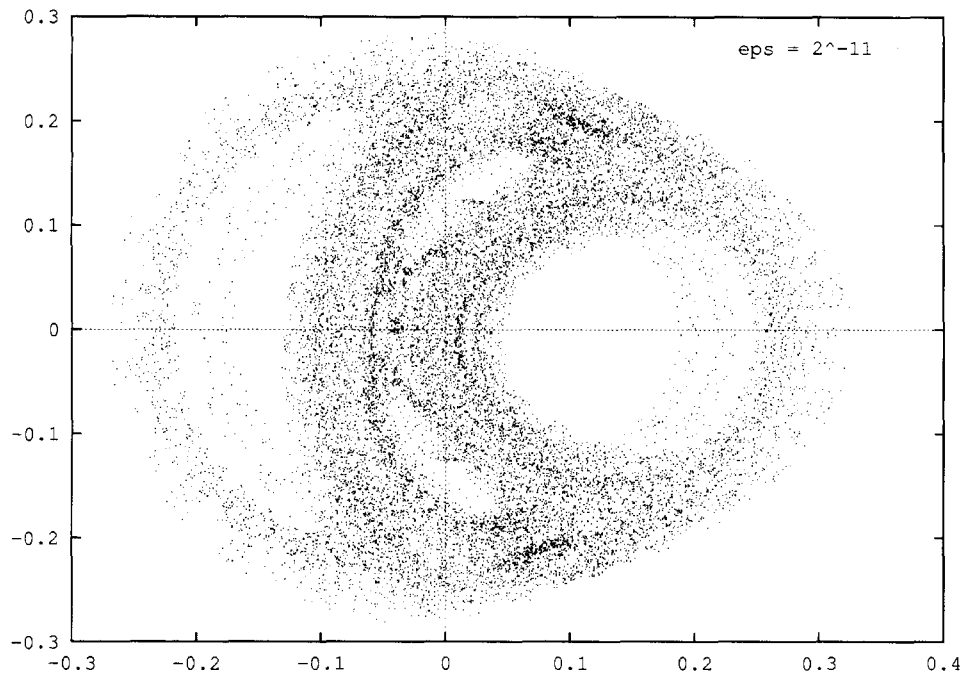

Fig. 9. Poincaré section of infinite orbit for Hénon–Heiles Hamiltonian with $\varepsilon = 2^{-11}$.

5. Effect of other errors

Recall the assumptions

$$\widehat{F}^n = \widehat{F}(q^n) \quad \text{and} \quad \left|\widehat{F}(q) - F(q)\right| \leqslant c\varepsilon.$$

It is important for symplecticness that $\widehat{F}^n$ depend only on q^n and not on p^n nor on q^{n-1}. However, the error in $\widehat{F}^n$ need not be due to roundoff; it could be due to any of the following:

- use of cutoffs, such as for (nonbonded) Lennard–Jones interactions,
- use of fast N-body solvers for inverse-square-law forces,
- incomplete iterations. If we formulate the implicit method as

  $$F^n = F\left(q^n + \alpha h^2 F^n\right),$$

  it is symplectic as long as we do not use outside information to solve it. If we use history to help us solve it, then it will not be symplectic unless we iterate to some limit defined independently of the iteration process.

Errors in the force greater than those due to roundoff alone may have implications for stability. Restrictions on h to ensure stability depend on the Jacobian matrix of the force vector. Linear stability for $\alpha = 0$ requires

$$\rho\left(h^2 \widehat{V}_{qq}\right) \leqslant 4.$$

Using error bound (5), we have

$$\rho\left(h^2 \widehat{V}_{qq}\right) \leqslant \rho\left(h^2 V_{qq}\right) + 8\left(ch^2 + 2h\overline{p}\right)\sqrt{N}\,\overline{q}^{-1}.$$

There is a possibility of instability if c is large, or, more specifically, if

$$\frac{|\widehat{F}(q) - F(q)|}{\varepsilon} \gg 1.$$

That this can happen in practice is suggested by the energy growth observed in molecular dynamics if too few terms are used in the multipole expansion of the fast multipole method [3].

We can make the constant c smaller by reducing the precision μ. In other words, it is suggested that increasing ε might thwart instability!

Acknowledgements

The author is grateful to George Thiruvathukal for Fig. 1, to Amy Ryan for Figs. 2–5, and to Tony Surma for Figs. 6–9.

References

[1] M.P. Allen and D.J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987). Reprinted in paperback in 1989 with corrections.
[2] G. Benettin and A. Giorgilli, On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms, *J. Statist. Phys.* 74 (1994) 1117–1143.
[3] T. Bishop, R.D. Skeel and K. Schulten, Difficulties with multiple timestepping and the fast multipole algorithm in molecular dynamics, *J. Comput. Chem.* 18 (1997) 1785–1791.
[4] P.J. Channell and J.C. Scovel, Symplectic integration of Hamiltonian systems, *Nonlinearity* 3 (1990) 231–259.
[5] D.J. Earn and S. Tremaine, Exact numerical studies of Hamiltonian maps: iterating without roundoff error, *Physica D* 56 (1992) 1–22.
[6] D.J.D. Earn, Symplectic integration without roundoff error, in: V.G. Gurzadyan and D. Pfenniger, eds., *Ergodic Concepts in Stellar Dynamics* (Springer, Heidelberg, 1994) 122–130; also: Weizmann Institute of Science, August 1993, preprint.
[7] Z. Ge and J.E. Marsden, Lie–Poisson Hamilton–Jacobi theory and Lie–Poisson integrators, *Phys. Lett. A* 133 (1988) 134–139.
[8] H. Grubmüller and P. Tavan, Efficient algorithms for molecular dynamics simulations of proteins: How good are they?, Manuscript, 1994.
[9] E. Hairer, Backward error analysis of numerical integrators and symplectic methods, *Ann. Numer. Math.* 1 (1994) 107–132.
[10] E. Hairer and C. Lubich, The life-span of backward error analysis for numerical integrators, *Numer. Math.* 74 (1997) 441–462.
[11] M. Hénon and C. Heiles, The applicability of the third integral of motion: some numerical experiments, *Astron. J.* 69 (1964) 73–79.
[12] C.F.F. Karney, Long-time correlations in the stochastic regime, *Physica D* 8 (1983) 360–380.
[13] D.H. Lehmer, *MTAC* 1 (1943) 31. An example from this article was given by W. Kahan in a lecture entitled "The tablemaker's dilemma and other quandaries" at the Second Conference on Mathematical Software at Purdue University, May 29–31, 1974.
[14] B. Mehlig, D.W. Heermann and B.M. Forrest, Hybrid Monte Carlo method for condensed-matter systems, *Phys. Rev. B* 45 (1992) 679–685.
[15] S. Reich, Backward error analysis for numerical integrators, Technical Report, 1997.

[16] J. Sanz-Serna and M. Calvo, *Numerical Hamiltonian Problems* (Chapman and Hall, London, 1994).
[17] J.M. Sanz-Serna, Two topics in nonlinear stability, in: W. Light, ed., *Advances in Numerical Analysis* (Clarendon Press, Oxford, 1991) 147–174.
[18] J.M. Sanz-Serna, Symplectic integrators for Hamiltonian problem: an overview, *Acta Numer.* 1 (1992) 243–286.
[19] T. Schlick, M. Mandziuk, R.D. Skeel and K. Srinivas, Nonlinear resonance artifacts in molecular dynamics simulations, *J. Comput. Phys.* 139 (1998) 1–29.
[20] C. Scovel, On symplectic lattice maps, *Phys. Lett. A* 159 (1991) 396–400.
[21] R.D. Skeel and J.B. Keiper, *Elementary Numerical Computing with Mathematica* (McGraw-Hill, New York, 1993).
[22] R.D. Skeel, G. Zhang and T. Schlick, A family of symplectic integrators: stability, accuracy, and molecular dynamics applications, *SIAM J. Sci. Comput.* 18 (1997) 203–222.
[23] D.M. Stoffer, Some geometric and numerical methods for perturbed integrable systems, Ph.D. Thesis, Swiss Federal Institute of Technology, Zürich (1988).
[24] M. Suzuki, Improved Trotter-like formula, *Phys. Lett. A* 180 (1993).
[25] G. Zhang and T. Schlick, Implicit discretization schemes for Langevin dynamics, *Mol. Phys.* 84 (1995) 1077–1098.