
Article:
Niesen, J and Moan, PC (2014) On an asymptotic method for computing the modified energy for symplectic methods. Discrete and Continuous Dynamical Systems - Series A, 34 (3). 1105 - 1120. ISSN: 1078-0947

https://doi.org/10.3934/dcds.2014.34.1105

Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

doi:10.3934/xx.xx.xx.xx

pp. **X–XX**

ON AN ASYMPTOTIC METHOD FOR COMPUTING THE MODIFIED ENERGY FOR SYMPLECTIC METHODS

Per Christian Moan

Centre of Mathematics for Applications
University of Oslo
Norway

Jitse Niesen

School of Mathematics
University of Leeds
United Kingdom

Abstract. We revisit an algorithm by Skeel *et al.* [5, 16] for computing the modified, or shadow, energy associated with symplectic discretizations of Hamiltonian systems. We amend the algorithm to use Richardson extrapolation in order to obtain arbitrarily high order of accuracy. Error estimates show that the new method captures the exponentially small drift associated with such discretizations. Several numerical examples illustrate the theory.

1. **Introduction.** Numerical simulation of conservative differential equations requires special care in order to avoid introducing non-conservative, or non-physical truncation error effects. For Hamiltonian ODEs or Euler–Lagrange equations originating from variational principles there exists much evidence [6, 9, 11, 14] that the proper discretization scheme should be *symplectic* [7, 9, 11, 18]. In the Hamiltonian case this can be achieved by imposing special conditions on classical methods or by methods based on generating functions [7]. In the variational formulation symplecticity is achieved by discretizing the action integral and carrying out a discrete variation [10]. In some cases these formulations and methods turn out to be equivalent by the Legendre transformation [8, 10].

Focusing on the Hamiltonian side, symplecticity implies that the trajectory produced by the numerical algorithm *is the exact solution* [12] of another, non-autonomous "modified" Hamiltonian system close to the original one. Various stability results for Hamiltonian ODEs then apply, leading to an understanding of the dynamics of such discretizations schemes [6, 9, 11, 15]. Early results on modified equations focused on the autonomous part [2, 4, 6, 13, 14] and established that its flow is exponentially close to the numerical trajectory. This work was motivated by the bounded error in energy observed in simulations with symplectic schemes. The early results are contained in the newer results since the time-dependent part is exponentially small due to analyticity. Despite its smallness the non-autonomous term excites instabilities through resonances, one consequence being a drift in the modified energy. In simulations requiring millions of steps such as in molecular

2010 *Mathematics Subject Classification.* Primary: 65P10; Secondary: 37J40, 37M15.

Key words and phrases. modified energy, symplectic integration, Hamiltonian systems, Richardson extrapolation.

dynamics [9, 17] and celestial mechanics [19] these effects become significant and it becomes important to understand and control them.

Constructing the modified Hamiltonian is equivalent to evaluating many terms in the Baker–Campbell–Hausdorff formula, or its continuous analogue [13], a combinatorially complicated task possible only for small systems and to a low order of accuracy. Recently, Skeel and coworkers [5, 16] devised a method for numerically computing the value of the modified Hamiltonian along the numerical trajectory, thus allowing us to track the possible drift in the modified energy. In this paper we modify their method to make it easier to implement and obtain high order, possibly at the cost of extra storage. The new method is based on the same idea, but it does not give identical results. We also provide exponentially small error bounds when the new method is applied in the asymptotic regime. It is then used to verify, and justify the theory of modified equation on several test equations and methods.

2. Modified equations.

As alluded to in the introduction, the numerical solution of an ODE $y' = f(y)$ is interpolated by the exact solution of a modified ODE $\overline{y}' = f(\overline{y}, t)$. The modified equation is non-autonomous, but the non-autonomous part is exponentially small in the step size. More precisely, given an analytic vector field f and a one-step method defined by an analytic mapping $\Psi_{h,f}$, there exists an analytic vector field $\overline{f}(y,t)$, h-periodic and analytic in t, whose exact flow exactly interpolates the numerical trajectory $\{x_n\}$, $x_{n+1} = \Psi_{h,f}(x_n)$. The construction in [12] starts by constructing a modified vector field $\tilde{f}(y,t)$ which is only C^∞ in t whose flow interpolates $\{x_n\}$. This vector field is then transformed by a time-dependent coordinate transformation into a vector field $\overline{f}$ analytic in t.

The domain of analyticity of $\tilde{f}$ plays an important role in the analysis, and we have found it useful to assume that $\tilde{f}$ is analytic for all y in a domain of the form

$$\mathcal{D}_y := \bigcup_{t>0} \{z \in \mathbb{C}^d : |\tilde{y}(t) - z|_\infty < \tilde{r}_y\} = \bigcup_{t>0} \{z \in \mathbb{C}^d : |\Im(\tilde{y}(t) - z)|_\infty < \tilde{r}_y\}$$

for some $\tilde{r}_y > 0$, where $\tilde{y}(t) = \phi_{t,\tilde{f}}(y_0)$ is the trajectory of the smooth modified vector field $\tilde{f}$. This domain is typically smaller than the domain of analyticity of f, and depends on the numerical method. In the following we will use the sup-norm $\|f\|_{\mathcal{D}} = \sup_{z \in \mathcal{D}_y} |f(z)|_\infty$. With these definitions the main result of [12] in the limit $h \to 0$ can be formulated as

Theorem 1. *Let $\Psi_{h,f}$ be a one-step method applied to the analytic vector field f, and $y_{n+1} = \Psi_{h,f}(y_n)$ be the approximations obtained by iterating Ψ. Then there exists a modified vector field $\overline{f}(\overline{y}, t) = f(\overline{y}) + r_1(\overline{y}) + r_2(\overline{y}, t)$ which is h-periodic in t and analytic in $(y,t) \in \overline{\mathcal{D}}'$ such that its exact flow satisfies $\overline{y}(nh) = \Phi_{nh,\overline{f}}(y_0) = x_n$. In the limit $h \to 0$ we have the estimates*

$$\|\overline{f}\|_{\overline{\mathcal{D}}'} \leq \frac{2}{1-\eta} \|f\|_{\mathcal{D}}$$

$$\|r_2\|_{\overline{\mathcal{D}}'} = \mathcal{O}\left(\frac{\|f\|_{\mathcal{D}}}{h} \exp\left(-\eta \frac{2\pi\delta}{\|f\|_{\mathcal{D}} e h} \right) \right)$$

for $0 < \eta < 1$, $0 < \delta < \tilde{r}_y$. The domain of analyticity of the modified vector field is

$$\overline{\mathcal{D}}' = \left\{ (z,\tau) \in \mathbb{C}^d \times \mathbb{C} : |\Im(z - \overline{y}(t))|_\infty < \tilde{r}_y - \delta,\ |\Im(\tau - t)| < \frac{\eta\delta}{\|f\|_{\mathcal{D}} e} \right\},$$

and the norm $\| \cdot \|_{\mathcal{D}}$ is defined by $\|\overline{f}\|_{\overline{\mathcal{D}}'} = \sup_{(z,\tau) \in \overline{\mathcal{D}}'} |\overline{f}(z,\tau)|_\infty$.

For a Hamiltonian vector field f and a symplectic numerical method [6, 7, 14], the modified vector field $\overline{f}$ is also Hamiltonian [4, 7, 9], with Hamiltonian $\overline{H} = H(y) + G_1(y) + G_2(y,t)$ where H, G_1 and G_2 are the Hamiltonians corresponding to the vector fields f, r_1 and r_2, respectively. The change in the *modified energy* along the numerical trajectory therefore satisfies

$$\frac{d}{dt}\overline{H} = \{\overline{H}, \overline{H}\} + \frac{\partial}{\partial t}\overline{H} = \frac{\partial}{\partial t}G_2,$$

where $\{F, G\} := \sum_{j=1}^{m} F_{q_j} G_{p_j} - F_{p_j} G_{q_j}$ is the Poisson bracket. By Theorem 1 this drift is very small for small h, thus motivating symplectic methods.

3. **The method of Skeel *et al.* for computing the modified energy.** Skeel and coworkers [5, 16] found an ingenious way of evaluating the modified energy $\overline{H}$ at the points $\{x_n\}$ for discretizations based on splittings [3]. Suppose we have an Hamiltonian given by $H = \frac{1}{2}p^T M^{-1} p + U(q)$. An explicit splitting algorithm with step size h is given by

$$\begin{aligned}
&\text{for } n = 0, 1, 2, ... \\
&\quad \hat{p}_0 = p_n, \quad \hat{q}_0 = q_n \\
&\quad \text{for } s = 1 : S \\
&\qquad \hat{p}_s = \hat{p}_{s-1} - h a_s U_q(\hat{q}_{s-1}) \\
&\qquad \hat{q}_s = \hat{q}_{s-1} + h b_s M^{-1} \hat{p}_s \\
&\quad \text{end} \\
&\quad p_{n+1} = \hat{p}_S, \quad q_{n+1} = \hat{q}_S \\
&\text{end}
\end{aligned} \tag{1}$$

leading to approximations $p_{n+1} = \hat{p}_S$, $q_{n+1} = \hat{q}_S$ when $\hat{q}_0 = q_n$, $\hat{p}_0 = p_n$ at $t_n = nh$. By choosing the coefficients a_s, b_s appropriately, a method of arbitrary high order can be found. The modified Hamiltonian can be found by representing the inner loop of (1) as a concatenation of exponential operators [7]

$$\begin{aligned}
\Psi_{h,f}(p,q) = \exp(-h a_1 U_q \partial_p)(p,q) \exp(h b_1 M^{-1} p \partial_q) \cdots \\
\exp(-h a_S U_q \partial_p) \exp(h b_S M^{-1} p \partial_q),
\end{aligned}$$

whereby the Baker–Campbell–Hausdorff (BCH) formula is used to find an expression so that $\Psi_{h,f}(x) \simeq \exp(h\overline{f}\partial)(x)$.

The approach of Skeel *et al.* for computing values of the modified energy is to append one scalar equation to the numerical integrator,

$$\begin{aligned}
&\text{for } n = 0, 1, 2, \dots \\
&\quad \hat{p}_0 = p_n, \quad \hat{q}_0 = q_n, \quad \hat{\beta}_0 = \beta_n \\
&\quad \text{for } s = 1 : S \\
&\qquad \hat{p}_s = \hat{p}_{s-1} - h a_s U_q(\hat{q}_{s-1}) \\
&\qquad \hat{\beta}_s = \hat{\beta}_{s-1} - h a_s (\hat{q}_{s-1}^T U_q(\hat{q}_{s-1}) + 2U(\hat{q}_{s-1})) \\
&\qquad \hat{q}_s = \hat{q}_{s-1} + h b_s M^{-1} \hat{p}_s \\
&\quad \text{end} \\
&\quad p_{n+1} = \hat{p}_S, \quad q_{n+1} = \hat{q}_S, \quad \beta_{n+1} = \hat{\beta}_S \\
&\text{end}
\end{aligned} \tag{2}$$

where $\beta_0 = 0$. To understand how the modified energy can be recovered from $\{p_n, q_n, \beta_n\}$, note that by Theorem 1 the numerical trajectory (p_n, q_n) is exactly interpolated by the flow of a Hamiltonian $\overline{H}$. The discretization (2) is the discretization of $H_\alpha = \alpha^2 H(\alpha^{-1}p, \alpha^{-1}q)$ where β is conjugate to α, the so-called *homogeneous extension* of $H(p,q)$. The discovery in [16] rests on the fact that homogeneous extension is a Lie algebra homeomorphism. Thus, since $\overline{H}$ is constructed by Poisson brackets as in the BCH formula, the modified Hamiltonian for (2), $\overline{H}_\alpha$, is the homogeneous extension of $\overline{H}$, hence the trajectory generated by (2) is interpolated by

$$\begin{aligned}
\overline{q}' &= \overline{H}_{\overline{p}}(\overline{p}, \overline{q}, t), \\
\overline{p}' &= -\overline{H}_{\overline{q}}(\overline{p}, \overline{q}, t), \\
\overline{\beta}' &= \overline{q}^T \overline{H}_{\overline{q}}(\overline{p}, \overline{q}, t) + \overline{p}^T \overline{H}_{\overline{p}}(\overline{p}, \overline{q}, t) - 2\overline{H}(\overline{p}, \overline{q}, t),
\end{aligned}$$

from which

$$\overline{H} = \frac{1}{2}(\overline{q}^T \overline{H}_{\overline{q}} + \overline{p}^T \overline{H}_{\overline{p}} - \overline{\beta}') = \frac{1}{2}(-\overline{q}^T \overline{p}' + \overline{p}^T \overline{q}' - \overline{\beta}'). \tag{3}$$

The equation for α is removed from (3) since $\overline{H}_\alpha$ does not depend on β, the conjugate variable of α, and hence $\alpha' = 0$ which is solved exactly by the methods we are considering.

Thus the value of $\overline{H}$ can be computed by finding the derivatives of the interpolating trajectory (which are not known since we do not have $\overline{H}_p, \overline{H}_q$). In [5, 16] estimates of the derivatives are computed using backward difference formulas and interpolating polynomials with stored values of $\{p_n, q_n, \beta_n\}$. These polynomials can be precomputed, but unfortunately the required expressions are very large, and they only provide expressions up to order 24. Their method does however have an advantage in requiring less stored values than one based on centered differences, which might be important if the modified energy is part of the simulation [5].

4. **Richardson extrapolation.**

Our suggestion is to use Richardson extrapolation in order to avoid the large expressions that arise in the method described in the previous section.

First consider the use of Richardson extrapolation to find the derivative of a function, say y, at zero given the function values on a grid. We define the central difference approximations

$$T_{j,1} = \frac{y(jh) - y(-jh)}{2jh}, \quad j = 1, \ldots,$$

and compute the Richardson table entries

$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(1 - k/j)^2 - 1}, \quad k = 1, \ldots, j-1.$$

We then have by standard results that $T_{j,j} = y'(0) + \mathcal{O}(h^{2j})$. In fact, it is straightforward to prove by induction that the $T_{j,k}$ satisfy

$$T_{j,k} = \sum_{i=1}^{k} \frac{2\,(-1)^{i+1}\,(j)_k^2}{(i-1)!\,(k-i)!\,(2j-k+i)_k\,(j-k+i)} T_{j-k+i,1}$$

where the Pochhammer symbol denotes the falling factorial:

$$(n)_k = n(n-1)(n-2)\ldots(n-k+1) = \frac{n!}{k!}.$$

It follows that the diagonal entries in the Richardson table are given by $T_{m,m} = D_m y(0)$ where $D_m(y)$ denotes the central-difference approximation to the derivative $y'(0)$ using $2m$ points, defined by

$$D_m y(0) = \sum_{j=1}^{m} \frac{(-1)^{j+1}(m!)^2}{jh(m-j)!(m+j)!}\big(y(jh) - [y(-jh)\big). \tag{4}$$

This approximation satisfies $D_m y(0) = y'(0) + \mathcal{O}(h^{2m})$. By choosing the index m appropriately, it is possible to find an exponentially accurate approximation for the derivative of analytic functions, as stated in the following lemma.

Lemma 2. *Let $y(t)$ be analytic in $\{t \in \mathbb{C} : |\Im(t)| < \rho\}$, then there exists an m^* (which depends on h) and a constant $C_1 > 0$ such that*

$$|y'(0) - D_{m^*} y(0)| \le C_1 \frac{\rho \exp\left(-\frac{\pi\rho}{h}\right)}{h^2} \|y\|_\rho$$

where $\|y\|_\rho = \sup_{|\Im(\tau)|<\rho} |y(\tau)|_\infty$.

The proof of this result and other results are found in the Appendix.

5. **Computing the modified energy using Richardson extrapolation.** Returning to the computation of the modified energy (3), the derivatives in this formula can be computed with Richardson extrapolation using the stored values of $\{p_n, q_n, \beta_n\}$. To compute the modified energy at $t = nh$, we define the central difference approximation

$$\begin{aligned} T^n_{j,1} &= \frac{1}{2}\left(-\overline{q}^T(nh)\,\frac{\overline{p}((n+j)h) - \overline{p}((n-j)h)}{2jh}\right. \\ &\qquad \left. + \overline{p}^T(nh)\,\frac{\overline{q}((n+j)h) - \overline{q}((n-j)h)}{2jh} - \frac{\overline{\beta}((n+j)h) - \overline{\beta}((n-j)h)}{2jh}\right) \\ &= \frac{1}{2}\left(-q_n^T\frac{p_{n+j} - p_{n-j}}{2jh} + p_n^T\frac{q_{n+j} - q_{n-j}}{2jh} - \frac{\beta_{n+j} - \beta_{n-j}}{2jh}\right) \end{aligned}$$

for $j = 1, \ldots$ and then compute the Richardson table entries as before:

$$T^n_{j,k+1} = T^n_{j,k} + \frac{T^n_{j,k} - T^n_{j-1,k}}{(1-k/j)^2 - 1}, \quad k = 1, \ldots, j-1, \tag{5}$$

The expression $T^n_{j,j}$ is a convenient way of computing the approximations and in addition it gives a way of estimating the error in the approximation $|\overline{H} - T_{j-1,j-1}| \approx |T_{j,j} - T_{j-1,j-1}|$, which is useful for finding a stopping criterion for the extrapolation process.

We mention in passing that accurate values of $\overline{H}$ might be obtained using Fourier series as well [1], however such methods seem most useful for quasi-periodic motions or scattering problems, while the approach taken here seems suitable for a broader range of problems.

The following corollary follows from Lemma 2.

Corollary 3. *Let p_n, q_n, β_n be given by the numerical scheme then there exists an m^* such that*

$$|\overline{H}(p_n, q_n, t_n) - T^n_{m^*,m^*}| \le \frac{C_1}{2}\frac{\rho \exp\left(-\frac{\pi\rho}{h}\right)}{h^2}\big(|q_n|_1\|\overline{p}\|_\rho + |p_n|_1\|\overline{q}\|_\rho + \|\overline{\beta}\|_\rho\big),$$

where ρ is such that the interpolating trajectory $(\overline{p}(t), \overline{q}(t), \overline{\beta}(t))$ is analytic for $|\Im(t)| < \rho$.

In Corollary 3 the parameter ρ related to the domain of analyticity of $\overline{y}(t)$ is undetermined. The following existence lemma will be useful for bounding ρ.

Lemma 4 (Domain of analyticity of the solution)**.** *Let $g(y,t)$ be an analytic vector field on the domain*

$$\begin{aligned}(z,\tau) &\in \mathcal{D}_y \times \mathcal{D}_t \\ \mathcal{D}_y &= \{z \in \mathbb{C}^d : |z - x_n|_\infty < r_y\} \\ \mathcal{D}_t &= \{t \in \mathbb{C}^d : |\Im(t)| < r_t\}\end{aligned}$$

Then $y(t)$ satisfying $y' = g(y,t)$, $y(0) = x_n \in \mathbb{R}^d$ is an analytic function of t on the domain

$$\mathcal{D} = \left\{ t \in \mathbb{C} : |t| < \min\left(\frac{r_y}{\|g\|_r}, r_t \right) \right\},$$

where $\|g\|_r := \sup_{(z,t)\in\mathcal{D}_y\times\mathcal{D}_t} |g(z,t)|_\infty$.

We can now combine the estimates of Corollary 3 and Lemma 4 to determine a bound on ρ, and hence on the error in the numerically computed modified energy.

Theorem 5 (Numerical modified energy)**.** *Let $H(p,q)$ be analytic in its arguments, and let p_n, q_n, β_n be computed by the algorithm (2). Then for each n there exists an m^* such that we have the error bound*

$$|\overline{H}(p_n,q_n,t_n) - T^n_{m^*,m^*}| \le \frac{C_1}{2h^2} \exp\left(-C_2 \frac{\delta}{h\|f\|_{\mathcal{D}}} \right) \left(|q_n| \|\overline{p}\|_\rho + |p_n| \|\overline{q}\|_\rho + \|\overline{\beta}\|_\rho \right),$$

where $C_2 < 2.14707$ (and $\rho = 0.6835\delta/\|f\|_{\mathcal{D}}$).

The error bound in Theorem 5 shows that we are able to track the modified energy exponentially accurately. Moreover the bound displays the same dependency on the parameters h, $\tilde{r}_y$ and $\|f\|_{\mathcal{D}}$ as Theorem 1.

The bounding-constant C_2 is however smaller than the $2\pi/e$ found in the proof of Theorem 1. It is unclear to us if this is due to the proof techniques applied, or if it is an actual weakness of the extrapolation method when applied to estimate the derivatives and thus the modified energy.

6. Numerical experiments.

We have implemented the extrapolation algorithm using the Arprec multiple-precision library in order to avoid pollution by round-off errors and to be able to verify the theory to high accuracy.[1] We set the precision to 120 digits. Most experiments are done using the standard Störmer–Verlet scheme (also known as the leap frog scheme). All the experiments were also repeated with two fourth-order splitting schemes to check for dependence on splitting scheme coefficients. No noteworthy dependence was found, and we only present these results for the Kepler experiment.

An early experiment verifying the exponentially small drift in modified energy was done by Benettin and Giorgilli [2] who used a Hamiltonian of the form $H = \frac{1}{2}(p_1^2+p_2^2) + U(q_1^2+q_2^2)$ with the potential function U vanishing fast as its argument becomes large. In this case, the exponentially small effects can be observed directly because methods of the form (1) preserve the energy H exactly when U is identically zero. To carry out this experiment the initial values $y_0 = (p_1(0), p_2(0), q_1(0), q_2(0))$

[1] The Arprec library is available from http://crd.lbl.gov/~dhbailey/mpdist/. The C++ source code for our experiments can be downloaded from http://www1.maths.leeds.ac.uk/~jitse/software.html

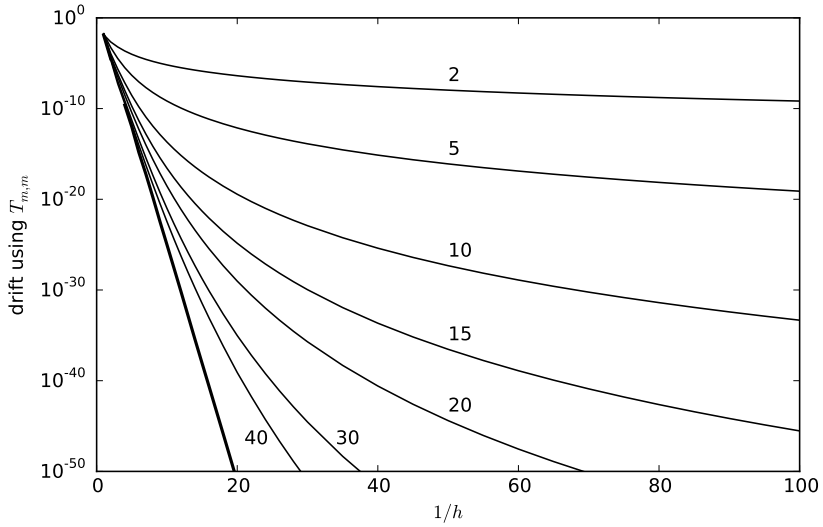

FIGURE 1. Drift in modified energy for the pendulum as a function of step size h for various values of m. The thick line indicates the limit.

are then chosen so that U vanishes and that the trajectory passes close to $(q_1, q_2) = (0, 0)$ before ending at some point $y_T = (p_1(T), p_2(T), q_1(T), q_2(T))$ where again U vanishes. Carrying out that simulation the difference $|H(y_0) - H(y_T)|$ is observed to be $\mathcal{O}(\exp(-C/h))$ where C is some unspecified constant.

We repeated this experiment using our method, and found that in this case it had *zero error* so the experiments we consider will not have this type of Hamiltonian. This matter warrants further investigations, but we have not pursued these in this paper.

6.1. **The pendulum.** In this experiment we apply the Störmer–Verlet method to the pendulum, which has Hamiltonian $H = p^2/2 - \cos(q)$, integrated over the time interval $[0, 100]$. Figure 1 reports the drift in the modified energy computed using $T_{m,m}$ for $m = 2, 5, 10, 15, 20, 30, 40$. Here, and in the other plots, the drift is defined as the difference between the maximum of the modified energy over the integration interval and its minimum. The figure suggests that the approximations $T_{m,m}$ converge for this problem. The limit is indicated by the thicker line in the left part of the plot, which shows that the drift follows the $\exp(-c/h)$ behaviour predicted by the theory.

The initial value for the experiment reported in Figure 1 is $q(0) = 0$ and $p(0) = 1$. Next we study the effect of varying the initial condition. The result is shown in Figure 2. The Störmer–Verlet method shows improved energy preservation near the equilibrium point, revealing the $\exp(-c/h\|f\|_{\mathcal{D}})$ dependency on step size and on $\|f\|_{\mathcal{D}}$ which decreases as $p \to 0$.

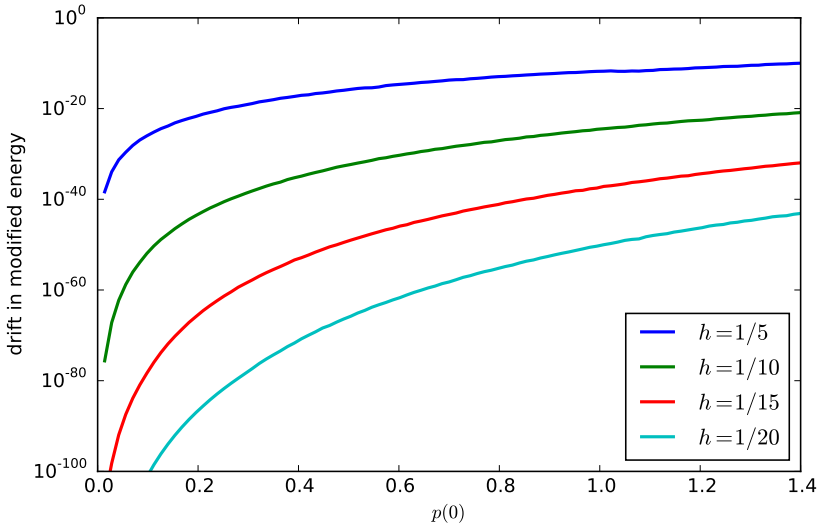

FIGURE 2. Drift in modified energy for the pendulum as a function of the initial momentum $p(0)$.

6.2. **Kepler problem.** The Kepler problem for one particle in a central force field is given by the Hamiltonian

$$H = \frac{1}{2}(p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}}.$$

We integrate over the interval $[0, 100]$ starting from the point

$$p_1 = 0, \qquad q_1 = 1 - ecc,$$
$$p_2 = \sqrt{\frac{1+ecc}{1-ecc}}, \qquad q_2 = 0.$$

where $0 \leq ecc < 1$ is the eccentricity of the orbit.

The left plot of Figure 3 shows the theoretical $\exp(-c/h)$ behavior. In contrast with the pendulum, for this problem the $T_{m,m}$ do not converge as $m \to \infty$, but the sequence has to be truncated at a suitably chosen point. To find the optimal m, we approximate the error in the m-th estimate as

$$|\overline{H} - T_{m,m}| \approx |T_{m,m} - T_{m-1,m-1}|. \tag{6}$$

We compute this estimate for $m = 2, 3, \ldots, 200$ and select the value of m for which the estimated error is minimized. This procedure recovers the expected exponential behaviour.

The right plot of Figure 3 shows the dependence of the drift in the modified energy on the eccentricity of the orbit. Almost circular orbits with a low eccentricity show much better preservation of the energy than highly elliptical orbits. An instability occurs at a critical eccentricity which depends on the step size. This can be explained by the fact the the topology of the energy levels of the modified energy changes with h, thus leading to unbounded trajectories and instability.

We used the second-order Störmer–Verlet method to produce Figure 3. We ran the experiments again with two fourth-order splitting methods: Yoshida's scheme

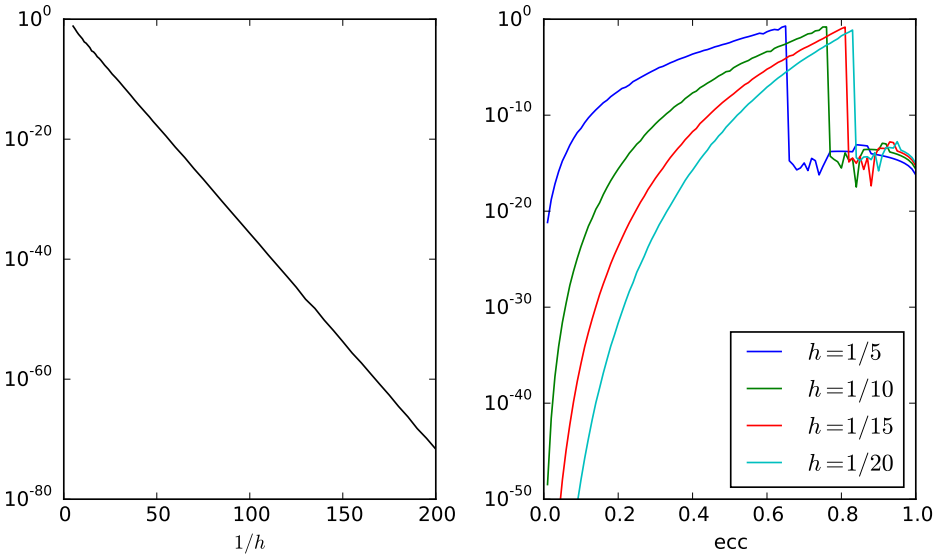

FIGURE 3. Drift in modified energy for the Kepler problem as a function of step size h for eccentricity $ecc = 0.6$ (left plot), and as a function of eccentricity (right plot).

based on extrapolation [20] and a fourth-order scheme due to Blanes and Moan [3]. The last scheme is optimized for problems of the type we have considered. It has very small error coefficients at the cost of many stages, leading to coordinate errors which are typically three orders of magnitude smaller than Yoshida's method at the same computational cost. The plots for the drift in modified energy of both Yoshida's method and the Blanes–Moan method look the same as for the Störmer–Verlet method. In particular, the constant c in $\exp(-c/h)$ is the same. However, the drift in the modified energy for the Störmer–Verlet method is approximately a factor of three smaller than Yoshida's method and a factor of four smaller than the method of Blanes and Moan.

The left plot in figure 4 illustrates for several different step sizes how the modified energy varies. There are peaks when the particle is near the singularity at the origin. The crucial point is that the energy essentially recovers its value after this point before another close encounter. The plot on the right compares the three different methods. It is seen that the methods give rather different results, even though the maximal variation in the modified energy is almost the same for the methods. The Blanes–Moan method seems to preserve the modified energy better after the close encounter, which might indicate a special advantage of this method when applied to the Kepler problem. If, however, the time steps are scaled so that the computational cost is the same for the three methods, the Störmer–Verlet method will preserve the modified energy better than the high-order methods.

Figure 5 shows the accumulated change in energy, $|\overline{H}(p_0, q_0, t_0) - \overline{H}(p_n, q_n, t_n)|$, and the instantaneous change in energy, $|\overline{H}(p_{n-1}, q_{n-1}, t_{n-1}) - \overline{H}(p_n, q_n, t_n)|$, together with the optimal m found by the error estimate (6). The graph shows that near the singularity quite a high order m (which we bounded by 200) is used. This

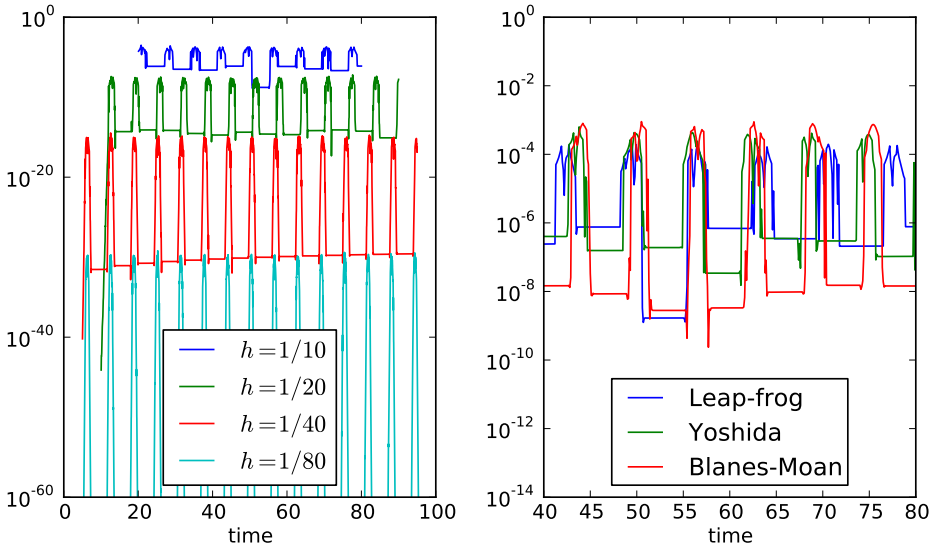

FIGURE 4. The difference between modified energy at a given time and the initial modified energy for the Kepler method with eccentricity $ecc = 0.6$. The left plot shows the results for the Störmer–Verlet method for different step sizes, while the right plot shows the results for different methods. All methods are run with step size $h = 0.1$, so Störmer–Verlet does considerably less work.

indicates that information from the smooth parts of the trajectory is used near singularities, and that it might be important to use very high order approximations to get a clear picture of the drift. The graph also indicates that the algorithm can track instantaneous changes in energy.

Away from the parts of the orbit where the singularity at the origin is approached most closely, a lower value of m suffices. It is thus useful to find a more efficient method for finding the optimal m instead of computing the error estimate for all m up to some large value (here, 200). We found good results with the following ad-hoc termination criterion: compute the error estimate (6) for all m up to the first value of m for which

$$\max_{j=m-11,\ldots,m-1} |T_{j,j} - T_{j-1,j-1}| \leq \max_{j=m-10,\ldots,m} |T_{j,j} - T_{j-1,j-1}|,$$

and then choose the m with the minimal error estimate. The plots produced by this criterion are nearly indistinguishable from the plots produced when all m up to 200 are considered.

6.3. **Hénon–Heiles system.** The Hamiltonian of the Hénon–Heiles system is given by

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

Skeel *et al.* investigated the theoretical $\exp(-c/h)$ behaviour of the drift in the modified energy for this system, and report that the results are "less convincing" [5, §2.4]. We revisit this problem, using instead the extrapolation method to achieve

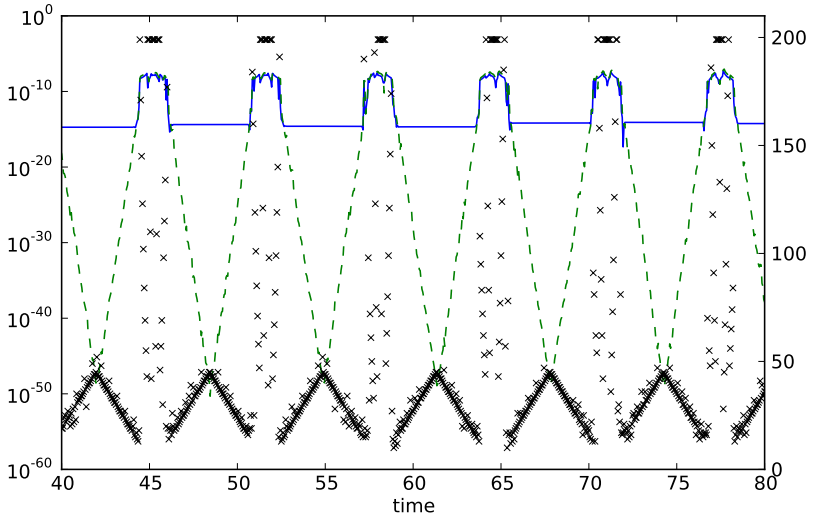

FIGURE 5. The change per step in the modified energy (dashed) and the accumulated change (solid), and the optimal order m (capped by 200, marked by 'x'). The axis for the energies is on the left, which the right axis is for the order m. This is for the Störmer–Verlet method applied to the Kepler problem with $h = 1/20$ and $ecc = 0.6$.

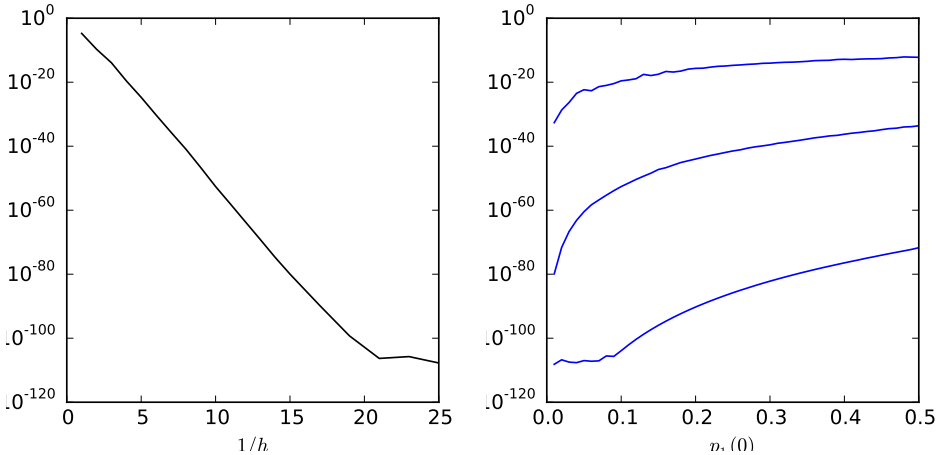

FIGURE 6. Drift in modified energy for the Störmer–Verlet method applied to the Hénon–Heiles problem as a function of step size h with initial condition $p_1(0) = 0.1$ (left plot), and as a function of p_1 for step sizes $h = 0.25, 0.1, 0.05$ (right plot).

arbitrary high orders. Figure 6 shows that the expected exponential smallness is indeed present, and that there is no problem in using the algorithm other than allowing for large values of m (we again capped m at 200). The effect of round-off error becomes visible when $1/h$ exceeds 20; remember that all computations are done with 120 digits.

The right plot shows how the maximal deviation of the modified energy changes as the initial condition for p_1 is varied; the initial conditions for the other variables are fixed as $q_1 = q_2 = p_2 = 0$. This plot shows that there is no abrupt change in energy preservation when moving from regular, integrable motions (the region with energy $H < 1/12$ or, equivalently, $p_1 < 1/\sqrt{6} \approx 0.4$) to the chaotic regime of phase space (where $H > 1/12$).

We also applied the fourth-order methods due to Yoshida and Blanes and Moan to this problem, with the same results as for the Kepler problems: the Störmer–Verlet method shows slightly better energy preservation, but the value of c in the $\exp(-c/h)$ dependence is the same.

7. **Conclusions.** We have supplied rigorous estimates for a numerical algorithm that computes the modified energy for methods based on operator splitting of Hamiltonian systems. The estimate shows that the procedure can recover exponentially small estimates, known to exist theoretically. The estimates exhibit the same dependence on the important parameters $\tilde{r}_y$, h and $\|f\|_{\mathcal{D}}$, and can therefore in principle be used to extract their values from simulations. When comparing different splitting algorithms, it seems that in the limit $h \to 0$ the exponential remainder term only weakly depends on the method coefficients. Thus when considering the additional cost of optimized, many-stage, methods these will have a larger drift than the second-order Störmer–Verlet algorithm. In other words, when it comes to preserving the modified energy, cheap, low-order methods are preferable. Although we have not considered ODEs originating from Hamiltonian semidiscretization of PDEs it seems likely that for long time simulations a low-order method such as Störmer–Verlet is preferable if energy preservation is important.

Appendix.

Proof of Lemma 2. Without loss of generality we assume that $n = 0$.

By representing (4) as a contour integral we have

$$D_m y(0) = \frac{1}{2\pi i} \sum_{j=1}^{m} \oint_\gamma \frac{(-1)^{j+1}(m!)^2}{hj(m-j)!(m+j)!} \left\{ \frac{1}{z - jh} - \frac{1}{z + jh} \right\} y(z) dz,$$

where the contour γ includes the points $-mh, \ldots, mh$ and excludes singularities of y, as sketched in Figure 2.

The derivative is given by $y'(0) = \frac{1}{2\pi i} \oint_\gamma \frac{y(z)dz}{z^2}$, so the error in the approximation becomes

$$E_m(y)(0) = D_m y(0) - y'(0) = \frac{1}{2\pi i} \oint_\gamma K_m(z) y(z) dz, \tag{7}$$

where the kernel is defined by

$$K_m(z) = \frac{(-1)^{m+1}(m!)^2 h^{2m}}{z^2(z^2 - h^2)(z^2 - (2h)^2) \cdots (z^2 - (mh)^2)}.$$

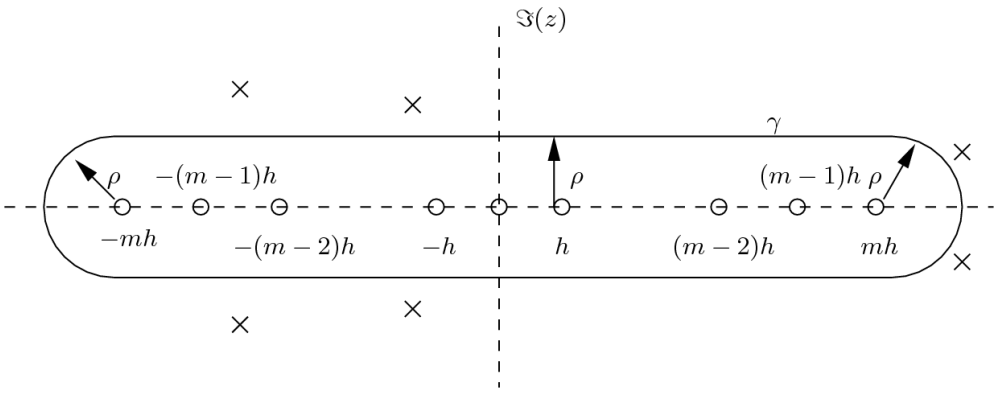

FIGURE 7. The contour of integration used in the proof of Lemma 2.

Along the curve γ, the kernel K_m achieves its maximum in modulus at $z = i\rho$, and the maximum is

$$\begin{aligned}|K_m(i\rho)| &= \frac{(m!)^2 h^{2m}}{\rho^2(\rho^2+h^2)\cdots(\rho^2+(mh)^2)} = \frac{1}{\rho^2(1+\frac{\rho^2}{h^2})\cdots(1+\frac{\rho^2}{(mh)^2})} \\ &= \frac{\pi}{\rho h \sinh(\frac{\pi\rho}{h})} \prod_{j=m+1}^{\infty}\left(1+\frac{\rho^2}{(hj)^2}\right)\end{aligned}$$

where the last equality follows from $\prod_{j=1}^{\infty}(1+\frac{\rho^2}{(hj)^2}) = \frac{h}{\pi\rho}\sinh(\pi\rho/h)$. The product can be bounded as

$$\begin{aligned}\log \prod_{j=m+1}^{\infty}\left(1+\frac{\rho^2}{(hj)^2}\right) &= \sum_{j=m+1}^{\infty}\log\left(1+\frac{\rho^2}{(hj)^2}\right) \\ &\le \int_m^{\infty}\log\left(1+\frac{\rho^2}{(hx)^2}\right)dx \le \frac{\rho^2}{h^2 m},\end{aligned}$$

yielding $\prod_{j=m+1}^{\infty}(1+\frac{\rho^2}{(hj)^2}) \le \exp(\frac{\rho^2}{h^2 m})$, and thus

$$|K_m(i\rho)| \le \frac{\pi}{\rho h \sinh(\frac{\pi\rho}{h})}\exp\left(\frac{\rho^2}{h^2 m}\right).$$

Since the length of the contour is $2\pi\rho+4mh$, the error expression (7) can be bounded as

$$|D_m y(0) - y'(0)| \le \frac{(\pi\rho+2mh)\exp\left(\frac{\rho^2}{h^2 m}\right)}{\rho h \sinh\left(\frac{\pi\rho}{h}\right)}\|y\|_\rho$$

where $\|y\|_\rho := \sup_{|\Im(z)|<\rho}|y(z)|_\infty$. This upper bound is minimized by choosing m so that $\frac{d}{dm}(\pi\rho+2mh)\exp(\rho^2/h^2 m)$ vanishes, i.e. $m \approx \frac{\rho^2}{h^2}$. This gives the bound

$$|D_m y(0) - y'(0)| \le \frac{e(\pi+2\rho^2/h^2)}{h\sinh\left(\frac{\pi\rho}{h}\right)}\|y\|_\rho \le C_1\frac{\rho\exp\left(-\frac{\pi\rho}{h}\right)}{h^2}\|y\|_\rho$$

for some constant $C_1 > 0$. □

Proof of Corollary 3. This follows from

$$\begin{aligned}
&|\overline{H}(p_n,q_n,t_n) - T^n_{m^*,m^*}| \\
&\quad \le \tfrac{1}{2}|\overline{q}^T(\overline{p}' - D_{m^*}\overline{p})| + \tfrac{1}{2}|\overline{p}^T(\overline{q}' - D_{m^*}\overline{q})| + \tfrac{1}{2}|\overline{\beta}' - D_{m^*}\overline{\beta}| \\
&\quad \le \tfrac{1}{2}|q_n|_1\|\overline{p}' - D_{m^*}\overline{p}\|_\rho + \tfrac{1}{2}|p_n|_1\|\overline{q}' - D_{m^*}\overline{q}\|_\rho + \tfrac{1}{2}\|\overline{\beta}' - D_{m^*}\overline{\beta}\|_\rho \\
&\quad \le \frac{C_1\rho\exp(-\frac{\pi\rho}{h})}{2h^2}\left(|q_n|_1\|\overline{p}\|_\rho + |p_n|_1\|\overline{q}\|_\rho + \|\overline{\beta}\|_\rho\right)
\end{aligned}$$

where the last inequality follows from Lemma 2. □

Proof of Lemma 4. We prove this by Picard iteration: set $\tilde{x}_1 = x_n$ and iterate $\tilde{x}_{k+1}(t) = x_n + \int_0^t g(\tilde{x}_k(s), s)\,ds$. Fix $t \in \mathcal{D}_t$, and assume at first that r_t is sufficiently large. For $\tilde{x}_{k+1}, \tilde{x}_k \in \mathcal{D}_y$

$$\begin{aligned}
&|g(\tilde{x}_{k+1},t) - g(\tilde{x}_k,t)|_\infty \\
&\quad = \left|\int_0^1 \frac{d}{ds} g\Big(\tilde{x}_{k+1} + s(\tilde{x}_k - \tilde{x}_{k+1}), t\Big)\,ds\right|_\infty \\
&\quad = \frac{1}{2\pi}\left|\int_0^1 \oint_{|z-s|=R} \frac{g\big(\tilde{x}_{k+1} + z(\tilde{x}_k - \tilde{x}_{k+1}), t\big)}{(z-s)^2}\,dz\,ds\right|_\infty \\
&\quad = \frac{1}{2\pi}\left|\int_0^1 \oint_{|w|=R} \frac{g\big(\tilde{x}_{k+1} + s(\tilde{x}_k - \tilde{x}_{k+1}) + w(\tilde{x}_k - \tilde{x}_{k+1}), t\big)}{w^2}\,dw\,ds\right|_\infty .
\end{aligned}$$

The radius R is restricted by the requirement that the argument of g lies within $\mathcal{D}_y$ or

$$\begin{aligned}
&|\tilde{x}_{k+1} + s(\tilde{x}_k - \tilde{x}_{k+1}) - x_n + w(\tilde{x}_k - \tilde{x}_{k+1})|_\infty \\
&\qquad \le |\tilde{x}_{k+1} + s(\tilde{x}_k - \tilde{x}_{k+1}) - x_n|_\infty + r|(\tilde{x}_k - \tilde{x}_{k+1}))|_\infty \\
&\qquad \le |t|\|g\|_r + R|\tilde{x}_k - \tilde{x}_{k+1}|_\infty < r_y
\end{aligned}$$

by using

$$\begin{aligned}
&|\tilde{x}_{k+1} + s(\tilde{x}_k - \tilde{x}_{k+1}) - x_n)|_\infty \\
&\qquad \le \sup_{|\tau|=|t|}\left|\int_0^\tau (1-s)\,|g(\tilde{x}_k(\tau),\tau)|_\infty + s\,|g(\tilde{x}_{k-1}(\tau),\tau)|_\infty\,d\tau\right| \le |t|\|g\|_r.
\end{aligned}$$

We may therefore choose

$$R = \eta\frac{r_y - |t|\|g\|_r}{|\tilde{x}_{k+1} - \tilde{x}_k|_\infty}, \quad 0 < \eta < 1$$

which gives the supremum-norm Lipschitz bound

$$|g(\tilde{x}_{k+1},t) - g(\tilde{x}_k,t)|_\infty \le \frac{\|g\|_r}{\eta(r_y - |t|\|g\|_r)}|\tilde{x}_{k+1} - \tilde{x}_k|_\infty.$$

Let $\Delta_{k+1}(|t|) = \sup_{|\tau|=|t|}|\tilde{x}_{k+1}(\tau) - \tilde{x}_k(\tau)|_\infty$, then the Picard iteration $\tilde{x}_1 = x_n$, $\tilde{x}_{k+1} = x_n + \int_0^t g(\tilde{x}_k(s), s)\,ds$, converges if $\Delta_k \to 0$ as $k \to \infty$, with

$$\Delta_{k+1}(t) \le \int_0^{|t|} \frac{\|g\|_r}{\eta(r_y - s\|g\|_r)}\Delta_k(s)\,ds, \quad \Delta_1(t) = |t|\|g\|_r$$

Introducing the generating function $G(\mu,|t|)=\sum_{k\geq 1}\mu^k\Delta_k$ we have

$$G(\mu,t)\leq \mu|t|\|g\|_r+\mu\int_0^{|t|}\frac{\|g\|_r}{\eta(r_y-s\|f\|_\rho)}G(\mu,s)\,ds.$$

Since the terms in this inequality are positive, an upper bound is the solution of

$$\frac{dG^+(\mu,|t|)}{d|t|}=\mu\|g\|_r+\frac{\|g\|_r}{\eta(r_y-|t|\|g\|_r)}G^+(\mu,|t|),\qquad G^+(\mu,|t|=0)=0,$$

i.e.

$$G^+(\mu,|t|)=\frac{\mu\eta\rho}{\eta+\mu}\left(\left(1-\frac{|t|\|g\|_r}{r_y}\right)^{-\mu/\eta}-\left(1-\frac{|t|\|g\|_r}{r_y}\right)\right).$$

Because G^+ is analytic in μ around $\mu=1$ provided $|t|<\frac{r_y}{\|g\|_r}$, the sequence $\Delta_k(|t|)$ converges uniformly to zero and hence $\tilde{x}_k(t)$ converges *uniformly* to the solution. Since each iterate $\tilde{x}_{k+1}(t)=x_n+\int_0^t f(\tilde{x}_k(s),s)\,ds$ is analytic in $t\in\{t\in\mathbb{C}:|t|<\min\{\frac{r_y}{\|g\|_r},r_t\}\}$ the uniform convergence gives by Weierstrass theorem that $y(t)=\tilde{x}_\infty(t)$ is analytic in this domain as well. □

Proof of Theorem 5. In Theorem 1 we take $g=\overline{f}$, thus $\|g\|_r\leq\frac{2}{1-\eta}\|f\|_{\mathcal{D}}$ with $\overline{f}$ analytic in $\overline{\mathcal{D}}'$. This gives that the $\overline{y}(t)$ is analytic in the domain

$$\left\{\tau\in\mathbb{C}:|\Im(\tau)|<\min\left(\frac{\tilde{r}_y-\delta}{\frac{2}{1-\eta}\|f\|_{\mathcal{D}}},\frac{\eta\delta}{e\|f\|_{\mathcal{D}}}\right)\right\}.$$

We find that the bound is optimized by picking $\tilde{r}_y=\eta\delta$ where $\eta=\frac{e-2}{e-\sqrt{2e}}$. Thus we may take $\rho=\frac{\eta\delta}{e\|f\|_{\mathcal{D}}}<0.6835\frac{\delta}{\|f\|_{\mathcal{D}}}$ in Corollary 3 giving the exponentially small bound ($t=nh$)

$$|\overline{H}(p_n,q_n,t)-T_{m^*,m^*}|\leq\frac{C_1}{2h^2}\exp\left(-C_2\frac{\delta}{h\|f\|_{\mathcal{D}}}\right)(|q_n|_1\|\overline{p}\|_\rho+|p_n|_1\|\overline{q}\|_\rho+\|\overline{\beta}\|_\rho),$$

where $C_2=0.6834\pi$. □

REFERENCES

[1] G. Benettin and F. Fasso. From Hamiltonian perturbation theory to symplectic integrators and back. *Appl. Numer. Math.*, 29:73–87, 1999.

[2] G. Benettin and A. Giorgilli. On the Hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms. *J. Stat. Phys.*, 74(5/6):1117–1143, 1994.

[3] S. Blanes and P. C. Moan. Practical symplectic Runge–Kutta and Runge–Kutta–Nyström methods. *J. Comput. Appl. Math*, 142(2):313–330, 2002.

[4] M. P. Calvo, A. Murua, and J. M. Sanz-Serna. Modified equations for ODEs. In *Chaotic Numerics*, volume 172 of *Contemp. Math.*, pages 53–74. Amer. Math. Soc., Providence, RI, 1994.

[5] R. D. Engle, R. D. Skeel, and M. Drees. Monitoring energy drift with shadow Hamiltonians. *J. Comput. Phys.*, 206(2):432–452, 2005.

[6] E. Hairer and C. Lubich. The life-span of backward error analysis for numerical integrators. *Numer. Math.*, 76:441–462, 1997.

[7] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Berlin, 2002.

[8] L. O. Jay. Beyond conventional Runge–Kutta methods in numerical integration of ODEs and DAEs by use of structures and local models. *J. Comput. Appl. Math*, 204(1):56–76, 2007.

[9] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2005.

[10] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001.
[11] P. C. Moan. On the KAM and Nekhoroshev theorems for symplectic integrators and implications for error growth. *Nonlinearity*, 17:67–83, 2004.
[12] P. C. Moan. On rigorous modified equations for discretizations of ODEs. Technical Report 2005-3, Geometric Integration Preprint Server, 2005. Available from http://www.focm.net/gi/gips/2005/3.html.
[13] P. C. Moan. On modified equations for discretizations of ODEs. *J. Phys. A*, 39(19):5545–5561, 2006.
[14] S. Reich. Backward error analysis for numerical integrators. *SIAM J. Numer. Anal.*, 36(5):1549–1570, 1999.
[15] Z. Shang. KAM theorem of symplectic algorithms for Hamiltonian systems. *Numer. Math.*, 83:477–496, 1999.
[16] R. D. Skeel and D. J. Hardy. Practical construction of modified Hamiltonians. *SIAM J. Sci. Comput.*, 23(4):1172–1188, 2001.
[17] R. D. Skeel, G. Zhang, and T. Schlick. A family of symplectic integrators: Stability, accuracy, and molecular dynamics applications. *SIAM J. Sci. Comput.*, 18(1):203–222, 1997.
[18] P. F. Tupper. Ergodicity and the numerical simulation of Hamiltonian systems. *SIAM J. Appl. Dynam. Systems*, 4(3):563–587, 2005.
[19] J. Wisdom and M. Holman. Symplectic maps for the n-body problem: Stability analysis. *Astron. J.*, 104(5):2022–2029, 1992.
[20] H. Yoshida. Construction of higher order symplectic integrators. *Phys. Lett. A*, 150(5–7):262–268, 1990.

E-mail address: pcmoan@gmail.com
E-mail address: jitse@maths.leeds.ac.uk