On modified equations for discretizations of ODEs

2006 J. Phys. A: Math. Gen. 39 5545

(http://iopscience.iop.org/0305-4470/39/19/S13)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **39** (2006) 5545–5561 doi:10.1088/0305-4470/39/19/S13

On modified equations for discretizations of ODEs

P C Moan

Department of Mathematics, University of Oslo, Oslo, Norway

Received 2 September 2005, in final form 2 February 2006
Published 24 April 2006
Online at stacks.iop.org/JPhysA/39/5545

Abstract

The theory of modified equations (MEs) for discretizations of ODEs is reconsidered. Obstructions to convergence of series expansions of MEs are pinpointed and alternative approaches are presented which provide more accurate descriptions of numerical approximations through MEs. We emphasize how structural assumptions on the ODE can be used to improve estimates. Then we give arguments for a slightly alternative approach based on time-dependent MEs which avoids the asymptotic nature traditionally associated with MEs. Some applications of the theory are also provided.

PACS numbers: 02.60.−x, 02.60.Jh, 02.60.Lj

1. Introduction

Numerical simulations are an invaluable tool for discovering properties of the dynamics of nonlinear differential equations. Long time-span simulations are for example used to compute approximate Lyapunov exponents, reveal invariant quantities etc in order to establish stability properties of the exact trajectories.

The theory of 'modified equations' [19, 21] is an approach for understanding the effect of unavoidable approximation errors significant in such simulations. There one considers a system of ordinary differential equations[1]

$$y' = f(y), \qquad y \in \mathbb{R}^d, \qquad y(0) = y_0, \tag{1.1}$$

and assumes that the numerical trajectory $x_n \approx y(n \cdot h)$ is produced by a consistent one-step method, i.e. a mapping $\Psi_{h,f} : \mathbb{R}^d \mapsto \mathbb{R}^d$ such that $x_{n+1} = \Psi_{h,f}(x_n) = x_n + hf(x_n) + \mathcal{O}(h^2), x_0 = y(0)$. The aim is to establish the existence of a perturbed or 'modified' differential equation[2]

$$\overline{y}' = \overline{f}_h(\overline{y}) = f(\overline{y}) + \epsilon \overline{r}_1(\overline{y}), \qquad \overline{y} \in \mathbb{R}^d, \quad \overline{y}(0) = y(0) \tag{1.2}$$

[1] It is assumed that f satisfies appropriate conditions guaranteeing the existence and uniqueness of trajectories $y(t)$.
[2] We have introduced a dummy parameter ϵ to indicate the smallness of the perturbation.

such that its solution[3] exactly reproduces the numerical trajectory, $\overline{y}(n \cdot h) = x_n$. Stability properties of the numerical approximations $\{x_j\}$ can then be analysed by classical perturbative stability results for differential equations applied to (1.2). In particular, there is no need to re-prove special stability results for discretizations.

It is, however, well known that in general no time-independent perturbation $\epsilon\overline{r}_1$ exists that can achieve this goal and the series expansions giving $\epsilon\overline{r}_1$ are only asymptotic [2, 19, 31, 37].

1.1. Outline

After covering important results relevant to modified equations, we introduce several alternative approaches. First, we consider a simple approach based on time-rescaling leading to smooth time-dependent MEs. Despite its simplicity this approach leads to suboptimal estimates for the time-dependent part of the ME. Then we present a novel time-averaging approach based on the so-called Magnus series which gives 'explicit' expressions for the MEs in terms of iterated integrals. This approach is then specialized to perturbed problems, a setting more useful in this context. We elucidate the importance of the $(\mathrm{d}\exp)^{-1}_{hF}$ operator and discuss two important cases where this is bounded. Then we introduce an alternative approach based on coordinate transformations which, in addition to the exponentially small estimates, shows that a ME for analytic $\Psi_{h,f}$ can be made analytic in time as well as in space. This ME has the added benefit of exactly interpolating the numerical trajectory. At the end we give some applications of the theory.

1.2. Connections to the theory of dynamical systems

Sensible one-step methods can be viewed as diffeomorphisms acting on phase space. Smale has pointed out that every diffeomorphism can be exhibited as the Poincare map of a global cross section of some flow, i.e. the *suspension*. When a diffeomorphism is isotropic to the identity[4] it is well known that there exists a time-periodic vector field $\tilde{f}(y,t)$ [5] such that its time-1 flow is the diffeomorphism[6]. Indeed differentiating $\Psi_{\tau,f}$ with respect to τ and composing it with its inverse we arrive at the vector field

$$\tilde{f}(y,\tau) = \left(\frac{\partial}{\partial\tau}\Psi_{\tau,f}\right) \circ \Psi^{-1}_{\tau,f}. \tag{1.3}$$

Clearly the time $\tau = h$-flow of (1.3), $\phi_{h,\tilde{f}}$, is exactly $\Psi_{h,f}$. The vector field (1.3) is then h-periodically extended in τ, giving $\tilde{f}(y,\tau;h)$ and in this way we have $\phi_{n\cdot h,\tilde{f}} = \Psi^n_{h,f}$—the approximation after n iterations of the numerical method[7].

On the other hand, the *embedding problem* is as follows: can we make this vector field time-independent? For a compact phase space the answer to the embedding problem is generally negative [33].

Proposition 1. *There exist vector fields f and one-step methods $\Psi_{h,f}$ for which no time-independent vector field, $\overline{f}_h$, exists with time-h flow equal to $\Psi_{h,f}$.*

[3] Time-independent vector fields be over-lined, and so will their corresponding solutions.
[4] A diffeomorphism $\Psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is an isotropy if it is smoothly connected to the identity.
[5] Time dependent vector fields be indicated by a tilde.
[6] By the consistency requirement of numerical methods we only need to consider diffeomorphisms isotropic to the identity($\Psi_{0,f}(x) = x$).
[7] We note that this construction does not even give continuous time-dependency.

Proof. We prove this by considering a special case. On the one hand, we note that arbitrarily close to the identity there exists a diffeomorphism $\Psi : \mathbb{T} \mapsto \mathbb{T}$ with expansive periodic points $\Psi^p(x_0) = \Psi^{p-1} \circ \Psi(x_0) = x_0$, $\left|\frac{\mathrm{d}}{\mathrm{d}x}\Psi^p(x_0)\right| > 1$. Therefore $\frac{\mathrm{d}}{\mathrm{d}x}\Psi^{np}(x_0) = \left(\frac{\mathrm{d}}{\mathrm{d}x}\Psi^p\right)^n(x_0)$ can be made arbitrarily large for sufficiently large n.

On the other hand, if $\Psi = \phi_{h,\overline{f}}$ is without fixed points, then the vector field $\overline{f}$ generating Ψ cannot have any zeros. Then $t_0 = \int_{\mathbb{T}} 1/|\overline{f}(s)|\,\mathrm{d}s$ is finite and $\phi_{t_0,\overline{f}}$ is the identity map. This implies that $\Psi^n = \phi_{n,\overline{f}} = \phi_{t_0\lfloor n/t_0\rfloor + r_n,\overline{f}} = \phi_{r_n,\overline{f}}$ where $0 \leqslant r_n \leqslant t_0$. Hence $\phi_{r_n,\overline{f}}$ is uniformly bounded in C^1 and so is $\frac{\mathrm{d}}{\mathrm{d}x}\Psi^n$ in contrast to the case for maps. □

In the general setting the Kupka–Smale theorem implies that most diffeomorphisms have periodic points that are isolated from other periodic points with the same period while a periodic point for a flow is isolated only if it is a stationary point. Palis [34] argued that the only embeddable diffeomorphisms are the Morse–Smale ones that only have fixed points and are orientation preserving on all of their invariant manifolds.

1.2.1. Smoothness of the suspension vector field and closeness to an embedding. With regard to regularity in time, Douady [7] proved for symplectic mappings that one can make $\tilde{f}(y,t)$ Hamiltonian and smooth in time. By introducing a time-rescaling, $\Psi_{h\chi(t/h),f}$, where $\chi(0) = 0$, $\chi(1) = 1$ with higher derivatives, $\chi^{(k)}(x) = 0$ for $x = 0, 1$ we find by differentiating and inverting Ψ the smooth vector field

$$\tilde{f}(y,t;h) = \frac{\mathrm{d}\Psi_{h\chi(t/h),f}}{\mathrm{d}t} \circ \Psi^{-1}_{h\chi(t/h),f}(y) = f(y) + \epsilon\tilde{r}(y,t;h), \tag{1.4}$$

with $\phi_{t,\tilde{f}} = \Psi_{h\chi(t/h),f}$ for $t \in [0,h]$, with $\Psi_{h\chi(h/h),f} = \Psi_{h,f}$. The periodic extension of (1.4) then gives a vector field valid for all time whose flow exactly interpolates the numerical trajectory. By splitting $\tilde{f}(y,t;h) = f(y) + \epsilon\tilde{r}_1(y) + \epsilon\tilde{r}_2(y,t;h)$ where $\int_0^h \epsilon\tilde{r}_2(y,s;h)\,\mathrm{d}s = 0$ the problem of making a suspension as close to an embedding becomes a question of how small we are able to make $\epsilon\tilde{r}_2$ as $h \to 0$. This is most easily studied by Fourier series expansions in t of $\tilde{f}$. The decay in the Fourier coefficients then reflects the regularity in time. More regularity in t gives faster decreasing Fourier coefficients, $\tilde{f}_k$, and hence smaller $\epsilon\tilde{r}_2$.

If, for example, $\tilde{f}$ is Gevrey-γ in t then $\|\tilde{f}_k\| \leqslant M\exp\left(-c\sqrt[\gamma]{\frac{|k|}{h}}\right)$ holds [27][8]. While if we can make $\tilde{f}$ analytic in $t \in \{z \in \mathbb{C} : |\mathrm{Im}(z)| < \rho\}$ we have $\|\tilde{f}_k\| \leqslant M\exp\left(-\frac{2\pi\rho|k|}{h}\right)$.

Analyticity in time cannot be achieved by choosing an appropriate χ, and more involved procedures are needed. Kuksin and Pöshel [6, 20] have shown that for perturbed integrable analytic symplectomorphisms isotropic to the identity an analytic (in time and space) Hamiltonian vector field exists so that its flow is equal to the mapping itself. They also state that the existence of an analytic suspension vector field has a positive answer via the Grauert embedding theorem. Pronin and Treschev [36] have shown by a more constructive approach that this is indeed possible for general symplectomorphisms isotropic to the identity, although their result does not quantify the magnitude of the time-dependent perturbation.

Another approach to the embedding problem is through time-averaging, obviously closely connected with the classical theory of MEs. There one carries out iterative procedures eliminating time-dependence in the vector field; see, e.g., [35]. But since embeddings generally do not exist, such procedures will only lead to asymptotic results.

[8] A function χ is Gevrey-γ regular provided the derivatives are bounded as $\|\chi^{(k)}\|_\infty \leqslant Mc^k k^{\gamma k}$. The choice $\chi(t/h) = \frac{1}{2} - \frac{1}{2}\tanh\left(\frac{\cos^\eta(\pi t/h)}{\sin^\eta(\pi t/h)}\right)$ provides a suitable time-rescaling for $\eta = 1, 3, \ldots$ with $\gamma = 1 + 1/\eta$.

1.3. Classical modified vector fields for numerical schemes

In seeking explanations of the energy preservation properties of symplectic discretization schemes several formal procedures for constructing $\epsilon r_1(y)$ have been derived [10, 12, 13, 16, 39, 40, 47, 48].

Feng [10] and Tang [47] used the methodology of generating functions. Ruth [39] and Yoshida [48] noted that for operator splitting methods the Campbell–Baker–Hausdorff formula(see, e.g., [19]) would give the *modified vector fields* (MVF). Hairer [13] derived similar results for B-series methods. Gonzalez *et al* [12] gave a general construction for one-step methods, while a non-classical approach used by Wisdom and Holman [51] (see also [14]) applied a time-dependent Dirac-delta function formalism.

Example 1. Let us consider the simplest of all methods, Euler's method, applied to

$$y' = f(y), \qquad y(0) = y_0.$$

That is, $x_{n+1} = x_n + hf(x_n)$, $n = 0, 1, 2\ldots$, $x_0 = y_0$. The Taylor series of the exact solution is

$$y(t+h) = y(t) + hf(y(t)) + \frac{h^2}{2}\,\mathrm{d}ff(y(t)) + \frac{h^3}{3!}(\mathrm{d}^2 f(f,f) + \mathrm{d}f\,\mathrm{d}f\,f)(y(t)) + \cdots, \tag{1.5}$$

where $\mathrm{d}f$ denotes the Jacobian of f w.r.t. y and d^j higher vector derivatives. By replacing f by $\overline{f}_h = f + \sum_{j\geqslant 1} h^j \overline{f}_j$ in (1.5) we have

$$x(t+h) = x(t) + hf(x(t)) + h^2\left(\frac{1}{2}\,\mathrm{d}ff + \overline{f}_1\right) + h^3\left\{\frac{1}{3!}(\mathrm{d}^2 f(f,f) + \mathrm{d}f\,\mathrm{d}f\,f) + \overline{f}_2 + \frac{1}{2}(\mathrm{d}f\overline{f}_1 + \mathrm{d}\overline{f}_1 f)\right\}(y(t)) + \cdots.$$

Setting $x(t+h) = x_{n+1}$ and collecting equal powers of h we arrive at a sequence of identities $\overline{f}_j = G_j(f, \overline{f}_1, \ldots, \overline{f}_{j-1})$, which leads to the MVF

$$\begin{aligned}\overline{f}_h = f + \epsilon\overline{r}_1 &= f - \frac{h}{2}\,\mathrm{d}ff + \frac{h^2}{12}\,\mathrm{d}^2 f(f,f) + \frac{h^2}{3}\,\mathrm{d}f\,\mathrm{d}ff + \cdots \\ &= f + \sum_{j=1}^{\infty} h^j \overline{f}_j.\end{aligned}$$

We note that, in general, $\overline{f}_j$ is expressed as terms involving derivatives of the original vector field, and in particular $\overline{f}_j$ will depend on the jth derivatives of f.

1.3.1. Optimal truncations and exponentially small estimates. Assuming that f is C^∞ a truncation $\epsilon\overline{r}_{1,N} = \sum_{j=p}^{N-1} h^j \overline{f}_j$ leads to a bound

$$\|\Psi_{h,f} - \phi_{h,f+\epsilon\overline{r}_{1,N}}\| \leqslant C_N h^N, \qquad \forall N \in \mathbb{N} \tag{1.6}$$

for some positive constant C_N. Due to the general non-existence of a MVF pointed out in section 1.2 we cannot expect that $\lim_{N\to\infty} h^N C_N = 0$ for any $h > 0$. For analytic vector fields the growth in C_N can be estimated through the Cauchy integral formula and the use of a supremum norm $\|f(x)\|_\delta = \sup_{z\in\mathcal{D}_\delta(x)} |f(z)|_\infty$ where $\mathcal{D}_\delta(x) \in \mathbb{C}^d$ is some open neighbourhood of $x \in \mathbb{R}^d$ with radius $\delta > 0$. This gives with $\delta, \delta' > 0$ the bound $\|\mathrm{d}ff(x)\|_\delta \leqslant \frac{1}{\delta'}\|f(x)\|_\delta \|f(x)\|_{\delta+\delta'}$. By iterating this bound one typically [15, 26, 37] finds that

$$\|\overline{f}_j\|_\delta \leqslant c\left(\frac{c' j \|f\|_{\delta+\delta'}}{\delta'}\right)^j, \tag{1.7}$$

for some constants $c, c' >$ that depend on the numerical method. Neishstadt [31] used a bound similar to (1.7) to show that $\epsilon r_{1,N}$ is bounded provided $N < \delta'/c'h\|f\|_{\delta+\delta'}$. With such a choice of N (1.6) translates, by using the maximum principle [2, 15] or Taylor series remainders [37, 26], to the now well-known result.

Theorem 1. *Let $\Psi_{h,f}$ be an analytic mapping close to the identity (i.e. $h\|f\|_{\delta+\delta'}$ sufficiently small). Then there exists a vector field $\overline{f}_h^* = f + \epsilon\overline{r}_1^*$ so that its time-h flow satisfies*

$$\left\|\phi_{h,\overline{f}_h^*} - \Psi_{h,f}\right\|_\delta = \mathcal{O}\left(\exp\left(-\frac{c\delta'}{h\|f\|_{\delta+\delta'}}\right)\right)$$

for some constant $c > 0$.

Such analysis has been carried out by several authors; Benettin and Giorgilli did analysis for symplectic schemes based on a formal scheme by J Moser [29]. Hairer and Lubich [15] generalized their analysis to B-series methods for general vector fields, while Reich [37, 21] derived an alternative scheme which allowed him to deduce, in great generality, the properties of the MEs (see also [9]).

We note that this analysis is carried out for one iteration of $\Psi_{h,f}$, and that global estimates require iterations of this bound. Hence there is a lifespan of validity for the analysis [15] which for systems with exponentially diverging trajectories can be quite short.

1.3.2. On the structural properties of modified equations. The development of MEs was motivated by the energy preservation of symplectic schemes applied to Hamiltonian problems. It is now known that essentially all symplectic methods applied to global Hamiltonian vector fields lead to a $\epsilon\overline{r}_1$ that is also a global Hamiltonian vector field [19]. Indeed, some [39, 48] work directly with the Hamiltonian themselves. More generally, one can say [37] that if $\Psi_{h,f}$ resides in some subgroup of $\mathrm{Diff}(\mathbb{R}^d)$, then $f + \epsilon\overline{r}_1$ belongs to the corresponding subalgebra of vector fields. Furthermore, if $\Psi_{h,f}$ possesses a symmetry or time-reversing symmetry, then $f + \epsilon r_1$ shares the same symmetry. The smooth vector field (1.4), constructed as the tangent vector field of Ψ, will similarly reflect the invariant quantities of Ψ.

2. Modified equations through time-averaging

Early results [48, 39] on MEs were for operator splitting methods and were direct applications of the Campbell–Baker–Hausdorff (CBH) [19] formulae. These formulas give expansions in terms of commutators for the logarithm of $\exp(A)\exp(B)$. Applying the so-called *continuous CBH formula* we will now show, for general one-step methods, that a MVF, $\overline{f}_h = f + \epsilon\overline{r}_1$, can be explicitly written down in terms of $\tilde{f}$ (1.4) [26].

2.1. Some results for linear systems

Much of the algebraic structure and the convergence properties of MEs are captured by studying real linear systems, $f(y) = Ay$, $A \in \mathbb{R}^{d\times d}$, and their discretization.

If we discretize $y' = Ay$, $y \in \mathbb{R}^d$ by some linear one-step method it is well known that $x_{n+1} = R(hA)x_n$, where R is the stability function of the method. Clearly, $\overline{y}(t) = R(hA)^{t/h}y_0$ exactly interpolates x_n when $t = n \cdot h$. But $\overline{y}(t) = R(hA)^{t/h}y_0 = \exp\left(\frac{t}{h}\ln(R(hA))\right)x_0$; thus $\overline{y}(t)$ satisfies the autonomous differential equation

$$\overline{y}' = \frac{1}{h}\ln(R(hA))\overline{y} = \overline{A}(h)\overline{y}, \qquad \tilde{y}(0) = x_0.$$

Thus for linear differential equations constructing the MVF essentially amounts to computing the real logarithm of a real matrix. There is however an obstruction to the existence of a real logarithm.

Theorem 2 (Culver [5]). *Let R be a real square matrix. Then there exists a real logarithm, $\log(R)$ if and only if R is non-singular and each Jordan block of R belonging to the negative eigenvalues occurs an even number of times.*

The non-existence of a time-independent MVF for non-linear systems is therefore a result of two factors. One is the increasing bounds on derivatives of vector fields, while the other is a logarithmic singularity originating from a nonlinear version of Culver's theorem (e.g., Krein signature theory). For the linear problem the construction (1.4) leads to $\tilde{x}' = \tilde{A}(\tau; h)\tilde{x}$, where $\tilde{A}$ is periodic in τ. In this case, theorem 2 gives [26, 28].

Theorem 3. *Suppose $\int_0^{h^*} \|\tilde{A}(s; h^*)\|_2 \,\mathrm{d}s < \pi$ then $R(h^*A)$ has a real logarithm.*

When $h < h^*$ a convergent series expansion for the logarithm of $R(hA)$ was derived by Magnus [23], also known as the *continuous CBH* formula [3]. We start by setting $R(\tau A) = \exp(\overline{A}(\tau))$, i.e. imposing that R is the time-1 flow of $x' = \overline{A}(h)x$. Differentiation with respect to τ then gives

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\tau}\exp(\overline{A}(\tau)) = \tilde{A}(\tau)\exp(\overline{A}(\tau)) \\ &\qquad\qquad \Downarrow \\ &\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\exp(\overline{A})\right)\exp(-\overline{A}) = \tilde{A}(\tau; h). \end{aligned} \tag{2.1}$$

Magnus's insight was that $\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\exp(\overline{A})\right)\exp(-\overline{A}) = \frac{\exp(\mathrm{ad}_{\overline{A}})-1}{\mathrm{ad}_{\overline{A}}}\frac{\mathrm{d}\overline{A}}{\mathrm{d}\tau} = \sum_{j\geqslant 0}\frac{1}{(j+1)!}\mathrm{ad}^j_{\overline{A}}\frac{\mathrm{d}\overline{A}}{\mathrm{d}\tau}$, where $\mathrm{ad}^j_{\overline{A}}\frac{\mathrm{d}\overline{A}}{\mathrm{d}\tau} = \mathrm{ad}^{j-1}_{\overline{A}}\left[\overline{A}, \frac{\mathrm{d}\overline{A}}{\mathrm{d}\tau}\right]$, with $[\cdot,\cdot]$ denoting the commutator. Solving for $\frac{\mathrm{d}\overline{A}}{\mathrm{d}\tau}$ and carrying out Picard iterations, he obtained the terms

$$\overline{A}(h) = \log(R(hA)) = \int_0^h \tilde{A}(s; h)\,\mathrm{d}s + \frac{1}{2}\int_0^h\int_0^{s_1}[\tilde{A}(s_2; h), \tilde{A}(s_1; h)]\,\mathrm{d}s_2\,\mathrm{d}s_1 + \cdots,$$

where higher order terms are given in terms of iterated commutators and integrals. Through integration by parts, using the skew-symmetry and Jacobi identity of Lie algebras other forms of these expansions can be found. A canonical form of the expansion is given by the explicit formula of I Bialynicki-Birula *et al* [3], later rediscovered by R S Stritchartz [43] and V A Vinokurov [50]. I M Gelfand *et al* [11] derived the expression as a consequence of a theory of non-commutative functions. Let the step-function $\theta_i = 1$ if $t_i > t_{i-1}$ and 0 otherwise and $\Theta_n = \theta_{n-1} + \theta_{n-2} + \cdots + \theta_2$. Defining the function

$$L_n(t_n, \ldots, t_1) = \frac{\Theta_n!(n-1-\Theta)!}{n!}(-1)^{\Theta_n+1-n}$$

the expansion is simply

$$h\overline{A}(h) = \log(R(hA)) = \sum_{n\geqslant 1}\underbrace{\int_0^h\cdots\int_0^h}_{n} L_n(\underline{s})\tilde{A}(s_1)\tilde{A}(s_2)\cdots\tilde{A}(s_n)\,\mathrm{d}s_1\cdots\mathrm{d}s_n. \tag{2.2}$$

The nested commutator form can be recovered by the Dynkin transform [3], simply by substituting $\tilde{A}(s_1)\tilde{A}(s_2)\cdots\tilde{A}(s_n) \mapsto \frac{1}{n}[\tilde{A}(s_1), [\tilde{A}(s_2)\cdots[\tilde{A}(s_{n-1}), \tilde{A}(s_n)]]]$. We note that explicit CBH formulae like those of Dynkins for $\log(\exp(A_1)\exp(A_2)\cdots\exp(A_n))$ are obtained by setting $\tilde{A}(t) = A_i$ when $t \in [i, i-1]$ and $h = n$.

2.2. *Transferring the results to nonlinear systems*

The algebraic structures in the linear setting carries over to nonlinear vector fields (1.4). It is well known that for a time-independent vector field $\overline{f}_h(y)$ and any smooth function $\rho : \mathbb{R}^d \mapsto \mathbb{R}$ that the time-1 flow satisfies $\rho \circ \phi_{1,\overline{f}_h} = \exp(\overline{F}_h)[\rho] = \rho + \sum_{j\geqslant 1} \frac{1}{j!}\overline{F}_h^k[\rho]$. Here capitals denote corresponding differential operators; $\overline{F}_h = \overline{f}_h \cdot \nabla = \sum_{i=1}^d \overline{f}_i \frac{\partial}{\partial y_i}$. We say $\overline{f}_h$ is an averaging vector field for $\tilde{f}(y,t)$ if $\rho \circ \phi_{1,\overline{f}_h} = \rho \circ \phi_{h,\tilde{f}} (= \rho \circ \Psi_{h,f})$, or $\exp(\overline{F}_h)[\rho] = \rho \circ \phi_{h,\tilde{f}}$. Differentiating this identity with respect to h, we obtain

$$\frac{\mathrm{d}\exp(\overline{F}_h)}{\mathrm{d}h}[\rho] = (\nabla\rho\tilde{f}) \circ \phi_{h,\tilde{f}} = (\tilde{F}[\rho]) \circ \phi_{h,\tilde{f}} = \exp(\overline{F}_h)\tilde{F}[\rho],$$

where $\tilde{F}(y,t;h) = \tilde{f}(y,t;h) \cdot \nabla$; hence

$$\exp(-\overline{F}_h)\frac{\mathrm{d}\exp(\overline{F}_h)}{\mathrm{d}h}[\rho] = (\mathrm{d}\exp)_{\overline{F}_h}\left(\frac{\mathrm{d}\overline{F}_h}{\mathrm{d}h}\right)[\rho] = \tilde{F}[\rho], \tag{2.3}$$

where $(\mathrm{d}\exp)_{\overline{F}_h} = \sum_{j\geqslant 0} \frac{(-1)^j}{(j+1)!}\mathrm{ad}^j_{\overline{F}_h}$. Apart from the difference in signs of this series, the equation is formally the same as (2.1). To account for this difference in (2.2) we define $K_n = (-1)^{1-n}L_n$, i.e. $K_n = \frac{\Theta_n!(n-1-\Theta_n)!}{n!}(-1)^{\Theta_n}$; hence the time-averaged vector field is formally given by

$$\begin{aligned}
h\overline{f}_h \cdot \nabla = \overline{F}_h &= \sum_{n\geqslant 1} \underbrace{\int_0^h \cdots \int_0^h}_{n} K_n \tilde{F}(s_1)\cdots\tilde{F}(s_n)\,\mathrm{d}s_1\cdots\mathrm{d}s_n,\\
&= \sum_{n\geqslant 1} \underbrace{\int_0^h \cdots \int_0^h}_{n} \frac{K_n}{n}[\tilde{F}(s_1),[\cdots,[\tilde{F}(s_{n-1}),\tilde{F}(s_n)]\cdots]]\,\mathrm{d}s_1\cdots\mathrm{d}s_n\\
&= \sum_{n\geqslant 1} \underbrace{\int_0^h \cdots \int_0^h}_{n} \frac{K_n}{n}[\tilde{f}(y,s_1),[\cdots,[\tilde{f}(y,s_{n-1}),\tilde{f}(y,s_n)]\cdots]]\\
&\quad \times \mathrm{d}s_1\cdots\mathrm{d}s_n \cdot \nabla,
\end{aligned}$$

where the latter bracket is the Lie–Jacobi bracket and $[\tilde{F}(s_i),\tilde{F}(s_{i+1})](y) = [\tilde{f}(s_i),\tilde{f}(s_{i+1})](y)\cdot\nabla = (d_y f(y,s_{i+1})f(y,s_i) - d_y f(y,s_i)f(y,s_{i+1}))\cdot\nabla$. We note that $\overline{f}_h$ is not of the form $f + \sum_{j\geqslant 1} h^j f_j(y)$, as is customary in MEs but rather $\sum_{j\geqslant 0} h^{j+1} f_j(y,h)$.[9] The advantage of this approach is that it is quite easy to obtain $\tilde{f}$ from a numerical scheme, from which an explicit expression for $\overline{f}_h$ follows by simply applying the continuous CBH formula. Exponentially small bounds are then obtained in a similar fashion to the traditional approach [26] using improved convergence estimates derived for linear problems. These, in turn, lead to improved estimates for the constant c in theorem 1 . Since the expansion is given in terms of Lie–Jacobi brackets the structural properties of the numerical method will be reflected in the MEs since the tangent vector field $\tilde{f}$ (1.4) inherits these.

2.3. *Time averaging for perturbed problems*

In the classical theory of MEs h is thought of as the small parameter and is used as the expansion parameter. More insight into time-averaging can be found if we instead find expansions in

[9] Dividing this MVF by h does, however, recover the traditional interpretation as a time-h flow.

powers of the dummy parameter ϵ. Since $\epsilon\tilde{r} = \mathcal{O}(h^p)$ each term in such an expansion will increase the order, in h, of the MVF by p, and not 1 as in the traditional approach.

We now let $\overline{F}_{h,n} = hF + \sum_{j=1}^{n} \epsilon^j \overline{F}_j$ and $\tilde{F} = F + \epsilon\tilde{R}$. Defining the vector field

$$G_{\epsilon,n}(\tau) = (\mathrm{d}\exp)_{\overline{F}_{\tau,n}}\left(\frac{\mathrm{d}\overline{F}_{\tau,n}}{\mathrm{d}\tau}\right),$$

obtaining a time-averaged vector field to order n in ϵ then amounts to finding $\overline{F}_j$ so that $G_{\epsilon,n}(\tau) = F + \epsilon R(\tau) + \epsilon^n R_n(\tau) + \mathcal{O}(\epsilon^{n+2})$. By Taylor expansion of $\mathrm{d}\exp$, it follows that

$$\begin{aligned} G_{\epsilon,n+1}(\tau) - G_{\epsilon,n} &= (\mathrm{d}\exp)_{\tau F}\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\epsilon^n\overline{F}_n\right) + \frac{1}{\tau}\sum_{j\geqslant 1}\frac{(-1)^{j+1}}{(j+1)!}\mathrm{ad}^j_{\tau F}(\epsilon^n\overline{F}_n) + \mathcal{O}(\epsilon^{n+1}) \\ &= (\mathrm{d}\exp)_{\tau F}\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\epsilon^n\overline{F}_n\right) + \frac{\mathrm{d}}{\mathrm{d}\tau}((\mathrm{d}\exp)_{\tau F})(\epsilon^n\overline{F}_n) + \mathcal{O}(\epsilon^{n+1}) \\ &= \frac{\mathrm{d}}{\mathrm{d}\tau}((\mathrm{d}\exp)_{\tau F}(\epsilon^n\overline{F}_n)) + \mathcal{O}(\epsilon^{n+1}). \end{aligned}$$

Rewriting this as $R_n(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}((\mathrm{d}\exp)_{\tau F}(\overline{F}_n(\tau)))$ an iteration can be carried out. For $n = 1$ we have, since $G_{\epsilon,0} = F$, that $\overline{F}_1(\tau)$ must satisfy the equation

$$R(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}((\mathrm{d}\exp)_{\tau F}(\overline{F}_1)) \quad \Rightarrow \quad (\mathrm{d}\exp)_{hF}\overline{F}_1(h) = \int_0^h R(s)\,\mathrm{d}s;$$

indeed every $\overline{F}_j(h)$ is given a solution of an equation of the form

$$(\mathrm{d}\exp)_{hF}\overline{F}_j(h) = \int_0^h R_j(s)\,\mathrm{d}s. \tag{2.4}$$

The invertibility of $(\mathrm{d}\exp)_{hF}$ is therefore central in the existence and convergence analysis of MEs[10]. Formally the inverse can be written as

$$(\mathrm{d}\exp)^{-1}_{hF} = \frac{\mathrm{ad}_{hF}}{1 - \exp(-\mathrm{ad}_{hF})}.$$

Since $\frac{z}{1-\mathrm{e}^{-z}}$ has singularities at $\sigma = \{z = 2\pi\mathrm{i}k, \forall k \in \mathbb{Z}\}$ such an inverse may exist if the spectrum of ad_{hF} does not intersect σ. We now consider two important cases where such a condition is satisfied.

Quasi-periodic $\phi_{t,f}$. If the flow of f is quasi-periodic with d' frequencies we can introduce new local coordinates $(\theta, I) \in \mathbb{T}^{d'} \times \mathbb{R}^{d-d'}$ so that $F = \sum_{i=1}^{d'} \omega_i \frac{\partial}{\partial\theta_i} = \omega \cdot \nabla_\theta$ and $R_n = r_n \cdot \nabla_{I,\theta}$. Expanding R_n in the Fourier series

$$R_n = \sum_{m\in\mathbb{Z}^{d'}} \exp(\mathrm{i}\theta \cdot m)\hat{R}_{n,m}(I,t),$$

we have by (2.4) the Fourier coefficients of $\overline{F}_n$

$$\hat{\overline{F}}_{n,m} = \frac{\mathrm{i}h\omega \cdot m}{1 - \exp(-\mathrm{i}h\omega \cdot m)}\int_0^h \hat{R}_{n,m}(s)\,\mathrm{d}s.$$

From this expression it is evident that if for given h, ω there exists $k \in \mathbb{Z}\backslash 0$ so that $h\omega \cdot m = 2\pi k$, then $\hat{\overline{F}}_{n,m}$ is not well defined, and averaging cannot be carried out[11]. If, on the other

[10] $(\mathrm{d}\exp)^{-1}_{hF}$ represents the conditioning for nonlinear ODEs $y' = f(y)$ in much the same way as the condition number for linear problems given in [1].

[11] The traditional MEs can be recovered by expanding $(\mathrm{d}\exp)^{-1}_{hF}$ in Taylor around $h = 0$. In a sense truncation of these Taylor series can be viewed as a regularization of $(\mathrm{d}\exp)^{-1}_{hF}$ giving boundedness.

hand, $h\omega \cdot m$ is not an multiple of 2π convergence of $\overline{F}_n$ for an analytic vector field R_n can be assured if ω, h satisfies the strong non-resonance condition [19]

$$\left| \frac{\mathrm{i}h\omega \cdot m}{1 - \exp(-\mathrm{i}h\omega \cdot m)} \right| \geqslant \gamma |m|_1^{-\beta},$$

for some $\gamma > 0$, $\beta > 1$.[12] In this case, we can expect that the error of an optimally truncated MVF satisfies an estimate of the form

$$\|\Psi_{h,f} - \phi_{1,\overline{f}}\| = \mathcal{O}(\exp(-c/\epsilon^{1/\beta})),$$

for some positive constant c, i.e. exponentially small in the perturbation.

Although rigorous analysis in this case has not been carried out the claim is supported by analysis of invariant tori when f is an integrable Hamiltonian and $\Psi_{h,f}$ is symplectic; see [19, 26, 42]. If additional non-degeneracy assumptions are made on f, KAM theorems [19, 25, 44] indicate that the series might converge. It would therefore be useful to establish what kind of non-degeneracy conditions are necessary for carrying out a complete time-averaging.

Hyperbolic $\phi_{t,f}$. Another case in which $(\mathrm{d}\exp)_{hF}$ is invertible is when an invariant subset $\Omega \subset \mathcal{M}$ has a hyperbolic structure, allowing for the invariant splitting of the tangent bundle $T\mathcal{M}|_\Omega = (E^u \oplus E^s \oplus E^c)|_\Omega$ where E^c is the span of f and there exist two constants λ^s, $\lambda^u > 0$ together with a metric $|||\cdot|||$ so that, for all $t \geqslant 0$,

$$\begin{aligned}
&(\phi_{t,f})^* E^\sigma = E^\sigma, \qquad \sigma = u, s, c \\
&|||(\phi_{-t,f})^* v^u||| \leqslant \exp(-\lambda^u t)|||v^u||| \\
&|||(\phi_{t,f})^* v^s||| \leqslant \exp(-\lambda^s t)|||v^s|||.
\end{aligned}$$

where the push forward $(\phi_{t,f})^* v = (\mathrm{d}\phi_{t,f} v) \circ \phi_{t,f}^{-1} = \exp(\mathrm{ad}_{tF}) v$.

Noting that we can write $(\mathrm{d}\exp)_{hF} = \frac{1-(\phi_{-h,f})^*}{\mathrm{ad}_{hF}}$ and assuming that $\phi_{t,f}$ satisfies a strong transversality condition [22], we may split $\int_0^h R_n(s)\,\mathrm{d}s = R_n^s + R_n^u + R_n^c$ into corresponding invariant tangent spaces. Then using $(\phi_{-h,f})^*(\phi_{h,f})^* = I$ we get when inserted into (2.4)

$$\begin{aligned}
\overline{F}_n &= R_n^c + \mathrm{ad}_{hF} \left\{ \frac{1}{1 - (\phi_{-h,f})^*} R_n^u - \frac{(\phi_{h,f})^*}{1 - (\phi_{h,f})^*} R_n^s \right\} \\
&= R_n^c + \mathrm{ad}_{hF} \left\{ \sum_{j \geqslant 0} (\phi_{-jh,f})^* R_n^u - \sum_{j \geqslant 1} (\phi_{jh,f})^* R_n^s \right\}.
\end{aligned}$$

Assuming that the metric $|||\cdot|||$ allows a commutator bound $|||\mathrm{ad}_{hF} G||| \leqslant c|||hf|||\,|||g|||$ for some $c > 0$, it follows that

$$|||\overline{f}_n||| \leqslant |||r_n^c||| + ch|||f||| \left\{ \sum_{j \geqslant 0} \exp(-hj\lambda^u) |||r_n^u||| + \sum_{j \geqslant 1} \exp(-hj\lambda^s) |||r_n^s||| \right\} < \infty,$$

where $f_n, r_n^s, \ldots$ are the components of $F_n, R_n^s \ldots$. This shows that $(\mathrm{d}\exp)_{hF}$ is invertible, and improvements to the bound of theorem 1 can be expected. Indeed M-C Li [22] proves:

Theorem 4. *Let $\phi_{t,f}$ be a C^{p+1} flow satisfying axiom A and the transversality condition, and let $\Psi_{h,f}$ be a method of order p. Then for sufficiently small h there exists a homeomorphism Ξ and a continuous real valued function τ_h on $\mathcal{M}$ such that $\Xi \circ \phi_{h+h\tau_h(x),f} = \Psi_{h,f} \circ \Xi$. In addition, $\max\{|\Xi(x) - x|, |\tau_h|\} \leqslant Ch^p$ for some positive constant C.*

[12] It is known that when $\beta > d'$ the measure of such orbits is dense.

Clearly, this theorem implies the existence of a MVF in the analytic category when allowing for a conjugacy. But this conjugacy is typically not analytic even for analytic $\Psi_{h,f}$ [30], leaving the optimal truncation error estimate an open issue. Similar to the non-resonant quasi-periodic case it would, however, be interesting to establish what assumptions are necessary for analytic $\Psi_{h,f}$ to establish the convergence of the MVF without conjugacy.

3. On the regularity of the remainder term in modified equations

In section 1.2.1 we argued that an analytic suspension will essentially recover the exponential estimate of theorem 1 simply by the exponential decay in the Fourier coefficients of $\tilde{f}$.

Motivated by the result of Neishstadt [31] on exponential small remainders in time-averaging and that of Kuksin and Pöshel [6, 20] on analytic suspensions a new approach to MEs was pursued in [27]; see also [36]. This approach takes as a starting point the smooth vector field (1.4) and, by a coordinate transformation, $\Xi : y \mapsto \overline{y}$, establishes the existence of an analytic and h-periodic perturbation ϵr_2. We start by defining the coordinate transformation $\Xi(s, t, y)$ as the s-flow of a vector field w which is to be determined;

$$\frac{\mathrm{d}}{\mathrm{d}s}\overline{y} = w(\overline{y}, t, s), \qquad \overline{y}(t, s = 0) = \tilde{y}(t), \tag{3.1}$$

with $\overline{y}(t, s) = \Xi(s, t, y)$, $\overline{y}(t, s = 0) = y(t)$ and the smooth h-periodic vector field $\tilde{f}(y, t; h)$ is transformed into a vector field $\overline{f}(\overline{y}, t)$. In other words, in the new coordinates $\overline{y}(t, s)$ satisfies the differential equation

$$\overline{y}' = \overline{f}(\overline{y}, t, s), \qquad \overline{y}(0, s) = \Xi(s, 0, \tilde{y}(0)). \tag{3.2}$$

By differentiating (3.1) with respect to t and (3.2) with respect to s, we obtain

$$\frac{\partial}{\partial s}\overline{f} = \frac{\partial}{\partial t}w + [w, \overline{f}], \qquad \overline{f}(\overline{y}, t, s = 0) = \tilde{f}(\overline{y}, t),$$

where $[\cdot, \cdot]$ denotes the Lie–Jacobi bracket. Using the Fourier series representation of vector fields and defining $w_k(\overline{y}, s) = \mathrm{i}\sigma(k)\overline{f}_k(\overline{y}, s)$, where σ is the sign function[13] we have

$$\frac{\partial}{\partial s}\overline{f}_k = -\frac{2\pi|k|}{h}\overline{f}_k + \sum_{p+q=k} \mathrm{i}\sigma(p)[\overline{f}_p, \overline{f}_q], \qquad \overline{f}_k(y, s = 0) = \tilde{f}_k(y).$$

The motivation behind this choice of transforming vector field w can be seen if we disregard the nonlinear Lie–Jacobi bracket. Then $\overline{f}_k(y, s) = \exp\left(-\frac{2\pi|k|s}{h}\right)\tilde{f}_k(y)$ and hence $\overline{f}_k$ represents an analytic function[14]. When taking the nonlinearity into account one can show [27] that $\overline{f}_k$ is essentially bounded as

$$\|\overline{f}_k\|_\delta < \frac{c's}{1 - c''s\|f\|_{\delta+\delta'}/\delta'} \exp\left(-\frac{2\pi|k|s}{h}\right)$$

for some constants $c', c'' > 0$. Hence for $s < \delta'/\|f\|_{\delta+\delta'}c''$, $\overline{f}_k(\overline{y}, s)$ represents an analytic vector field. A result similar to the classical asymptotic result of theorem 1 is then obtained by choosing the transformation parameter s so that the leading term of the time-dependent part $\epsilon\tilde{r}_2 = \sum_{k\neq 0}\overline{f}_k(y, s)\exp\left(\frac{2\pi\mathrm{i}kt}{h}\right)$ is minimized.

Theorem 5. *[27] Let $\Psi_{h,f}$ be an analytic one-step method. Then there exists a MVF $\overline{f}(y, t; h) = f(y) + \epsilon\overline{r}_1(y) + \epsilon\tilde{r}_2(y, t; h)$, where $\epsilon\tilde{r}_2$ is analytic and h-periodic in t so that its*

[13] This transform will naturally preserve possible Lie algebraic and time-reversing structures $\tilde{f}$ might possess.
[14] In this case, we may let $s \to +\infty$ and remove the time-dependency all together since $\overline{f}_{k\neq 0} \to 0$.

flow exactly interpolates the numerical trajectory for all time. If the step-size is sufficiently small (e.g. $h\|f\|_{\delta+\delta'} < 2\pi\delta'/e$), then $\epsilon\tilde{r}_2$ is exponentially small[15] *in h:*

$$\|\epsilon r_2\|_\delta \leqslant C \exp\left(-\frac{2\pi\delta'}{eh\|f\|_{\delta+\delta'}}\right).$$

The main advantage of theorem 5 over theorem 1 is that it provides a vector field which exactly represents the numerical trajectory. Indeed, classical stability results from dynamical systems theories can now been used directly to explain the dynamics of numerical approximations. One way of achieving this is by extending the phase space to take account of the time variable ($\overline{y}_\tau = t$)

$$\underbrace{\begin{pmatrix}\overline{y}\\ \overline{y}_\tau\end{pmatrix}'}_{y'_+} = \underbrace{\begin{pmatrix}f(\overline{y})\\ 1\end{pmatrix}}_{f_+(y_+)} + \underbrace{\begin{pmatrix}\epsilon r_1(\overline{y}) + \epsilon r_2(\overline{y}, \overline{y}_\tau)\\ 0\end{pmatrix}}_{\epsilon r_+(y_+)}, \qquad \overline{y}(0) = y(0), \quad \overline{y}_\tau(0) = 0.$$

It is then interesting to establish if when $y' = f(y)$ satisfies some conditions guaranteeing stability under small perturbations does $y'_+ = f_+(y_+)$ satisfy conditions guaranteeing stability as well. One such attempt of establishing a general stability theory for symplectic discretization of close to integrable Hamiltonian systems can be found in [25]. One can in principle through theorem 5 establish numerical stability theorems for systems where *structural stability theories* such as KAM theory for Hamiltonian, time-reversible and divergence free vector fields and hyperbolic systems hold. In classical versions of these theories structural properties of the perturbation ϵr, and thus the numerical method Ψ are also central, thus motivating the construction of geometric integration methods [19, 21, 41].

4. Results derived from modified equations

Example 2 (Almost preservation of modified energy)**.** By theorem 5 the numerical trajectory produced by applying a symplectic method to a hamiltonian H is generated by a time dependent Hamiltonian $\overline{H} = H + R_1 + R_2(t)$. Along this trajectory the variation in the energy $\overline{H}$ is $\mathrm{d}\overline{H}/\mathrm{d}t = \{\overline{H}, \overline{H}\} + \partial/\partial t \overline{H} = \partial/\partial t R_2(t)$, where $\{\cdot,\cdot\}$ is the Lie-Poisson bracket. By theorem 5 R_2 is exponentially small for small h, thus by analyticity $\partial/\partial t R_2$ is exponentially small as well, and this extremely slow drift in energy is one of the hallmarks of symplectic discretizations.

Example 3 (Drift in energy)**.** We will consider the Hamiltonian [21]

$$H = \underbrace{\frac{1}{2}\left(p_1^2 + q_1^2\right)}_{H_{\text{Slow}}} + \underbrace{\frac{1}{2\epsilon}\left(p_2^2 + \left(1 + \alpha q_1^2\right)q_2^2\right)}_{H_{\text{Fast}}},$$

with parameters $\epsilon = 0.025, \alpha = 0.01$, and initial values $p_1(0) = 1, q_1(0) = 0, p_2(0) = 2\sqrt{\epsilon}, q_2(0) = 0$.

We discretize using the symplectic leap-frog method evolving the fast and slow parts exactly, i.e. $\Psi_{h,H} = \phi_{h/2,H_{\text{Slow}}} \circ \phi_{h,H_{\text{Fast}}} \circ \phi_{h/2,H_{\text{Slow}}}$. We study the problem for time-steps ranging from small to quite large, our theory explaining the behaviour mainly in the limit $h \to 0$. For time steps with h/ϵ bounded away from 0, the alternative theory of modulated Fourier expansions [19] can provide additional insight.

Using the CBH formula to compute the modified Hamiltonians H_j to order $j+1$ in h, we first monitor the preservation of modified energies to verify the theory[16].

[15] A Murua has noted that it is possible to remove the factor e in the estimates through a more refined analysis than that of [27], achieving what we believe is an optimal bound on $\epsilon\tilde{r}_2$.

[16] Since the leap-frog scheme is symmetric the modified energy will be of the form $\overline{H} = H + h^2\Delta H_2 + h^4\Delta H_4 + \cdots$.

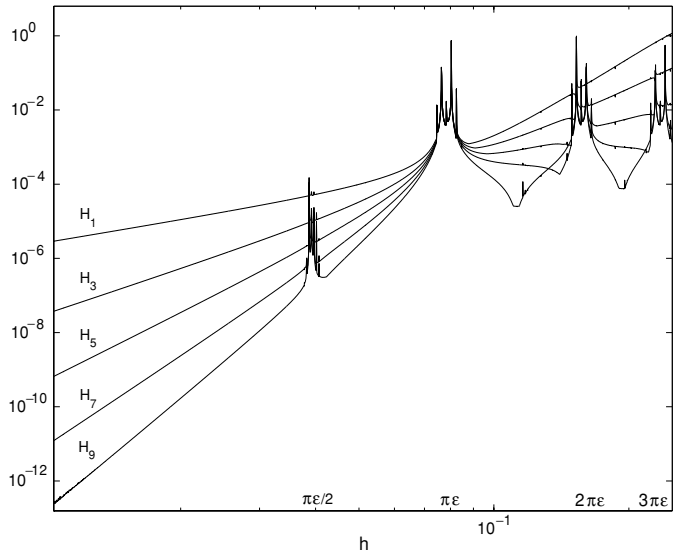

Figure 1. Preservation of modified energies to order 2, 4, 6, 8, 10 up to time $t = 100$. Vertical axis is $|(H_j - H_j^0)/H_j^0|$.

Figure 1 shows a convergence of preserved energy according to the order of the truncated modified Hamiltonians. There are, however, small peaks corresponding to step size resonances. Theorem 5 show that for this problem there exists a modified Hamiltonian of the form $\overline{H}(p, q, t; h) = H(p, q) + \epsilon R_1(p, q) + \epsilon R_2(p, q, t; h)$, for which ϵR_2 is exponentially small when h is small. Thus the effect of resonances corresponding to small h will be exponentially small as well, and we can expect that if higher order modified Hamiltonians were computed smaller peaks would be revealed for $h < \epsilon\pi/2$. Whenever a frequency of the dynamics is an integer multiple of the period, h in t of R_2 numerical resonances may appear leading to a faster deterioration of energy preservation and accuracy in the simulation. A similar argument can be found from the results in section 2.3; see also [51].

Next we monitor the drift in the modified energy H_9 which is $\mathcal{O}(h^{10})$ accurate near a resonance close to $h = \epsilon\pi/2$. Figure 2 shows that the energy seems to grow up to some value where it remains bounded for almost all values of h in this neighbourhood. Only for step sizes in a neighbourhood (shrinking as time increases) of the peak do we observe a continued growth indicating that even for such small steps that resonances lead to a possible source of instability.

Figure 3 shows for a fixed step $h \approx 0.0391$ that the energy can grow rather fast near a resonant step size. We note that for systems of high dimensions (e.g. PDE discretizations) the occurrence of a numerical resonance becomes more likely [45], and thus the possibility for energy growth increases, eventually leading to approximations that are polluted by resonance effects unless there is some counteracting effect present such as convexity of H explained in example 5.

Example 4 (Non-preservation of exact energy)**.** A result by Ge [52] and Feng and Wang [8] is that symplectic methods cannot conserve the energy H if H has no other conserved quantities but H itself. This result follows quite easily from MEs, in fact we can say more. From theorem 5 [17] we know that the numerical trajectory produced by applying a symplectic

[17] Actually the construction (1.4) is sufficient here.

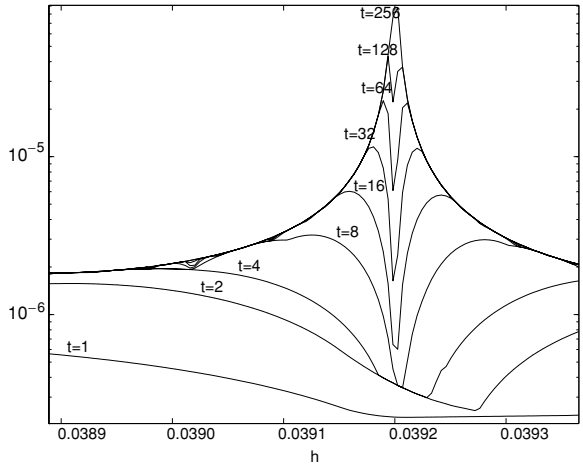

Figure 2. Drift in energy near $h = \pi\epsilon/2$. Vertical axis is $|(H_9^0 - H_9)/H_9^0|$.

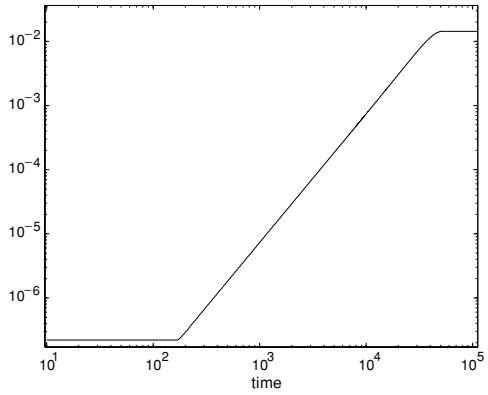

Figure 3. Drift in energy close to a resonance for a step size $h \approx \pi\epsilon/2$. Vertical axis is $|(H_9^0 - H_9)/H_9^0|$.

method to a Hamiltonian $H(p, q) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is given by the flow of a modified Hamiltonian $\overline{H}(p, q, t) = H(p, q) + \epsilon R(p, q, t; h)$ where ϵR is h-periodic in t. Along the numerical trajectory the Hamiltonian H varies as

$$\frac{\mathrm{d}}{\mathrm{d}t} H(p(t), q(q)) = \{H, \overline{H}\} = \{H, \epsilon R\}. \tag{4.1}$$

Situation 1. If, like Ge, we assume that the flow of H has no conserved quantities other than H then $\mathrm{d}/\mathrm{d}t\, H = 0$ only if $R = \alpha(H, t)$ where α is some smooth function. In this case the modified Hamiltonian vector field is $J\nabla\overline{H} = (1 + \partial_H \alpha(H, t))J\nabla H$, thus since H is constant ($=H(p(0), q(0)) = H^0$) along the trajectory, the numerical trajectory (generated by $\overline{H}$) is equal to the exact trajectory but at a time equal to $\int_0^{n\cdot h}(1 + \partial_H \alpha(H^0, s))\,\mathrm{d}s = n \cdot h + \int_0^h \partial_H \alpha(H^0, s)\,\mathrm{d}s \cdot n = (h + \overline{\alpha}) \cdot n$, since $\alpha(H^0, s)$ is periodic in s, i.e. $\Psi_{h,H}^n = \phi_{\overline{\alpha}\cdot n, H}$- the numerical trajectory is the exact solution up to a reparameterization of time.

Situation 2. If H in addition has $I_1, I_2, \ldots, I_j$, $j \leqslant d - 1$, as the only invariants, then (4.1) implies that $\epsilon R = \alpha(H, I_1, \ldots, I_j, t; h)$, and thus $\overline{H}$(i.e. the method) has $I_1, \ldots, I_j$ as invariants as well.

Using I_j as new variables (θ_j as their conjugates) in H and $\overline{H}$. Then $H = H'(\underline{p}, \underline{q}, I_1, \ldots, I_j), \overline{H} = \overline{H}'(\underline{p}, \underline{q}, I_1, \ldots, I_j, t)$. Considering the Hamiltonians on the reduced phase-space $(\underline{p}, \underline{q}) \in \mathbb{R}^{d-j} \times \mathbb{R}^{d-j}$ where H' has no other invariants than H' itself. Since $\overline{H}'$ must again Poisson-commute with H' we are back to *Situation 1*. Therefore the numerical approximation on the reduced space is a time-rescaling (by α' say) of the exact solution. The variables θ_j evolve at different rates for the exact solution and numerical approximation because of truncation errors, i.e. for the numerical solution we have

$$I_i(t) = I_i(0), \qquad i = 1, \ldots, j$$

$$\theta_i(t) = \theta_i(0) + \int_0^t \partial_{I_j} \overline{H}(\underline{p}(s), \underline{q}(s), I_1(0), \ldots, I_j(0), s)\, \mathrm{d}s, \qquad i = 1, \ldots, j$$

$$(\underline{p}, \underline{q})(t) = \phi_{\alpha' t, H'}(\underline{p}(0), \underline{q}(0)).$$

So in this case individual rescalings of the component of the numerical trajectory $\theta_i(t)$, $i = 1, \ldots, j$ and $(\underline{p}(t), \underline{q}(t))$ will again give the exact solution. Such numerical methods are however impossible in general since the method gives an exact solution of a Hamiltonian (H') without conserved quantities, while one knows that such formulae do not exist for non-integrable (i.e. without invariants) systems. See [49] for a discussion of the case when $j \geqslant d$ and [24] for a problem where exact energy preservation and symplecticity is possible, making higher order methods possible through time-rescaling of a low order method.

Example 5 (Preservation of other invariants)**.** Numerical simulations with close to integrable Hamiltonian systems $H = H_0(I) + \epsilon R_0(I, \theta), I \in \mathbb{R}^d, \theta \in \mathbb{T}^d$ have revealed that invariants are very well preserved. KAM theory has been applied to show that the variation in the I_j is bounded for *all time* [19, 25, 44] by assuming a strong-non resonance condition on $H_0(I)$;

$$|\partial_I H_0(I) \cdot m| \geqslant \gamma |m|_1^{-\tau}, \qquad \forall m \in \mathbb{Z}^d \backslash 0, \quad \gamma, \tau > 0$$

for $I = I^0$ given by the initial condition. Such non-resonance assumptions are impossible to check. The alternative *Nekhoroshev theory* [32] removes the non-resonance condition at the price of only giving boundedness of $|I_j(t) - I_j(0)|$ over exponentially long time-intervals, i.e.

$$\|I(t) - I(0)\| \leqslant \mathcal{O}(\epsilon^b), \qquad |t| \leqslant T = \mathcal{O}(\exp(-c/\|\epsilon R_0\|^a)) \tag{4.2}$$

where $c > 0, 0 < a, b < 1$. If ϵ is sufficiently small and H_0 satisfies one of the two convexity conditions

$$(i) \quad \left|\left\langle \partial_I^2 H_0(I) v, v \right\rangle\right| > m\|v\|_2^2, \quad m > 0 \quad \forall v \in \mathbb{R}^d$$

$$(ii) \quad \left|\left\langle \partial_I^2 H_0(I) v, v \right\rangle\right| > \hat{m}\|v\|_2^2, \quad \hat{m} > 0 \quad \forall v \in \mathbb{R}^d, v \perp \partial_I H^0$$

for I in some open neighbourhood of the trajectory, then one can show that $a = b = 1/2n$.

To establish Nekhoroshev stability of numerical approximations we consider $\overline{H}$ as given by theorem 5[18] and extend the phase-space to remove the time-dependency, i.e.

$$\overline{H} = \{H_0(I) + e\} + \{R_1(I, \theta) + R_2(I, \theta, \tau)\},$$

with e and τ conjugate variables, and R_1, R_2 are assumed small. If we assume that condition (ii) holds for H_0 then we can show [25] that $H_0 + e$ satisfies condition (i)(with H_0 replaced by $H_0 + e$ and $v \in \mathbb{R}^{d+1}$) with $m = \hat{m} \big/ \left(1 + h^2 \|\partial H_0(I)\|_2^2\right)$. Thus for the numerical approximation (4.2) holds with $a = b = 1/(2n + 2)$, a slight decrease in the stability time. It seems likely that such stability results are closer to the true explanation of conservation of invariants in symplectic discretizations than KAM theory provides[19].

[18] Smooth versions of Nekhoroshev theory exist, thus it is actually sufficient to use the construction (1.4).

[19] Note that e.g. RK methods based on Gauss–Legendre quadratures preserve linear and quadratic invariants [19] exactly, and this might affect the constants a, b in a beneficial way.

For discussion and analysis of adiabatic invariants see [46, 38], and [18] for a proof of a Virial theorem. Preservation of invariants is crucial to reducing global errors in numerical simulations. Using the almost preservation of invariants I as in example 5 linear bounds in $t = n \cdot h$ on error growth for symplectic integration schemes can be proved under general assumptions [25].

5. Conclusions

We have shown that the classical theory of MEs for discretizations of ODEs can be improved. In section 2 we presented an explicit formula for the traditional MVF, avoiding the traditional recursive schemes. In section 2.3 we pointed out some assumptions on f that can be used to improve the estimates, and related these to the inversion of $(\mathrm{d}\exp)_{hF}$.

In section 3 we provided an exponentially small analytic time-dependent perturbation, ϵr_2, which when added to the traditional MEs gives vector fields that exactly recover the numerical trajectories. This perturbation turns out to be important in explaining numerical resonance effects as in example 3 (see also [21]). In a sense ϵr_2 explains why we have the estimate of theorem 1. We believe that theorem 5 forms a basis for developing a more comprehensive understanding of the stability of numerical discretizations than the traditional estimate.

Acknowledgments

I would like to thank the anonymous referees for valuable comments, in particular for supplying me with the important reference [36].

References

[1] Ascher U M, Mattheij R M M and Russell R D 1988 *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations* (Englewood Cliffs, NJ: Prentice-Hall)
[2] Benettin G and Giorgilli A 1994 On the hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms *J. Stat. Phys.* **74** 1117–43
[3] Bialynicki-Birula I, Mielnik B and Plebanski J 1969 Explicit solution of the continuous Baker–Campbell–Hausdorff problem and a new expression for the phase operator *Ann. Phys.* **51** 187–200
[4] Calvo M P, Murua A and Sanz-Serna J M 1994 Modified equations for ODEs *Chaotic Numerics (Contemporary Mathematics* vol 172*)* (Providence, RI: American Mathematical Society) pp 63–74
[5] Culver W J 1966 On the existence and uniqueness of the real logarithm of a matrix *Proc. Am. Math. Soc.* **17** 1146–51
[6] Douady R 1982 Applications du théorème des tores invariantes *Thesis* Université Paris VII
[7] Douady R 1982 Une démonstration directe de l'équivalence des théoremes de tores invariants pour difféomorphismes et chapms de vectors *C. R. Seances Acad. Sci.* **295** 201–4
[8] Feng K and Wang 1991 A note on conservation laws of symplectic difference schemes for Hamiltonian systems *J. Comput. Math.*
[9] Feng K 1993 Symplectic, contact and volume preserving algorithms *Proc. 1st China–Japan Conf. on Numerical Math.* ed Z C Shi and T Ushijima (Singapore: World Scientific) pp 1–23
[10] Feng K 1991 The calculus of generating functions and the formal energy of Hamiltonian algorithms *Preprint* ASCC
[11] Gelfand I M, Krob D, Lascoux A, Leclerc B, Retakh V S and Thibon J-Y 1995 Noncommutative symmetric functions *Adv. Math.* **112** 218–348
[12] Gonzalez O, Higham D J and Stuart A M 1999 Qualitative properties of modified equations *IMA J. Numer. Anal.* **19** 169–90
[13] Hairer E 1994 Backward analysis of numerical integrators and symplectic methods *Ann. Numer. Math.* **1** 107–32
[14] Lichtenberg A J and Lieberman M A 1983 *Regular and Stochastic Motion (Applied Mathematical Sciences* vol 38*)* (New York: Springer)

[15] Hairer E and Lubich C 1997 The life-span of backward error analysis for numerical integrators *Numer. Math.* **76** 441–62
[16] Hairer E and Lubich C 2000 Asymptotic expansions and backward analysis for numerical integrators *Dynamics of Algorithms (IMA Vol. Math. Appl.* vol 118*)* (Minneapolis, MN: IMA) pp 91–106
[17] Hairer E and Lubich C 1999 Invariant tori of dissipatively perturbed Hamiltonian systems under symplectic discretization *Proc. NSF/CBMS Regional Conf. on Numerical Analysis of Hamiltonian Differential Equations (Golden, CO, 1997) Appl. Numer. Math.* **29** 57–71
[18] Hairer E, Lubich C and Wanner G 2003 Geometric numerical integration illustrated by the Störmer-Verlet method *Acta Numer.* 399–450
[19] Hairer E, Lubich C and Wanner G 2002 *Geometric Numerical Integration: Structure Preserving Algorithms for Ordinary Differential Equations (Springer Series in Computational Mathematics* vol 31*)* (Berlin: Springer)
[20] Kuksin S B and Pöschel J 1994 On the inclusion of analytic symplectic maps in analytic Hamiltonian flows and its applications *Seminar on Dynamical Systems (St Petersburg, 1991) (Progr. Nonlinear Differential Equations Appl.* vol 12*)* pp 96–116
[21] Leimkuhler B and Reich S 2004 *Simulating Hamiltonian Dynamics (Cambridge Monographs on Applied and Computational Mathematics* vol 14*)* (Cambridge: Cambridge University Press)
[22] Li M-C 2004 Qualitative property between flows and numerical methods *Nonlinear Anal.* **59** 771–87
[23] Magnus W 1954 On the exponential solution of differential equations for a linear operator *Commun. Pure Appl. Math.* **7** 649–73
[24] McLachlan R I and Zanna A 2005 The discrete moser-veselov algorithm for the free rigid body, revisited *Found. Comput. Math.* **5** 87–123
[25] Moan P C 2004 On the KAM and nekhoroshev theorems for symplectic integrators and implications for error growth *Nonlinearity* **17** 67–83
[26] Moan P C 2002 On backward error analysis and Nekhoroshev stability in the numerical analysis of conservative systems of ODEs *PhD Thesis* University of Cambridge
[27] Moan P C 2005 On rigorous modified equations for discretizations of ODEs, at press
[28] Moan P C and Niesen J On the convergence of the Magnus expansion, in preparation
[29] Moser J 1968 Lectures on hamiltonian systems *Mem. Am. Math. Soc.* **81** 1
[30] Moser J 1969 On a theorem Of anosov *J. Diff. Eqns* **5** 411–40
[31] Neishtadt A I 1984 The separation of motions in systems with rapidly rotating phase *J. Appl. Math. Mech.* **48** 133–9
[32] Nekhorošev N N 1977 An exponential estimate of the time of stability of nearly integrable hamiltonian systems *Russ. Math. Surv.* **32** 1–65
[33] Palis J 1974 Vector fields generate few diffeomorphisms *Bull. Am. Math. Soc.* **80** 503–5
[34] Palis J 1968 On Morse–Smale dynamical systems *Topology* **8** 385–404
[35] Perko L M 1968 Higher order averaging and related methods for perturbed periodic and quasi-periodic systems *SIAM J. Appl. Math.* **17**
[36] Pronin A V and Treschev D V 1997 On the inclusion of analytic maps into analytic flows *Regul. Khaoticheskaya Din.* **2** 14–24
[37] Reich S 1999 Backward error analysis for numerical integrators *SIAM J. Numer. Anal.* **36** 1549–70
[38] Reich S 1999 Preservation of adiabatic invariants under symplectic discretization *Appl. Numer. Math.* **29** 45–56
[39] Ruth R D 1983 A canonical integration technique *IEEE Trans. Nucl. Sci.* **NS-30** 2669–71
[40] Sanz-Serna J M 1995 Solving numerically Hamiltonian systems *Proc. Int. Congress of Mathematicians (Zürich, 1994)* vol 1 and 2 (Basle: Birkhauser) pp 1468–72
[41] Sanz-Serna J M and Calvo M P 1994 Numerical Hamiltonian problems *Applied Mathematics and Mathematical Computation* vol 7 (London: Chapman and Hall)
[42] Stoffer D 1998 On the qualitative behaviour of symplectic integrators: II. Integrable systems *J. Math. Anal. Appl.* **217** 501–20
[43] Stritchartz R S 1987 The Campbell–Baker–Hausdorff–Dynkin formula and solutions of differential equations *J. Funct. Anal.* **72** 320–45
[44] Shang Z-j 1999 KAM theorem of symplectic algorithms for hamiltonian systems *Numer. Math.* **83** 477–96
[45] Shang Z-j 2000 Resonant and diophantine step sizes in computing invariant tori of hamiltonian systems *Nonlinearity* **13** 299–308
[46] Shimada M and Yoshida H 1996 Long-term conservation of adiabatic invariants by using symplectic integrators *Publ. Astron. Soc. Japan* **58** 147–55
[47] Tang Y-F 1994 Formal energy of a symplectic scheme for hamiltonian systems and its applications *Comput. Math. Appl.* **27** 31–9
[48] Yoshida H 1990 Conserved quantities of symplectic integrators for Hamiltonian systems *Preprint* unpublished

[49] Yoshida H 2001 Non-existence of the modified first integral by symplectic integration methods *Phys. Lett.* A **282** 276–83

[50] Vinokurov V A 1992 Logarithm of the solution of a linear differential equations, the Hausdorff formula, and conservation laws *Sov. Math.—Dokl.* **44** 200–5

[51] Wisdom J and Holman M 1992 Symplectic maps for the n-body problem—stability analysis *Astron. J.* **104** 2022–9

[52] Marsden J and Ge Z 1988 Lie–Poisson Hamilton–Jacobi theory and Lie–Poisson integrators *Phys. Lett.* A **133** 134–9

Ge Z 1988 Geometry in symplectic difference schemes and generating functions *Preprint*