

# On rigorous modified equations for discretizations of ODEs

Per Christian Moan\*

July 4, 2005

## Abstract

We construct in a novel way the modified vector field used in backward error analysis of discretizations of ODEs. In addition to giving improvements of known results, the new result includes a small, analytic time-dependent perturbation so that the flow of the modified vector field exactly reproduces the numerical approximations produced by one-step methods. This simplifies analysis of resonance effects prevalent in conservative discretizations, and also paves the way to easy proofs of stability results such as the KAM theorem, which previously was proved by ad-hoc methods for numerical methods.

## 1 Introduction

We consider so-called one-step methods applied with constant step size  $h > 0^1$ , to a real analytic vector field  $F(y) : \mathcal{M} \subset \mathbb{R}^d \mapsto T\mathcal{M} \simeq \mathbb{R}^d$  i.e. approximations  $\Psi_{h,F}$  to the flow map(exact solution)  $\phi_{h,F} : \mathcal{M} \mapsto \mathcal{M}$ . The methods are assumed consistent  $\Psi_{h,F} - \phi_{h,F} = \mathcal{O}(h^r)$ ,  $r \geq 2$ , and that they respect  $\mathcal{M}$ ,  $\Psi_{h,F} : \mathcal{M} \mapsto \mathcal{M}$ . The *numerical trajectory* is then a sequence of vectors  $x_n \in \mathcal{M}$  given by iterating the numerical method

$$x_{n+1} = \Psi_{h,F}(x_n), \quad n = 0, 1, \dots, \quad x_0 = x(0).$$

We will make the assumption that the method  $\Psi_{h,F}(x)$  is analytic in  $x$  and  $h$ , and study the question of existence of a *modified vector field* such that its flow exactly interpolates the numerical trajectory.

Modified equations have since their inception [21] been applied in the analysis of discretizations of differential equations. In particular for discretization of ODEs one has through rigorous analysis of modified equations explained global effects such as energy-preservation of symplectic schemes, error growth, and parasitic effects of discretizations such as separatrix splitting and resonances[7, 16, 22] and many others. While mainly a technique for analyzing dynamics, it has also been used to improve considerably the efficiency of certain algorithms[8, 1]. Due to divergences occurring in the traditional constructions, the theory has up till now been of an asymptotic character, and the standard result can roughly be formulated as follows [2, 6, 16]

---

\*Centre of Mathematics for applications/University of Oslo. email:pcmoan@math.uio.no

<sup>1</sup>Runge-Kutta methods, splitting methods naturally belong to this class. Multi-step methods might be considered in this class by enlarging the phase-space, although in this case further complications occur.[7]

**Theorem 1.** *Let  $F$  be an analytic vector field in some open domain  $\mathcal{D} \subset \mathbb{C}^d$  around the trajectory with corresponding bound  $\|F\|_{\mathcal{D}}$ . Let  $x_1 = \Psi_{h,F}(x_0) = x_0 + hF(x_0) + \mathcal{O}(h^2)$  be an approximation produced by a one-step method. Then there exists a autonomous modified vector field  $\tilde{F}$ , bounded on the smaller domain  $\mathcal{D}' \subset \mathcal{D}$  so that*

$$\|x_1 - \phi_{h,\tilde{F}}(x_0)\| = \mathcal{O}(h \exp(-h_0/h\|F\|_{\mathcal{D}}))$$

for a sufficiently small step size  $h$ . Here  $h_0$  is a positive constant depending on the distance between  $\mathcal{D}$  and  $\mathcal{D}'$  and the method.

By iterating this bound one finds that the flow  $\phi_{n,h,\tilde{F}}$  stays exponentially close to  $x_{n+1} = \Psi_{h,F}(x_n)$  up to some finite time (which might be rather short for systems with exponentially diverging trajectories). In some applications such as molecular dynamics these can be too short to be of much use. Also when establishing proper stability via e.g. KAM theory[6] (i.e. for systems with only linearly diverging trajectories) the approach of going via a modified equation will considerably weaken the estimates.

In this report we show that by allowing *time-dependent modified vector fields* we avoid asymptotics, and still the traditional estimates for backward error analysis hold. Essentially we show that the numerical approximation can be viewed as the exact solution of a system of the form

$$y' = F(y) + R_1(y) + R_2(y, t) \in T\mathcal{M}.$$

In our analysis we put particular emphasis on the regularity in  $t$  of  $R_2$  as this turns out to be closely related to the estimate in Theorem 1. We have tried to keep the analysis focused on the problem of constructing this type of modified vector fields without letting particular structure of the methods enter.

## 1.1 Outline of the paper

In Section 2 we construct a smooth time-dependent modified vector field that exactly interpolates the numerical trajectory. We show that it is possible to make this Gevrey-smooth and  $h$ -periodic in time which can be viewed as a rough modified vector field. In Section 3 we take this modified vector field as a starting point, and construct a time-dependent coordinate transformation, so that in the new coordinates the time-dependent modified vector field is analytic in time. In the last section some applications of the result are presented. For readability technical proofs are collected in the appendix.

## 2 Smooth non-autonomous modified vector fields

Let  $x_0 \in \mathcal{M}$  be an arbitrary point in phase space, and consider the set of curves,  $\mathcal{C}$ , parameterized as  $x_1(\tau) \in \mathcal{M}$ ,  $\tau \in [0, h]$  smoothly connecting  $x_0$  with  $x_1 = \Psi_{h,F}(x_0)$ . We note that differentiating an element of  $\mathcal{C}$  by  $\tau$  we get a tangent vector. We start our discussion by simply letting  $x_1(\tau) = \Psi_{\tau,F}(x_0) \in \mathcal{C}$ . Differentiating  $x_1(\tau)$  with respect to  $\tau$  we find that the curve satisfies the differential equation

$$\begin{aligned} \frac{d}{d\tau} x_1 &= \frac{d}{d\tau} \Psi_{\tau,F} \circ x_0 = (\Psi_{\tau,F})_{\tau} \circ \Psi_{\tau,F}^{-1}(x_1) \\ (2.1) \quad &= F(x_1) + R(x_1, \tau) =: \tilde{F}(x_1, \tau) \in T\mathcal{M}, \quad x_1(0) = x_0 \end{aligned}$$

by assuming invertibility and consistency of the method. The vector field  $R = \tau^{r-1}E_r(x_1) + \mathcal{O}(\tau^r)$  is a non-autonomous perturbation with  $\tau^r E_r$  the local truncation error of the scheme.<sup>2</sup> A recursive scheme for expressing  $R$  for B-series methods can be found in [3] where the emphasis was on symplectic methods. We note that for all one-step methods the functional inverse  $\Psi_{h,F}^{-1}$  is well defined for sufficiently small step. Invertibility is a natural assumption since flow maps are invertible and a numerical method that is not invertible can therefore not be represented as a flow map.

Since the point  $x_0 \in \mathcal{M}$  was arbitrary the same construction holds for all steps  $n = 0, 1, \dots$  with the modification that we periodically extend  $\tilde{F}$  to the whole real line  $\tau \in \mathbb{R}$ .

$$\frac{d}{d\tau}\tilde{x} = \tilde{F}(\tilde{x}, \tau), \quad \tilde{x}(0) = x_0, \quad \tilde{x}(n \cdot h) = x_n.$$

In order to remove the discontinuity at  $\tau = n \cdot h$  a non-linear scaling of “time”  $\tau = \chi(t/h)$  is introduced, where  $\chi$  is smooth 1-periodic function such that

$$\begin{aligned} \chi(0) &= 0 & \chi(1) &= 1, \\ \chi^{(k)}(0, 1) &= 0, & k &= 1, 2, \dots \end{aligned}$$

Then it is not difficult to see that we can repeat the above argument with the time-scaled curve  $\mathcal{C}_\chi$ , given by  $\Psi_{h\chi(t/h), F}$  and obtain a modified equation similar to (2.1) which is smooth in  $t$ . Unfortunately the time-scaling introduces a time-dependency in the “unperturbed part”. To avoid this we instead introduce parametrization

$$\Phi_{t,F} = \phi_{t-h\chi(t/h), F} \circ \Psi_{h\chi(t/h), F} \in \mathcal{C}$$

which is equivalent to the numerical method when iterated with time step  $t = h$ . The modified vector field now becomes

$$\begin{aligned} \left(\frac{d}{dt}\Phi_t\right) \circ \Phi_t^{-1} &= \tilde{F}_\chi = F + R_\chi, \\ R_\chi &= \chi'(t/h)(\phi_{t-h\chi(t/h), F})_* R(y, h\chi(t/h)) \\ (2.2) \quad &= \chi'(t/h) \sum_{j=0, k=r}^{\infty} (t - h\chi(t/h))^j \chi(t/h)^k R_{j,k} \end{aligned}$$

where  $R$  is defined as in (2.1),  $\phi_* G = d\phi \cdot G \circ \phi^{-1}$ . The coefficients  $R_{j,k}$  consists of higher derivatives of  $F$  and their products. By periodically extending  $R_\chi$  for all  $t$  we have the desired modified vector field. By construction it follows that  $R_\chi = \mathcal{O}(h^{r-1})$ ,  $r$  being the order of the method. See [5] for a similar, but less geometrical idea.

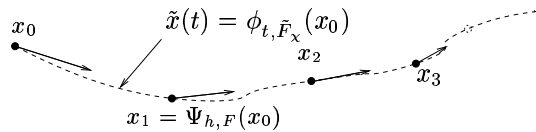


Figure 1: The smooth interpolating trajectory.

<sup>2</sup>We make the assumption that  $R$  has a convergent Taylor expansion around  $\tau = 0$

**Note 1.** (Structural properties of the vector fields) Due to algebraic constraints etc. the manifold  $\mathcal{M}$  need not be  $\mathbb{R}^d$ , but rather be defined as the level-sets or intersections of a set of functions on  $\mathcal{M}$ . Several numerical methods have been constructed preserving such constraints, we assume that  $\tilde{F} \in T\mathcal{M}$  for such methods. Phases-spaces may well be endowed with more structure than that of just  $\mathcal{M}$ . E.g. if  $\mathcal{M}$  is even dimensional we may define a symplectic two-form on  $T^*\mathcal{M}$ -the cotangent space, and using this two form define a Hamiltonian vector field in  $T\mathcal{M}$ . Or if  $T^*\mathcal{M}$  is endowed with a volume form, we may define divergence free vector fields in  $T\mathcal{M}$ . Both these two cases represent vector fields that form Lie-sub algebras, and correspondingly their flows form Lie sub-groups. There are numerical methods that preserve these differential forms (such as operator splitting or some Runge-Kutta methods for Hamiltonian vector fields) [7, 11], and hence produce elements in the Lie-groups. Since time-scalings respect group operations,  $\Phi_{t,F}$  will be in the subgroup of diffeomorphisms generated by the one-parameter elements  $\phi$  and  $\Psi$ . The implication of this is that the tangent-vector field will reside in the corresponding infinite dimensional Lie algebra of smooth vector fields [16]. In particular in the Hamiltonian case the modified vector field  $\tilde{F}_\chi$  is also Hamiltonian. A third, important type, of structure on  $\mathcal{M}$  are time-reversing symmetries. A vector field  $F$  on  $\mathcal{M}$  is said to have a time-reversing symmetry if there exists a mapping  $I$  so that  $I_*F(y) = -F(y)$ . It is known that e.g. symmetric Runge-Kutta methods preserve linear time-reversing symmetries. In order for our time-dependent modified vector field to retain such symmetries one can replace  $\Phi_t$  by

$$\Phi_{t,F} = \phi_{t/2-h/2\chi(t/h)} \circ \Psi_{h\chi(t/h)} \circ \phi_{t/2-h/2\chi(t/h)},$$

with a similar expression for  $R_\chi$  for which  $I_*R_\chi(y, t) = -R_\chi(y, -t)$ .

## 2.1 Gevrey regularity

The exponentially small estimates associated with modified vector fields is closely related to the smoothness of  $\tilde{F}$ . To make  $\tilde{F}$  as smooth as possible in  $t$  we now make a choice of  $\chi \in C^\infty(\mathbb{R})$ . We say a function  $\chi$  is *Gevrey- $\gamma$*  [9] provided

$$(2.3) \quad \|\chi^{(k)}\|_\infty \leq M c^k k^{\gamma k}$$

holds for some constants  $M, c > 0$ . We denote the set of functions  $h$ -periodic and Gevrey- $\gamma$  in  $t$  by  $\Gamma_{h,\gamma}$ . Gevrey- $\gamma = 1$  functions are the analytic functions. We achieve high smoothness by choosing

$$(2.4) \quad \chi_\eta(t) = \frac{1}{2} - \frac{1}{2} \tanh\left(\frac{\cos^\eta(\pi t)}{\sin^\eta(\pi t)}\right),$$

$\eta = 1, 3, 5, \dots$ , with the properties that it is periodic,  $\chi_\eta(0) = 0$ ,  $\chi_\eta(1) = 1$ ,  $\chi_\eta^{(k)}(1, 0) = 0$ ,  $k > 0$ . It turns out that the first derivative of  $\chi$  needed in  $\tilde{R}_\chi$  grows as  $\mathcal{O}(\eta)$  so that there is a limit to how smooth we can make the time-dependency without making  $\|R_\chi\|$  too large.

**Proposition 1.**  $\chi_\eta(t/h) \in \Gamma_{h,\gamma=1+1/\eta}$ .

It is well known that Gevrey functions form an algebra, stable under differentiation and composition. Applying the same proof technique as in the proof of Proposition 1 one can show that indeed the coefficients  $\chi'(t/h)(t - h\chi(t/h))^j (h\chi(t/h))^k$  of the Taylor expansion of  $R_\chi$  (2.2) are also Gevrey- $\gamma$ , thus  $R_\chi$  is in the same class as  $\chi_\eta$ .

**Corollary 1.** *Let  $\Psi_{h,F}$  be a one-step method. Then there exists a vector field  $\hat{F}(y, t) = F(y) + \hat{R}(y, t; h)$  in  $\Gamma_{h,\gamma>1}$ , so that the solution of*

$$x' = \hat{F}(x, t), \quad x(0) = x_0$$

*exactly interpolates the numerical trajectory,  $x(n \cdot h) = x_n$ .*

**Lemma 1.** *The Fourier coefficients of  $R(t) \in \Gamma_{h,\gamma}$  are bounded as*

$$|R^k| \leq M e \exp\left(-\frac{\gamma}{e} \sqrt[3]{\frac{2\pi|k|}{ch}}\right),$$

*where  $M, c$  are the constants appearing in (2.3).*

The following result then follows immediately by splitting out the constant term  $R_1$  of  $\hat{R}$  in Corollary 1.

**Corollary 2.** *Let  $\Psi_{h,F}$  be some one-step method. Then there exists a non-autonomous vector field  $\hat{F}$  in  $\Gamma_{h,\gamma>1}$ ,*

$$(2.5) \quad \hat{F}(y, t) = F(y) + R_\chi(y, t) = F(y) + R_1(y) + R_2(y, t)$$

*whose flow exactly interpolates the numerical trajectory. Furthermore the Fourier coefficients of the non-autonomous perturbation behaves for small  $h$  as*

$$\|R_2^k\| \simeq \mathcal{O}\left((1-\gamma)^{-1} \exp\left(-\frac{\gamma}{e} \sqrt[3]{\frac{2\pi|k|}{ch}}\right)\right)$$

*since  $\chi' = \mathcal{O}(\eta) = \mathcal{O}(1/(1-\gamma))$ .*

**Note 2.** *For the important class of close to integrable Hamiltonian vector fields  $F$  discretized by symplectic numerical methods, the above result can be used together with smooth versions of the KAM[14] and Nekhoroshev theory[17] to prove results regarding preservation of invariants and through that linear bounds on the error growth by the same argument as in [12].*

Numerical experiments show that the bound in Corollary 2 is too pessimistic[2, 13], and the main result of this article is the following

**Theorem 2.** *Let  $\Psi_{h,F}$  be a one-step method, and assume that  $F(y)$  and  $R_\chi(y, t)$  as given by (2.5) are analytic for  $y \in \mathcal{D} \subset \mathbb{C}^d$ ,  $\mathcal{D}$  containing the trajectory  $x(t)$ . Then there exists an analytic coordinate transformation  $\Xi_t : \mathcal{M} \mapsto \mathcal{N}$  and a modified vector field  $\bar{F} = F(y) + \bar{R}(y, t) \in T\mathcal{N}$ , analytic in  $\mathcal{D}' \subset \mathcal{D}$ ,  $h$ -periodic and analytic in  $t$ , whose flow exactly interpolates the numerical trajectory after transforming the trajectory by  $\Xi_t$ . I.e.*

$$\Xi_t \circ \phi_{t,\bar{F}} \circ \Xi_t^{-1}(x_0) = x_n = \Psi_{h,F}^n(x_0), \quad \text{for } t = n \cdot h.$$

*The domain of convergence of  $\bar{F}$  for the time-variable is  $\{t \in \mathbb{C} : |\Im(t)| < \xi\delta/\rho e\}$ , where  $\rho, \delta$  and  $\xi > 0$  are given in the following section.*

Thus by allowing a coordinate transformation we have shown that the numerical trajectory can be represented as the flow of a time-dependent vector field,  $h$ -periodic and analytic in  $t$ .

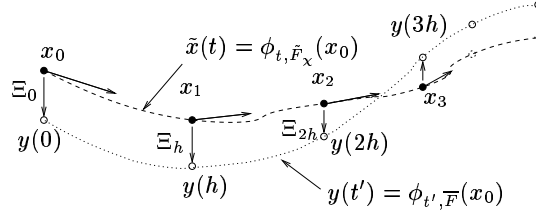


Figure 2: The interpolating flow in transformed coordinates.

### 3 A transform scheme for non-autonomous systems

We want to find a coordinate transform  $\Xi_t : \mathcal{M} \ni x \mapsto y \in \mathcal{N}$  so that in the new coordinates (2.5) is  $h$ -periodic and *analytic* in  $t$ , and we will do this by an iterative scheme. We start by making the assumption that  $\Xi_{t,s}$  is the solution of the differential equation

$$(3.1) \quad \frac{d}{ds}y = W(y, t, s), \quad y(t, s = 0) = x(t),$$

where the vector field  $W$  is to be chosen, and thus the coordinate transformation  $\Xi_{t,s} = \phi_{s,W}$ . The variable  $s \in \mathbb{R}$  is a transform parameter. The transformed variable  $y(t, s) = \Xi_{t,s}(x(t))$  then satisfies a differential equation

$$(3.2) \quad y' = \bar{F}(y, t, s) \in TN, \quad y(t = 0, s) \in \mathcal{N} \text{ given,}$$

where  $\bar{F} = (\Xi_t)_* \hat{F}$ . Differentiating (3.1) and (3.2) with respect to  $t$  and  $s$  respectively leads to

$$(3.3) \quad \begin{aligned} \frac{\partial}{\partial t}W + d_y W \cdot y' &= \frac{\partial}{\partial s}\bar{F} + d_y \bar{F} \frac{d}{ds}y \\ &\Downarrow \\ \frac{\partial}{\partial s}\bar{F}(y, t, s) &= \frac{\partial}{\partial t}W(y, t, s) + [\bar{F}, W](y, t, s), \quad \bar{F}(y, t, s = 0) = F(y) + R(y, t), \end{aligned}$$

where the Lie-Jacobi bracket  $[\bar{F}, W](y, t, s) = d_y W(y, t, s)\bar{F}(y, t, s) - d_y \bar{F}(y, t, s)W(y, t, s)$ .

<sup>3</sup> Introducing Fourier series for  $\bar{F}$  and  $W$ ;

$$\bar{F}(y, t, s) = \sum_{k \in \mathbb{Z}} \bar{F}^k(y, s) \exp\left(\frac{2\pi i k t}{h}\right), \quad W(y, t, s) = \sum_{k \in \mathbb{Z}} W^k(y, s) \exp\left(\frac{2\pi i k t}{h}\right),$$

gives from (3.3) the system of equations

$$(3.4) \quad \frac{\partial}{\partial s}\bar{F}^k = \frac{2\pi i k}{h}W^k + \sum_{p+q=k} [\bar{F}^p, W^q], \quad \forall k \in \mathbb{Z}.$$

Since analyticity in time of  $\bar{F}$  is equivalent to the Fourier coefficients satisfying a bound  $|\bar{F}^k| \leq M \exp(-\tilde{\delta}|k|)$  we choose  $W^k = i \text{sign}(k)\bar{F}^{k-4}$  and have the PDEs

$$(3.5) \quad \frac{\partial}{\partial s}\bar{F}^k = -\frac{2\pi|k|}{h}\bar{F}^k + \sum_{p+q=k} i \text{sign}(q)[\bar{F}^p, \bar{F}^q], \quad \forall k \in \mathbb{Z},$$

<sup>3</sup> $d_y F$  denotes the Jacobian of the vector field  $F$  with respect to  $y$ .

<sup>4</sup> $\text{sign}(k)$  is the sign function,  $\text{sign}(0) = 0$ .

with initial conditions

$$\overline{F}^k(y, s = 0) = \hat{F}^k(y) = \begin{cases} F(y) + R^0(y) & k = 0 \\ R^k(y) & k \neq 0 \end{cases}$$

The motivation behind the choice of  $W^k$  can be seen if we neglect the non-linearity in (3.5), then the solutions  $\overline{F}^k$  clearly represent an analytic function. The equation (3.5) evolves with  $s$  a non-analytic initial value  $\hat{F}$  into an analytic one  $\overline{F}$ , and the larger  $s$  is, the larger radius of convergence in  $t \in \mathbb{R}$   $\overline{F}(y, t, s)$  has. To take account of the non-linearity for rigorous estimates we let

$$(3.6) \quad \overline{F}^k = \exp\left(-\frac{2\pi|k|s}{h}\right) G^k$$

and inserting into (3.5) we arrive at the following system of equations

$$(3.7) \quad \begin{aligned} \frac{d}{ds} G^0 &= -2i \sum_{l=1}^{\infty} [G^l, G^{-l}] \exp(-4\pi ls/h) \\ \frac{d}{ds} G^{\pm k} &= \pm i [G^0, G^{\pm k}] \pm 2i \sum_{l=1}^{\infty} [G^{\mp l}, G^{\pm k \pm l}] \exp(-4\pi ls/h). \end{aligned}$$

An iterative scheme is then found by applying Picard iterations to (3.7), giving for  $G^k = G_0^k + G_1^k + \dots$ , the following recursive scheme;  $n > 0$ :

$$(3.8) \quad \begin{aligned} G_n^0(y, s) &= -2i \sum_{p+q=n-1} \sum_{l \geq 1} \int_0^s [G_p^l(\sigma), G_q^{-l}(\sigma)](y) \exp(-4\pi l\sigma/h) d\sigma \\ G_n^{\pm k}(y, s) &= \pm i \sum_{p+q=n-1} \int_0^s [G_p^0(\sigma), G_q^{\pm k}(\sigma)](y) d\sigma \\ &\quad \pm 2i \sum_{p+q=n-1} \sum_{l \geq 1} \int_0^s [G_p^{\mp l}(\sigma), G_q^{\pm k \pm l}(\sigma)](y) \exp(-4\pi l\sigma/h) d\sigma \end{aligned}$$

where  $G_0^0(y, s) = F(y) + R^0(y)$ ,  $G_0^k(y, s) = R^k(y)$  for  $k \neq 0$  while  $G_n^k(y, s = 0) = 0$ ,  $n \geq 1$ . Thus in order to prove analyticity in  $t$  we need to show that  $G^k(y, s)$  is bounded for some  $s > 0$ . Indeed our aim in the next section is to find the largest  $s$  which guarantees this.

**Note 3.** *The transformation, being represented by Lie brackets leaves the Lie algebra generated by  $\tilde{F}$  invariant. By considering the Fourier coefficients of the vector field, one can also show that time-reversing symmetries are retained by  $\overline{F}$ . This implies that the structural properties of the method  $\Psi_{h,F}$  are properly retained in  $\overline{F}$ . Thus structural properties such as preservation of integrals, Lie group and reversing symmetries of the numerical algorithm are reflected in  $\overline{F}$  as in the standard modified vector fields[16].*

### 3.1 Rigorous estimates of the transformation

Let  $|z|_{\infty} := \max_{j=1,2,\dots,n} |z_j|$  the max-norm. Let  $\mathcal{D}(x) \subset \mathbb{C}^n$ ,  $i = 0, 1, \dots$  be an open tubular neighborhood of the trajectory  $x(t)$ ,  $t \in \mathbb{R}$  of (2.5) and  $\partial\mathcal{D}(x)$  its boundary

while  $\text{dist}_{\mathcal{D}}(z) := \inf_{w \in \partial \mathcal{D}} |w - z|_{\infty}$ . We found the Nagumo-norm [20] useful in our estimates;

$$(3.9) \quad \|F\|_l := \max_j \sup_{z \in \mathcal{D}(x)} |F_j(z)| \text{dist}_{\mathcal{D}}(z)^l, \quad l = 1, 2, \dots$$

where  $F_j$  denotes the components of the vector field<sup>5</sup>. The following lemma is then proved by the Cauchy integral formula

**Lemma 2.** *Let the vector fields  $F$  and  $G$  be analytic in  $\mathcal{D}(x)$  then*

$$\|dGF\|_n \leq (m+1) \left(1 + \frac{1}{m}\right)^m \|G\|_m \|F\|_{n-m-1}.$$

for  $m = 0, 1, \dots, n-1$  with  $(1 + 1/m)^m = 1$  when  $m = 0$ .

Clearly the Lie–Jacobi bracket of two real analytic vector fields is real analytic, moreover we have the bound

**Corollary 3.** *Let the vector fields  $F$  and  $G$  be analytic in  $\mathcal{D}(x)$  then*

$$\|[F, G]\|_{n+m+1} \leq C_{m,n} \|F\|_n \|G\|_m,$$

where  $C_{m,n} = (1+n)(1+1/n)^n + (1+m)(1+1/m)^m$ .

For convergence analysis we apply weighted Fourier norms:

$$(3.10) \quad \begin{aligned} \rho_n^+(s) &:= \|G_n\|_n^* = \sum_{k \neq 0} w(k, h) \|G_n^k(s)\|_n, \\ \rho_n^0(s) &:= \|G_n^0(s)\|_n, \end{aligned}$$

where  $w(k, h) = \exp(c' \sqrt[3]{\frac{|k|}{h}})(1 + |k|)^2$  for some constant  $c' > 0$ . By Lemma 1  $\rho_0^+, \rho_0^0$  are bounded for a sufficiently small constant  $c' > 0$ .

**Lemma 3.** *For  $n = 1, 2, \dots$  the functions  $\rho_n^+(s), \rho_n^0(s)$ , satisfy the following inequalities*

$$\begin{aligned} \rho_n^0(s) &\leq \epsilon^2/10 \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^+(\sigma) \rho_q^+(\sigma) d\sigma \\ \rho_n^+(s) &\leq \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^0(\sigma) \rho_q^+(\sigma) d\sigma + \epsilon \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^+(\sigma) \rho_q^+(\sigma) d\sigma, \end{aligned}$$

with  $\epsilon := 2/3 \exp(-c' \sqrt[3]{1/h}) < 2/3$ .

**Lemma 4.** *Let  $\rho := \max\{\|F + R^0\|_0, \|R\|_0^*\}$ <sup>6</sup>, then  $\rho_n^0, \rho_n^+$  are bounded as*

$$\rho_n^+(s) \leq \rho \left(\frac{\rho es}{\xi}\right)^n \tilde{\xi}, \quad \rho_n^0(s) \leq \rho \left(\frac{\rho es}{\xi}\right)^n \left(1 + \frac{\epsilon \xi \tilde{\xi}^2}{3}\right),$$

where  $\xi(\epsilon)$  and  $\tilde{\xi}(\epsilon)$  are given in the proof. We note that  $\xi(\epsilon)$  converges very fast to 1 as  $h \rightarrow 0$ .

<sup>5</sup> $\|F\|_0$  is the usual sup-norm often applied in rigorous perturbation theory.

<sup>6</sup>We note that for reasonable numerical methods (and choice of  $c'$  in  $w(k, h)$ )  $\|R\|_0^* \leq \|F + R^0\|_0$  so we may assume that  $\rho = \|F + R^0\|_0$ .

The emphasis in Lemma 4 was on finding a sharp upper bound on  $s$  so that  $\sum_{n \geq 0} \rho_n^0$ ,  $\sum_{n \geq 0} \rho_n^+$  are bounded, and this is the case provided  $s < \frac{\xi}{\rho e}$ . If sharper bounds on the individual  $\rho_n s$  are sought, iteratively solving the inequalities in Lemma 3 with the corresponding initial conditions will give a significant improvement, although the convergence bound on  $s$  is largely unaltered by this.

We are now in position to derive bounds on the Fourier coefficients,  $G^k$ . If  $\mathcal{D}' \subset \mathcal{D}$  and  $\|F\|'_l$  denotes the Nagumo-norm with respect to  $\mathcal{D}'$  it follows from the definition that

$$(3.11) \quad \|F\|'_0 \leq \frac{1}{\delta^l} \|F\|_l, \quad \delta = \inf_{w \in \mathcal{D}'} \text{dist}_{\mathcal{D}}(w).$$

To avoid that  $\mathcal{D}'$  is empty we assume that  $\delta < \text{dist}_{\mathcal{D}}(\mathbb{R}(\mathcal{D}))$ .<sup>7</sup>

**Proposition 2.** *Assume that  $s < \frac{\xi \delta}{\epsilon \rho}$ , then on the smaller domain  $\mathcal{D}'$ , we have the bounds*

$$(3.12) \quad \begin{aligned} |k| \neq 0: \quad \|G^k\|'_0 &\leq \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \rho \tilde{\xi}(\epsilon) \frac{z}{1-z} \Big|_{z=\rho \epsilon s / \xi \delta} \right\} \\ \|G^0\|'_0 &\leq \rho_0^0 + \rho \left( 1 + \frac{\epsilon \xi \tilde{\xi}^2}{3} \right) \frac{z}{1-z} \Big|_{z=\rho \epsilon s / \xi \delta} \end{aligned}$$

*Proof of Theorem 2.* By Prop. 2 we choose  $0 < s < \xi \delta / \rho e$ , recalling (3.6) we have the modified equation

$$\overline{F}(y, t, s) = \sum_{k \in \mathbb{Z}} G^k(y, s) \exp\left(-\frac{2\pi s |k|}{h}\right) \exp\left(\frac{2\pi i k t}{h}\right).$$

By construction the vector fields  $G^k$  are analytic with the bounds given in Prop 2. For  $\overline{F}$  we have the bound

$$\begin{aligned} \|\overline{F}(\cdot, t)\|'_0 &\leq \rho_0^0 + \rho \left( 1 + \frac{\epsilon \xi \tilde{\xi}^2}{3} \right) \frac{z}{1-z} \Big|_{z=\rho \epsilon s / \xi \delta} \\ &+ \left\{ \rho_0^+ + \rho \tilde{\xi}(\epsilon) \frac{z}{1-z} \Big|_{z=\rho \epsilon s / \xi \delta} \right\} \sum_{|k| \neq 0} \frac{\left| \exp\left(-\frac{2\pi(s|k| - itk)}{h}\right) \right|}{2w(k, h)} \\ &\leq \rho_0^0 + \rho \left( 1 + \frac{\epsilon \xi \tilde{\xi}^2}{3} \right) \frac{z}{1-z} \Big|_{z=\rho \epsilon s / \xi \delta} \\ &+ \frac{3\epsilon}{2} \left\{ \rho_0^+ + \rho \tilde{\xi}(\epsilon) \frac{z}{1-z} \Big|_{z=\rho \epsilon s / \xi \delta} \right\} \sum_{|k| \neq 0} \frac{\left| \exp\left(-\frac{2\pi(s|k| - itk)}{h}\right) \right|}{2(1 + |k|)^2}. \end{aligned}$$

since  $w(k, h) \geq \frac{2}{3\epsilon}(1 + |k|)^2$ . The sum converges for  $|\Im(t)| \leq s$ , and a simple calculation shows that it is bounded for  $\epsilon \in [0, 2/3]$ . By the relation  $W^k = i \text{sign}(k) \overline{F}^k$  the vector field  $W$  is both real and analytic in  $t$  with the same domain of convergence as  $\overline{F}$ .  $\square$

<sup>7</sup> $\text{dist}_{\mathcal{D}}(\mathbb{R}(\mathcal{D}))$  denotes the distance from  $\partial \mathcal{D} \subset \mathbb{C}^d$  to the real subset of  $\mathcal{D}$ .

## 4 Connections with the traditional estimate

Although the Fourier coefficients  $\overline{F}^k$  decrease geometrically, we have no guarantee that the time-dependent part of  $\overline{F}$  is small, as  $G^k$  might be correspondingly large for small  $k$ . The bound on  $G^k$  blows up for  $s \geq \xi\delta/e\rho$  while  $\exp(-\frac{2\pi|k|s}{h})$  decreases for increasing  $s > 0$  thus there is an optimal value of  $s \in (0, \xi\delta/e\rho)$  minimizing our bounds on the Fourier-term,  $\overline{F}^{k \neq 0}$ .  $\arg \min_s \frac{\rho e s / \xi \delta}{1 - \rho e s / \xi \delta} \exp(-2\pi|k|s/h) \approx \arg \min_s \frac{1}{1 - \rho e s / \xi \delta} \exp(-2\pi|k|s/h) = \frac{\xi\delta}{\rho e} - \frac{h}{\pi|k|}$ . If we require that the leading term  $\overline{F}^{\pm 1}$  of the time-dependent part is small we get by the requirement  $s > 0$  a step size restriction

$$(4.1) \quad h < \frac{\pi\xi\delta}{e\rho}, \quad \text{when } s = \frac{\xi\delta}{\rho e} - \frac{h}{\pi}.$$

By these considerations we recover a refinement of the traditional estimates for modified vector fields

**Theorem 3.** *Let  $\Psi_{h,F}$  be a one-step method. Suppose the step size satisfies the bound  $h < \pi\xi\delta/e\rho$  then there exists a real analytic coordinate transformation  $\Xi_t$ , and modified vector field  $\overline{F}(y, t) = F(y) + R_1(y) + R_2(y, t)$   $h$ -periodic in  $t$  whose flow exactly interpolates the numerical trajectory up to the coordinate transformation. The non-autonomous perturbation satisfies the bound ( $t \in \mathbb{R}$ ):*

$$\|R_2(\cdot, t)\|'_0 \leq 8\epsilon \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{eh} \right\} \exp\left(-\frac{2\pi\xi\delta}{\rho eh}\right)$$

where  $\rho_0^+ = \|R\|'_0$ ,  $\rho = \|F + R^0\|'_0$ . In the limit  $h \rightarrow 0$  we may insert  $\xi = 1$  and  $\rho = \|F\|'_0$ .

Compared to the traditional estimate of Theorem 1 this result contains a few improvements. Firstly it does give a vector field interpolating the numerical trajectory for all time. Second the bound on the time-dependent part is sharper than the exponentially small discrepancy would indicate in previous results[6, 16, 13] in addition to relaxing the step size requirement slightly.

### 4.1 Eliminating the coordinate transformation

In our construction we showed the existence of  $\overline{F}$  by carrying out a close to identity transformation  $\Xi_t$ . To show that there exist an analytic vector field,  $\overline{F}^* \in T\mathcal{M}$  whose flow interpolates  $x_n$  at  $t = n \cdot h$  it is sufficient to compute the pull-back  $(\Xi_0)^* : T\mathcal{N} \mapsto T\mathcal{M}$ , i.e. the transform given by the  $s$ -flow of  $W$  frozen at  $t = 0$ ;

$$\overline{F}^* = (\Xi_0)^*\overline{F} = (d\phi_{s,W(t=0,s)})^{-1} \cdot \overline{F} \circ \phi_{s,W(t=0,s)}$$

Differentiating  $\overline{F}^*$  with respect to  $s$  we find (in a similar way to (3.3))

$$(4.2) \quad \frac{d}{d\sigma} \overline{F}^* = [W(\sigma, t=0), \overline{F}^*], \quad \overline{F}^*(\sigma=0) = \overline{F}$$

which is to be solved for  $\overline{F}^*(\sigma = s)$ , where  $s$  is given by Prop. 2.

**Theorem 4.** Assume  $h < \pi\xi\delta/\rho$  and let  $\mathcal{D}'' \subset \mathcal{D}' \subset \mathbb{C}^d$  be a non-empty domain with  $\delta' = \inf_{z \in \mathcal{D}''} \text{dist}_{\mathcal{D}'}(z)$  such that  $\rho h < 27\delta'/e$ . Then there exists a vector field  $\overline{F}^*(y, t) \in TM$  analytic in  $y \in \mathcal{D}'' \subset \mathcal{D}'$ ,  $t \in \{t \in \mathbb{C} : |\Im(t)| \leq \xi\delta/\rho e\}$  and  $h$ -periodic in  $t$  so that its flow exactly interpolates the numerical trajectory

$$x^*(t) = \phi_{t, \overline{F}^*}(x_0) = x_n = \Psi_{h, F}^n(x_0), \quad \text{for } t = n \cdot h.$$

The time-dependent part of  $\overline{F}^*$  is exponentially small.

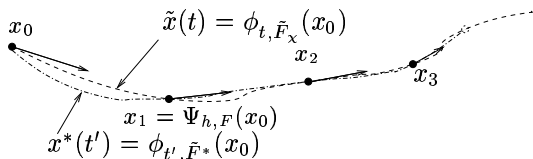


Figure 3: The analytic interpolating flow.

Theorem 4 then implies Theorem 1 by the non-linear variation of constants formula. It seems possible to remove the coordinate transformation also in Theorem 2 but we will not pursue this possibility here.

## 5 Applications to symplectic integration schemes

### 5.1 Energy preservation of symplectic schemes

If a symplectic scheme is applied to a Hamiltonian system with Hamiltonian  $H(p, q)$  it follows from Theorem 3 and Note 3 that there exists a modified Hamiltonian  $\overline{H}$  and a symplectic coordinate transformation, such that the numerical approximation is the exact flow generated by  $\overline{H}(p, q, t) = H(p, q) + R_1(p, q) + R_2(p, q, t)$  where  $R_2$  is analytic and  $h$ -periodic in  $t$ . As is well known the energy  $\overline{H}$  along the (numerical) trajectory varies as

$$\overline{H}(p(t), q(t), t) = \overline{H}(p(0), q(0), 0) + \int_0^t \frac{\partial}{\partial s} R_2(p(s), q(s), s) ds,$$

thus if  $R_2$  is bounded we have the result that the  $\overline{H}$  grows at most linearly with a proportionality factor that is exponentially small in  $h$  if  $h$  satisfies the smallness assumption  $\rho h < \pi\xi\delta/e$ . If  $h$  is larger than this, Theorem 2 still applies, but now the time-dependent part is only  $\mathcal{O}(h^{r-1})$ . The above integral may still be relatively well bounded, unless there are resonances between the frequencies of the trajectory  $(p(t), q(t))$  and the  $h/|k|$ -periodic time-dependency, in  $R_2^k(p, q) \exp(2\pi i |k|t/h)$ . For large enough  $|k|$ ,  $R_2^k$  will be exponentially small by the analyticity thus it is the lower Fourier modes that might cause instabilities. Suppose the trajectory  $(p(t), q(t))$  is quasi-periodic, with frequency vector  $\omega \in \mathbb{R}^{d'}$  i.e.  $p(t) = \sum_{m \in \mathbb{Z}^{d'}} p_k \exp(i\omega \cdot mt)$ ,  $q(t) = \sum_{m \in \mathbb{Z}^{d'}} q_k \exp(i\omega \cdot mt)$ ,  $d' \leq d$  with some constants  $p_k, q_k \in \mathbb{C}^d$ . Inserted into the integral and collecting terms the integral takes the form

$$(5.1) \int_0^t \frac{\partial}{\partial s} R_2(p(s), q(s), s) ds = \sum_{k \in \mathbb{Z} \setminus \{0\}, m \in \mathbb{Z}^d} r_{m, k} \int_0^t \exp(i(\omega \cdot m + 2\pi k/h)s) ds,$$

with  $r_{k,m} \in \mathbb{C}^d$  some constants. Because of analyticity of  $R_2$  the sum is convergent. The integral is

$$\int_0^t \exp(i(\omega \cdot m + 2\pi k/h)s) ds = \frac{\exp(i(\omega \cdot m + 2\pi k/h)t) - 1}{i(\omega \cdot m + 2\pi k/h)}.$$

We note on one hand that if the trajectory itself is resonant i.e.  $\omega \cdot m = 0$ , this causes no problems. If, on the other hand, we have a *numerical resonance*,  $\omega \cdot m + 2\pi k/h = 0$  for some  $m \in \mathbb{Z}^d, k \in \mathbb{Z}$  the integral simplifies to  $t$ ,- thus we get a secular drift in the energy whose proportionality factor  $r_{m,k}$  is exponentially small if  $\rho h < \pi \xi \delta / \epsilon$ , and might be large if this condition is not satisfied. In that case if  $\omega, h$  are such that resonances are avoided the energy drift might be bounded, but due to abundant “near resonances” their total contribution can be significant. Typically strong non-resonances conditions such as  $|\omega \cdot m + 2\pi k/h| > \psi / (|m|_1 + |k|)^\theta$  with some  $\psi, \theta > 0$  guarantees that the sum (5.1) still is bounded, see [4] for analysis along these lines. Numerical experiments [22, 18, 7] clearly show the effect of numerical resonances. Conservative discretizations are particularly prone to such effects since resonance effects are not damped out. In celestial mechanics such effects were observed for symplectic schemes[22] and by careful step size selection dominant resonances were avoided, and significant improvements in accuracy achieved. Also in molecular dynamics applications resonance induced instabilities have been observed and quite some efforts have been carried out to circumvent these problems without introducing dissipation[18]

## 5.2 KAM and Nekhoroshev’s theorems for symplectic schemes

KAM [19, 7] and Nekhoroshev [13] theorems have been proved for symplectic schemes in an ad-hoc fashion. These are theorems guaranteeing the stability of discretizations of close to integrable Hamiltonians  $H = H_0 + \epsilon H_1$ , given some resonance and non-degeneracy conditions on  $H_0$  and a smallness assumption on  $\epsilon H_1$ . Theorem 2 simplifies these proofs significantly, as we know that the numerical trajectory is given as the flow generated by  $\overline{H}(p, q) = H_0(p, q) + \epsilon H_1(p, q) + H_2(p, q, t)$ , with  $p, q$  conjugate variables. Simply by extending the phase space to remove the time-dependency we arrive at an Hamiltonian

$$\overline{H} = \{H_0 + e\} + \{\epsilon H_1(p, q) + H_2(p, q, \tau)\},$$

where  $e$  and  $\tau$  are conjugate variables. Standard KAM and Nekhoroshev theorems are now valid by assuming that the non-resonance and non-degeneracy condition holds for  $H_0 + e$  and that the smallness assumption holds for  $\epsilon H_1(p, q) + H_2(p, q, \tau)$ . Such theorems lead to proofs of the celebrated linear error growth estimates for symplectic schemes applied to close to integrable Hamiltonian systems, see [6, 12] for details.

## 6 Conclusions

We have by a coordinate transformation shown that the trajectory produced by iterating a one-step numerical approximation is exactly represented by the flow of an analytic time  $h$ -periodic vector field. By making a step size restriction we showed that the exponentially small error committed in autonomous modified equations is a direct consequence of the analyticity. Our bounds appear to be close to optimal, in

terms of constants entering the exponential bounds. In this way we paved the way for applying known theorems for differential equations to understand the properties of numerical discretizations. The finite-time restriction that follows from Theorem 1 can then be avoided, and proper stability theorems such as “the KAM theorem” can be proved for infinite times.

Further generalizations and improvements of the results may include :

- In order to identify particular methods a more detailed analysis similar to that in [3, 6] for e.g. B-series methods can be carried out within the framework of B-series mappings. This would allow one to identify how e.g. the coefficients of RK methods enter the estimates.

-Obstructions both in terms of  $F$  and the numerical methods to letting  $s \rightarrow \infty$  or at least further than we achieved are interesting to pursue, as this might lead to new numerical schemes. In particular for systems with  $F$  having a hyperbolic structure such results seem possible, see [10] and the references therein. One first avenue to improvement is in the commutator bound of Corollary 3. By introducing coordinates adapted to the trajectories  $x(t)$ , and a corresponding norm an improvement can be achieved. For quasi-periodic orbits [15] Pöschel used Fourier weighted norms which in our setting can remove the factor  $e$  (originating from  $C_{p,q} \leq e(p+q+1)$  in Lemma 4), in the exponential estimates.

-The construction might be generalized to the analysis of discretizations of vector fields,  $F(y)$  that are only Gevrey- $\gamma$  in  $y$  provided we can prove the equivalent to Lemma 2. In this case an optimal truncation might have to be carried out as the bounds on  $G^k$  diverge due to worse upper bound on  $C_{p,q}$  for such vector fields.

## Acknowledgments

I would like to thank J. Niesen for carefully reading the first draft, and making many valuable comments. E. Hairer and C. Lubich questioned the need for including the coordinate transform in Theorem 2, and Theorem 4 is the result of this.

## Appendix

*Proof of Proposition 1.* Periodicity follows directly. The proof of Gevrey class is rather technical, and we only sketch the idea. The function  $\chi_n$  is analytic except at  $x \in \mathbb{Z}$ . Considering the interval  $x \in [0, 1]$  we estimate the derivatives by the Cauchy integral formula

$$\chi_\eta^{(k+1)}(x) = \frac{k!}{2\pi i} \oint_{|x-z|=r_x} \frac{\chi'_n(z)}{(x-z)^{k+1}} dz \Rightarrow |\chi_\eta^{(k+1)}(x)| \leq \frac{k!}{r_x^k} \sup_{|x-z|=r_x} |\chi'_n(z)|.$$

Because  $\chi_\eta$  is not analytic at  $x = 0$  (the argument for the other points is similar) the radius  $r_x$  shrinks as  $x \rightarrow 0^+$  so we consider bounds in a wedge in  $\mathbb{C}$  with its vertex at  $x = 0$ , thus  $r_x = \mathcal{O}(x)$  while  $\sup_{|x-z|=r_x} |\chi'_n(z)| = \mathcal{O}(\exp(-1/r_x^\eta)/r_x^{\eta+1})$  when  $x \rightarrow 0$ . Minimizing these contributions, using Stirling’s formula to bound the factorial gives the desired bound on the derivatives of  $\chi_\eta$ , with  $\gamma = 1 + 1/\eta$ .  $\square$

*Proof of Lemma 1.* The Fourier coefficients are given by

$$\begin{aligned}
R^k &= \frac{1}{h} \int_{-h/2}^{h/2} R(t) \exp(2\pi ikt/h) dt \\
\text{Int. by parts } m \text{ times} &= \frac{h^{m-1}}{(-2\pi ik)^m} \int_{-h/2}^{h/2} R^{(m)}(t) \exp(2\pi ikt/h) dt \\
\Rightarrow |R^k| &\leq \frac{h^m}{(2\pi|k|)^m} \|R^{(m)}\|_\infty \leq M \frac{h^m c^m m^{\gamma m}}{(2\pi|k|)^m}.
\end{aligned}$$

Choosing  $m = \lfloor \frac{1}{e^{\frac{1}{\gamma\sqrt{ch/2\pi|k|}}}} \rfloor$  we obtain  $|R^k| \leq Me \exp\left(-\frac{\gamma}{e} \sqrt{\frac{2\pi|k|}{ch}}\right)$ .  $\square$

*Proof of Lemma 2.* By Cauchy's integral formula  $d_z GF(z) = \frac{1}{2\pi i} \oint_{|s|=\tilde{r}} \frac{G(z+sF(z))}{s^2} ds$

$$\begin{aligned}
\Rightarrow \|d_z GF\|_n &= \frac{1}{2\pi} \left\| \oint_{|s|=\tilde{r}} \frac{G(z+sF(z))}{s^2} ds \right\|_n \\
&= \frac{1}{2\pi} \inf_{\tilde{r}} \max_j \sup_{z \in \mathcal{D}} \left| \oint_{|s|=\tilde{r}} \frac{G_j(z+sF(z))}{s^2} ds \right| \text{dist}_{\mathcal{D}}(z)^n \\
&\leq \inf_{\tilde{r}} \max_j \sup_{z \in \mathcal{D}, |s|=\tilde{r}} \left| \frac{1}{\tilde{r}} G_j(z+sF(z)) \right| \text{dist}_{\mathcal{D}}(z)^n.
\end{aligned}$$

By the definition of the Nagumo norm  $\max_j |G_j(z)| \leq \|G\|_m / \text{dist}_{\mathcal{D}}(z)^m$  for  $m = 0, \dots, n-1$  and furthermore  $\text{dist}_{\mathcal{D}}(z+sF(z)) \geq \text{dist}_{\mathcal{D}}(z) - |s| \max_j |F_j| = \text{dist}_{\mathcal{D}}(z) - \tilde{r} \max_j |F_j|$  when  $z \in \mathcal{D}$ . From this we have

$$\begin{aligned}
&\inf_{\tilde{r}} \max_j \sup_{z \in \mathcal{D}, |s|=\tilde{r}} \left| \frac{1}{\tilde{r}} G_j(z+sF(z)) \right| \text{dist}_{\mathcal{D}}(z)^n \\
&\leq \inf_{\tilde{r}} \sup_{z \in \mathcal{D}, |s|=\tilde{r}} \frac{1}{\tilde{r}} \frac{\|G\|_m}{(\text{dist}_{\mathcal{D}}(z) - \tilde{r} \max_j |F_j(z)|)^m} \text{dist}_{\mathcal{D}}(z)^n \\
&\leq (1+m) \left(1 + \frac{1}{m}\right)^m \|G\|_m \|F\|_{n-m-1},
\end{aligned}$$

where in the last inequality we minimized the bound by choosing  $\tilde{r} = \text{dist}_{\mathcal{D}}(z) / (\max_j |F_j(z)|(1+m))$ .  $\square$

*Proof of Corollary 3.* Applying Lemma 2 twice, swapping indexes.  $\square$

*Proof of Lemma 3.* Since  $G_n$  is real  $\|G_n^k\|_l = \|G_n^{-k}\|_l$  it follows that  $\|G_n^k\|_l \leq \frac{\rho_n^+}{2w(k,h)}$ .

$$\begin{aligned}
\rho_n^0 = \|G_n^0\|_n &\leq 2 \sum_{p+q=n-1} \sum_{l \geq 1} \int_0^s \|[G_p^l, G_q^{-l}]\|_n \exp(-4\pi l\sigma/h) d\sigma \\
(\text{Cor. 3}) &\leq 2 \sum_{p+q=n-1} C_{p,q} \sum_{l \geq 1} \int_0^s \|G_p^l\|_p \|G_q^{-l}\|_q d\sigma \\
&\leq \frac{1}{2} \sum_{p+q=n-1} C_{p,q} \sum_{l \geq 1} \frac{1}{w(l,h)^2} \int_0^s \rho_p^+(\sigma) \rho_q^+(\sigma) d\sigma \\
&\leq c_1 \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^+(\sigma) \rho_q^+(\sigma) d\sigma,
\end{aligned}$$

with  $c_1 = \frac{\zeta(4)-1}{2} \exp(-2c\sqrt{1/h})$  since  $\sum_{l \geq 1} w(l, h)^{-2} \leq \exp(-2c\sqrt{1/h}) \sum_{l \geq 1} \frac{1}{(l+1)^4} = \exp(-2c\sqrt{1/h})(\zeta(4) - 1)$ .<sup>8</sup>

For  $\rho_n^+$  we consider first the positive Fourier indexes (the negative contribution has the same upper bound).

$$\begin{aligned}
\sum_{k \geq 1} w(k, h) \|G_n^k\|_n &\leq \sum_{k \geq 1} w(k, h) \sum_{p+q=n-1} \int_0^s \|[G_p^0(\sigma), G_q^k(\sigma)]\|_n d\sigma \\
&+ 2 \sum_{k \geq 1} w(k, h) \sum_{p+q=n-1} \sum_{l \geq 1} \int_0^s \|[G_p^{-l}(\sigma), G_q^{k+l}(\sigma)]\|_n d\sigma \\
\text{Cor. 3} &\leq \sum_{k \geq 1} w(k, h) \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^0(\sigma) \|G_q^k\|_q d\sigma \\
w(k, h) \text{ increasing in } k &+ 2 \sum_{p+q=n-1} C_{p,q} \sum_{k,l \geq 1} \int_0^s \|G_p^{-l}\|_p w(k+l, h) \|G_q^{k+l}\|_q d\sigma \\
&\leq \frac{1}{2} \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^0(\sigma) \rho_q^+(\sigma) d\sigma \\
&+ \sum_{p+q=n-1} C_{p,q} \sum_{l \geq 1} \int_0^s \|G_p^{-l}\|_p \rho_q^+ d\sigma \\
&\leq \frac{1}{2} \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^0(\sigma) \rho_q^+(\sigma) d\sigma \\
&+ \frac{1}{2} \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^+ \rho_q^+ d\sigma \sum_{l \geq 1} \frac{1}{w(l, h)} \\
&\leq \frac{1}{2} \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^0(\sigma) \rho_q^+(\sigma) d\sigma \\
&+ \frac{1}{2} c_2 \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^+ \rho_q^+ d\sigma.
\end{aligned}$$

since  $\sum_{l \geq 1} \frac{1}{w(l, h)} \leq \exp(-c\sqrt{1/h}) \sum_{l \geq 1} \frac{1}{(l+1)^2} = \exp(-c\sqrt{1/h})(\pi^2/6 - 1) = c_2$ . Adding the contribution from the negative indexes we have

$$\rho_n^+ \leq \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^0(\sigma) \rho_q^+(\sigma) d\sigma + c_2 \sum_{p+q=n-1} C_{p,q} \int_0^s \rho_p^+ \rho_q^+ d\sigma$$

Since  $c_1 \leq c_2^2/10$  we obtain the bound by letting  $\epsilon = c_2$ .  $\square$

*Proof of Lemma 4.* We are interested in the asymptotic behavior of  $\rho_n, n > 0$ . The bound  $\max_{p+q=n-1} C_{p,q} \leq en$  follows from  $C_{p,q} \leq C_{(p+q)/2, (p+q)/2}$ . To simplify our analysis we also use  $\epsilon^2/10 \leq \epsilon/3$  for  $\epsilon < 2/3$ . Let  $\rho = \max\{\|F + R^0\|_0, \|R\|_0^*\}$ . Then the inequalities in Lemma 3 are satisfied by  $\rho_n^0 = D_n^0 s^n e^n$  and  $\rho_n^+ = D_n^+ s^n e^n$  provided

$$\begin{aligned}
D_n^0 &= \epsilon/3 \sum_{p+q=n-1} D_p^+ D_q^+, \quad D_0^+ = \rho \\
(6.1) \quad D_n^+ &= \sum_{p+q=n-1} D_p^0 D_q^+ + \epsilon \sum_{p+q=n-1} D_p^+ D_q^+, \quad D_0^0 = \rho,
\end{aligned}$$

---

<sup>8</sup> $\zeta(j) = \sum_{n \geq 1} n^{-j}$

Introducing the generating functions  $D^+ = \sum_{n \geq 0} D_n^+ x^n$  and  $D^0 = \sum_{n \geq 0} D_n^0 x^n$  we find by multiplying (6.1) by  $x^n$  and summing over  $n = 0, 1, \dots$

$$\begin{aligned} D^0 &= \rho + x\epsilon/3(D^+)^2, \quad D^+ = \rho + xD^+D^0 + x\epsilon(D^+)^2 \\ &\Downarrow \\ (6.2) \quad 0 &= \frac{x^2\epsilon}{3}(D^+)^3 + x\epsilon(D^+)^2 + (x\rho - 1)D^+ + \rho \end{aligned}$$

The singularity in  $x$  of the real solution  $D^+(x)$  closest to the origin determines the radius of convergence of  $D^+$  around  $x = 0$ . The dominant singularity occurs when the discriminant is zero

$$4(\rho x)^3 - 12(1 + \epsilon)(\rho x)^2 + 12(1 + \epsilon)^2(\rho x) - 4 - 3\epsilon = 0,$$

and is given by  $\rho x = \xi(\epsilon) := 1 + \epsilon - \frac{1}{2}(18\epsilon + 24\epsilon^2 + 8\epsilon^3)^{1/3} \leq 1$ . By Decartes' rule of signs it follows that there is one negative and two positive roots of (6.2), and the singularity occurs when the two positive roots coalesce. This double root is

$$d^+ = \rho \tilde{\xi}(\epsilon) = \rho \frac{-2\epsilon + \sqrt{4\epsilon(1 + \epsilon - \xi(\epsilon))}}{2\epsilon\xi(\epsilon)}$$

Since the coefficients of  $D^+$  are positive we have by the Cauchy integral formula the bound  $D_n^+ = \frac{1}{2\pi i} \oint_{|z|=\xi/\rho} \frac{D^+(z)}{z^{n+1}} dz \leq \frac{\rho^n}{\xi^n} d^+$ , giving  $\rho_n^+ \leq \rho \left(\frac{\rho\epsilon s}{\xi}\right)^n \tilde{\xi}(\epsilon)$ . Since the coefficients of  $D^0$  are positive, we again have by the Cauchy formula  $D_n^0 = \frac{1}{2\pi i} \oint_{|z|=\xi/\rho} \frac{D^0(z)}{z^{n+1}} dz \leq \frac{\rho^n}{\xi^n} (\rho + \frac{\epsilon\xi}{3\rho} d^{+2}) = \rho \frac{\rho^n}{\xi^n} (1 + \frac{\epsilon\xi\tilde{\xi}^2}{3})$ , thus  $\rho_n^0 \leq \rho \left(\frac{\rho\epsilon s}{\xi}\right)^n (1 + \frac{\epsilon\xi\tilde{\xi}^2}{3})$ ,  $\square$

*Proof of Proposition 2.* Recalling inequality (3.11) we have by Lemma 4

$$\begin{aligned} |k| \neq 0: \quad \|G^k\|'_0 &\leq \frac{\rho_0^+}{2w(k, h)} + \sum_{l \geq 1} \frac{\|G_l^k\|_l}{\delta^l} \leq \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \sum_{l \geq 1} \frac{\rho_l^+}{\delta^l} \right\} \\ &\leq \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \rho \tilde{\xi}(\epsilon) \sum_{l \geq 1} \left( \frac{\rho\epsilon s}{\xi\delta} \right)^l \right\} \\ &= \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \rho \tilde{\xi}(\epsilon) \frac{z}{1-z} \Big|_{z=\frac{\rho\epsilon s}{\xi\delta}} \right\} \\ k = 0: \quad \|G^0\|'_0 &\leq \rho_0^0 + \sum_{l \geq 1} \frac{\|G_l^0\|_l}{\delta^l} \leq \rho_0^0 + \sum_{l \geq 1} \frac{\rho_l^0}{\delta^l} \\ &\leq \rho_0^0 + \rho \left(1 + \frac{\epsilon\xi\tilde{\xi}^2}{3}\right) \sum_{l \geq 1} \left( \frac{\rho\epsilon s}{\xi} \right)^l = \rho_0^0 + \rho \left(1 + \frac{\epsilon\xi\tilde{\xi}^2}{3}\right) \frac{z}{1-z} \Big|_{z=\frac{\rho\epsilon s}{\xi\delta}} \end{aligned}$$

which is bounded for  $s < \xi\delta/\rho\epsilon$ .  $\square$

*Proof of Theorem 3.* By inserting  $s = \xi\delta/\rho\epsilon - h/\pi$  into (3.12) we obtain

$$\begin{aligned} \|G^k\|'_0 &\leq \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \rho \tilde{\xi} \frac{z}{1-z} \Big|_{z=\rho\epsilon s/\xi\delta} \right\} \\ &= \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \rho \tilde{\xi} \frac{1 - \rho\epsilon h/\pi\xi\delta}{\rho\epsilon h/\pi\xi\delta} \right\} \leq \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{\epsilon h} \right\} \end{aligned}$$

giving

$$\begin{aligned} \|\overline{F}^k\|'_0 &\leq \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{eh} \right\} \exp\left(-\frac{2\pi|k|}{h}\left(\frac{\xi\delta}{\rho e} - h/\pi\right)\right) \\ \text{By } h < \pi\xi\delta/\rho e &\leq \frac{1}{2w(k, h)} \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{eh} \right\} \exp\left(-\frac{2\pi}{h}\left(\frac{\xi\delta}{\rho e} - h/\pi\right)\right) \\ &= \frac{e^2}{2w(k, h)} \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{eh} \right\} \exp\left(-\frac{2\pi\xi\delta}{\rho h}\right) \end{aligned}$$

Thus by  $w(k, h) > \frac{2}{3\epsilon}(1 + |k|)^2$

$$\begin{aligned} \|R_2\|'_0 &\leq \sum_{k \neq 0} \|\overline{F}^k\|'_0 \leq \sum_{k \neq 0} \frac{3\epsilon e^2}{4(1 + |k|)^2} \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{eh} \right\} \exp\left(-\frac{2\pi\xi\delta}{\rho h}\right) \\ &= \left(\frac{\pi^2}{6} - 1\right) \frac{3\epsilon e^2}{2} \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{eh} \right\} \exp\left(-\frac{2\pi\xi\delta}{\rho h}\right) \leq 8\epsilon \left\{ \rho_0^+ + \frac{\pi\xi\tilde{\xi}\delta}{eh} \right\} \exp\left(-\frac{2\pi\xi\delta}{\rho h}\right) \end{aligned}$$

□

*Proof of Theorem 4.* There is an explicit formula available for the solution of (4.2)

$$\overline{F}^*(s) = \overline{F} + \sum_{k \geq 1} \int_0^s \int_0^{\sigma_1} \cdots \int_0^{\sigma_{k-1}} [W(\sigma_k), [\cdots, [W(\sigma_1), \overline{F}]] \cdots]] d\sigma_k \cdots d\sigma_1,$$

where  $s$  is chosen as in (4.1). Since  $W$  is frozen at  $t = 0$  we have for the Fourier coefficients of  $\overline{F}^*$

$$\overline{F}^{*k}(s) = \overline{F}^k + \sum_{k \geq 1} \int_0^s \int_0^{\sigma_1} \cdots \int_0^{\sigma_{k-1}} [W(\sigma_k), [\cdots, [W(\sigma_1), \overline{F}^k]] \cdots]] d\sigma_k \cdots d\sigma_1.$$

Introducing the domain  $\mathcal{D}'' \subset \mathcal{D}' \subset \mathcal{D}$  with corresponding Nagumo norm  $\|F\|''_0$ , and  $\delta' = \inf_{z \in \mathcal{D}''(x)} \text{dist}_{\mathcal{D}'}(z)$ , whave have by recursively applying Corollary 3 the bound

$$\begin{aligned} \|[W(\sigma_k), [\cdots, [W(\sigma_1), \overline{F}^k]] \cdots]]\|'_k &\leq C_k \|W(\sigma_k)\|'_0 \cdots \|W(\sigma_1)\|'_0 \|\overline{F}^k\|'_0 \\ C_1 = 2, \quad C_k &= 2 \prod_{j=2}^k \left(1 + j\left(1 + \frac{1}{j-1}\right)^{j-1}\right) \leq \frac{5}{e^2} e^k k!, \end{aligned}$$

using (3.11), i.e.  $\|F\|''_0 \leq \frac{1}{\delta'^k} \|F\|'_k$  we therefore have

$$\begin{aligned} \|\overline{F}^{*k}\|''_0 &\leq \|\overline{F}^k\|'_0 + \frac{5}{e^2} \sum_{k \geq 1} \int_0^s \int_0^{\sigma_1} \cdots \int_0^{\sigma_{k-1}} \frac{k! e^k}{\delta'^k} \|W(\sigma_k)\|'_0 \cdots \|W(\sigma_1)\|'_0 \|\overline{F}^k\|'_0 d\sigma_k \cdots d\sigma_1 \\ (6.3) \quad &= \|\overline{F}^k\|'_0 \left\{ 1 + \frac{5}{e^2} \sum_{k \geq 1} \left( \frac{e}{\delta'} \int_0^s \|W(\sigma)\|'_0 d\sigma \right)^k \right\} \end{aligned}$$

So we get the requirement  $\int_0^s \|W\|'_0 d\sigma < \delta'/e$ . Since  $W(\cdot, t=0, \sigma) = \sum_{|k| \neq 0} i \text{sign}(k) \overline{F}^k(\cdot, \sigma)$ , recalling Prop. 2 and (3.6),

$$\begin{aligned} \|W(\sigma)\|'_0 &\leq \sum_{|k| \neq 0} \|\overline{F}^k(\sigma)\|'_0 \leq \left\{ \rho_0^+ + \rho \tilde{\xi} \frac{z}{1-z} \Big|_{z=\rho e\sigma/\xi\delta} \right\} \sum_{k \geq 1} \frac{1}{w(k, h)} \exp\left(-\frac{2\pi k\sigma}{h}\right) \\ &\leq \left\{ \rho_0^+ + \rho \tilde{\xi} \frac{z}{1-z} \Big|_{z=\rho e\sigma/\xi\delta} \right\} \frac{3\epsilon}{8} \sum_{k \geq 1} \frac{1}{(1+k)^2} \exp\left(-\frac{2\pi|k|\sigma}{h}\right) \\ &= \frac{3\epsilon(\pi^2/6-1)}{8} \left\{ \rho_0^+ + \rho \tilde{\xi} \frac{z}{1-z} \Big|_{z=\rho e\sigma/\xi\delta} \right\} \exp\left(-\frac{2\pi\sigma}{h}\right). \end{aligned}$$

Since  $\int_0^s \sigma \exp(-2\pi\sigma/h) d\sigma \leq \frac{h^2}{(2\pi)^2}$ ,  $\int_0^s \exp(-2\pi\sigma/h) d\sigma \leq \frac{h}{2\pi}$ , we have

$$\begin{aligned} \int_0^s \|W(\sigma)\|'_0 d\sigma &\leq \frac{3\epsilon(\pi^2/6-1)}{8} \left\{ \frac{\rho_0^+ h}{2\pi} + \frac{\rho^2 \tilde{\xi} e h^2}{\xi \delta (2\pi)^2} \frac{1}{1-z} \Big|_{z=\rho e\sigma/\xi\delta} \right\} \\ (\rho_0^+ \leq \rho, 4.1) &= \frac{3\epsilon(\pi^2/6-1)}{8} \left\{ \frac{1}{2\pi} + \frac{\tilde{\xi}(\epsilon)\pi}{e(2\pi)^2} \right\} \rho h \leq \frac{\rho h}{27} \end{aligned}$$

by using the observation that  $\epsilon \tilde{\xi}(\epsilon)$  reaches its maximum for  $\epsilon = 2/3$ . Thus  $\int_0^s \|W\|'_0 d\sigma < \delta'/e$  provided  $\rho h \leq 27/e\delta'$ , and  $\|\overline{F}^{*k}\|''_0$  is bounded. By the step size assumption and Theorem 2 analyticity and exponential smallness from the same corresponding properties of  $\overline{F}$ .  $\square$

## References

- [1] Akhmatskaya, E.; Reich, S. "The targeted shadowing hybrid Monte Carlo (TSHMC) method." *Advances in Molecular Simulation Algorithms, Lecture Notes in Computational Science and Engineering*, to appear.
- [2] Benettin, G.; Giorgilli, A. "On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms". *J. Stat. Phys.* 74 (1994) 1117-1143.
- [3] Calvo, M.P.; Murua, A.; Sanz-Serna, J.M. "Modified equations for ODEs" *Chaotic numerics*, American Mathematical Society, Providence, 1994, *Contemporary Mathematics*, Vol. 172, 63-74.
- [4] Cohen, D.; Hairer, E.; Lubich, C. "Modulated Fourier expansions of highly oscillatory differential equations." *Found. Comput. Math.* 3 (2003), no. 4, 327-345.
- [5] Fiedler, B.; Scheurle, J. "Discretization of homoclinic orbits, rapid forcing and "invisible" chaos.", 1996 *Mem. Amer. Math. Soc.* 119, no 570.
- [6] Hairer, E.; Lubich, C. "The life-span of backward error analysis for numerical integrators." *Numer. Math.* 76 (1997), no. 4, 441-462.
- [7] Hairer, E.; Lubich, C.; Wanner, G. "Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations." *Springer Series in Computational Mathematics*, 31. Springer-Verlag, Berlin, 2002.
- [8] Izaguirre, J.A.; Scott, S. "Shadow Hybrid Monte Carlo: An Efficient Propagator in Phase Space of Macromolecules." *Journal of Computational Physics*, Vol. 200, No. 2, (2004) 581-604
- [9] John, F. "Partial differential equations" 4th ed. Springer Verlag 1982.

- [10] Li, M.-Ch. "Qualitative property between flows and numerical methods." *Nonlinear Anal.* 59 (2004), no. 5, 771–787.
- [11] McLachlan, R.I.; Quispel, G.R.W. "Six lectures on the geometric integration of ODEs", *Foundations of Computational Mathematics*, (2001) 155-210.
- [12] Moan, P.C. "On the KAM and Nekhoroshev theorems for symplectic integrators and implications for error growth." *Nonlinearity* 17 (2004), no. 1, 67–83.
- [13] Moan, P.C. "On backward error analysis and Nekhoroshev stability in the numerical analysis of conservative systems of ODEs". PhD thesis, University of Cambridge 2002.
- [14] Moser, J. "On the construction of almost periodic solutions for ordinary differential equations." 1970 *Proc. Internat. Conf. on Functional Analysis and Related Topics* (Tokyo, 1969) pp. 60–67 Univ. of Tokyo Press, Tokyo
- [15] Pöschel, J. "Nekhoroshev estimates for quasi-convex Hamiltonian systems" *Math. Z.* 213 (1993) 187-216.
- [16] Reich, S. "Backward error analysis for numerical integrators." *SIAM J. Numer. Anal.* 36 (1999), no. 5, 1549–1570
- [17] Sauzin, D. "Nekhoroshev estimates and instability for Gevrey class Hamiltonians." *Dynamical systems. Part I*, (2003)*Pubbl. Cent. Ric. Mat. Ennio Giorgi*, 199–217.
- [18] Schlick, T.; Mandziuk, M.; Skeel, R. D.; Srinivas, K. "Nonlinear resonance artifacts in molecular dynamics simulations." *J. Comput. Phys.* 140 (1998), no. 1, 1–29.
- [19] Shang, Z.-j. "KAM theorem of symplectic algorithms for Hamiltonian systems." *Numer. Math.* 83 (1999), no. 3, 477–496.
- [20] Walter, W. "An elementary proof of the Cauchy-Kowalevsky theorem" *Amer. Math. Monthly* 92 (1985) no.2, 115-126.
- [21] Warming, R. F.; Hyett, B. J. "The modified equation approach to the stability and accuracy analysis of finite-difference methods." *J. Computational Phys.* 14 (1974), 159–179.
- [22] Wisdom, J.; Holman, M. "Symplectic maps for the n-body problem - Stability analysis" *Astronomical Journal*, vol. 104, no. 5 (1992), 2022–2029.