BACKWARD ERROR ANALYSIS FOR NUMERICAL INTEGRATORS

SEBASTIAN REICH*

Abstract. Backward error analysis has become an important tool for understanding the long time behavior of numerical integration methods. This is true in particular for the integration of Hamiltonian systems where backward error analysis can be used to show that a symplectic method will conserve energy over exponentially long periods of time. Such results are typically based on two aspects of backward error analysis: (i) It can be shown that the modified vector fields have some qualitative properties which they share with the given problem and (ii) an estimate is given for the difference between the best interpolating vector field and the numerical method. These aspects have been investigated recently, for example, by BENETTIN & GIORGILLI in [5], by HAIRER in [16], and by HAIRER & LUBICH in [18]. In this paper we aim at providing a unifying framework and a simplification of the existing results and corresponding proofs. Our approach to backward error analysis is based on a simple recursive definition of the modified vector fields that does not require explicit Taylor series expansion of the numerical method and the corresponding flow maps as in [5, 16, 18]. As an application we discuss the long time integration of chaotic Hamiltonian systems and the approximation of time averages along numerically computed trajectories.

Key words. numerical integrators, differential equations, error analysis, Hamiltonian systems, long-time dynamics

AMS subject classifications. 65L05, 34A50, 65L70, 70H05

1. Introduction. In this paper, we consider the relationship between solutions to a given system of ordinary differential equations (vector fields)

$$\frac{d}{dt}\boldsymbol{x} = \boldsymbol{Z}(\boldsymbol{x})\,,$$

numerical approximations

$$\boldsymbol{x}_{n+1} = \boldsymbol{\Psi}_{\delta t}\,(\boldsymbol{x}_n)$$

to them, and solutions to associated modified equations

$$\frac{d}{dt}\boldsymbol{x} = \tilde{\boldsymbol{X}}_i\,(\boldsymbol{x};\delta t) \qquad (i \geq 1)\,.$$

The vector fields $\tilde{\boldsymbol{X}}_i(\delta t)$ are formulated in terms of an asymptotic expansion in the step-size δt, i.e., are chosen such that the numerical solution can formally be interpreted, with increasing index i, as the more and more accurate solution of the modified equation. Previous papers on backward error analysis for differential equations include those by WARMING & HYETT [38], GRIFFITHS & SANZ-SERNA [12], BEYN [6], FENG [10], FIEDLER & SCHEURLE [11], and SANZ-SERNA [31]. Another early reference to related ideas is MOSER [24] who discusses the approximation of a symplectic map near an equilibrium by the flow map of a Hamiltonian vector field.

More recently, general formulas for the computation of the modified vector fields $\tilde{\boldsymbol{X}}_i(\delta t)$ have been derived by HAIRER [16], CALVO, MURUA & SANZ-SERNA [7], BENETTIN & GIORGILLI [5], and REICH [26]. In papers by NEISHTADT [25], BENETTIN & GIORGILLI [5], and HAIRER & LUBICH [18], the question of closeness of the numerical approximations and the solutions of the modified equations has been addressed.

*Konrad-Zuse-Zentrum, Takustr. 7, D-14195 Berlin, current address: Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey GU2 5XH, UK (S.Reich@surrey.ac.uk).

In particular, it has been shown in these papers that the difference can be made exponentially small in the step-size δt, i.e.

$$|| \mathbf{\Phi}_{\delta t, \tilde{\boldsymbol{X}}_{i_*}}(\boldsymbol{x}) - \mathbf{\Psi}_{\delta t}(\boldsymbol{x}) || \leq c_1 \, \delta t \, e^{-c_2/\delta t} \,, \tag{1.1}$$

provided the vector field $\boldsymbol{Z}$ and the numerical one step method $\mathbf{\Psi}_{\delta t}$ are real analytic [22]. Here $c_1, c_2 > 0$ are appropriate constants, $\mathbf{\Phi}_{\delta t, \tilde{\boldsymbol{X}}_{i_*}}$ denotes the time-δt-flow map of the vector field $\tilde{\boldsymbol{X}}_{i_*}$, and the index $i_*(\delta t)$ has been chosen such that the difference is minimized.

Backward error analysis is of utmost importance for an understanding of the qualitative behavior of symplectic methods [32] for Hamiltonian problems. It has been shown by Hairer [16], Calvo, Murua & Sanz-Serna [7], Reich [26], and Benettin & Giorgilli [5] that for symplectic discretizations, the modified vector fields $\tilde{\boldsymbol{X}}_i(\delta t)$ are Hamiltonian. For special cases see also the papers by Auerbach & Friedman [4] and Yoshida [40]. The Hamiltonian structure of the modified equations implies that a symplectic integrator almost preserves the total energy over an exponentially long period of time [25, 23, 5, 18]. Similarly, the adiabatic invariant of a Hamiltonian system with rapidly rotating phase is also preserved over an exponentially long periods of time provided a symplectic method is used [29].

The fact that symplectic methods lead to modified equations that are Hamiltonian is a special instance of the so called geometric properties of backward error analysis. By this we mean the following: If the vector field $\boldsymbol{Z}$ belongs to a certain class of vector fields, like integral preserving or divergence-free vector fields, and the numerical approximation $\mathbf{\Phi}_{\delta t}$ also preserves the corresponding quantities, then the modified vector fields $\tilde{\boldsymbol{X}}_i(\delta t)$ will be in the same class as $\boldsymbol{Z}$. Besides symplectic methods, special instances of these geometric aspects have been discussed before. See, for example, the papers by Reich [26], Hairer & Stoffer [20] and Gonzalez, Higham & Stuart [13].

In this paper, we revisit backward error analysis by using a simple recursive scheme for the definition of the modified vector fields $\tilde{\boldsymbol{X}}_i(\delta t)$ as first proposed by the author in the unpublished report [26]. The main advantage of this formulations is that it does not require Taylor series expansions of the numerical one step method $\mathbf{\Psi}_{\delta t}$ and the flow maps $\mathbf{\Phi}_{\delta t, \tilde{\boldsymbol{X}}_i}$ in terms of the step size δt as it used in the papers by Benettin & Giorgilli [5], Hairer [16], and Hairer & Lubich [18]. This in turn allows for a simple characterization[1] of the geometric properties of the modified vector fields $\tilde{\boldsymbol{X}}_i(\delta t)$ and a rather simple proof for the exponentially small truncation error (1.1). Our approach is close to the one discussed by Benettin & Giorgilli [5] in the sense that we consider general one step methods[2] and that we use a "direct" approach[3]. However, different techniques are used and we will discuss this in more detail in §4.

In §5, we consider the numerical integration of a "chaotic" Hamiltonian system by a symplectic method and discuss the approximation of time-averages along numerically computed trajectories. We assume that a Poincaré section [14] can be defined and that the corresponding Poincaré section is uniformly hyperbolic. Backward error

[1]The general idea can already been found in the report [26].

[2]Hairer & Lubich [18] consider methods that can be represented by P-series [19]. Note that Runge-Kutta and partitioned Runge-Kutta methods fall under this category.

[3]This is in contrast to the "indirect" approach used by Neishtadt [25] where the one step method is first interpolated by the flow of a time-dependent vector field and averaging in time is then used to obtain an optimal approximating time-independent vector field.

analysis and the shadowing lemma [33] will be used to show that a numerically computed trajectory stays close to an exact solution over exponentially long periods of time. This and a large deviation theorem [36] allow us to discuss the convergence of long time averages along numerically computed trajectories. The anisotropic Kepler problem [15] will serve us as a numerical illustration. This problem requires the application of a symplectic variable step-size method as first discussed by the author in the technical report [27] and independently by HAIRER in [17].

2. The Modified Vector Field Recursion. Let us consider a smooth vector field

$$\frac{d}{dt}\boldsymbol{x} = \boldsymbol{Z}(\boldsymbol{x}), \tag{2.1}$$

$\boldsymbol{Z} : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$ and its discretization by a one step method [19]

$$\boldsymbol{x}_{n+1} = \boldsymbol{\Psi}_{\delta t}(\boldsymbol{x}_n) \;=\; \boldsymbol{x}_n + \delta t\,\boldsymbol{\psi}(\boldsymbol{x}_n, \delta t). \tag{2.2}$$

We assume that $\boldsymbol{\Psi}_{\delta t} : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$ is a smooth map and a method of order $p \geq 1$, i.e.

$$||\,\boldsymbol{\Phi}_{\delta t,\boldsymbol{Z}}(\boldsymbol{x}) - \boldsymbol{\Psi}_{\delta t}(\boldsymbol{x})\,|| = \mathcal{O}(\delta t^{p+1})$$

for all $\boldsymbol{x} \in \mathcal{U}$ where $\boldsymbol{\Phi}_{\delta t,\boldsymbol{Z}}$ is the time-δt-flow map of the differential equation (2.1).

As in described in the Introduction, we look for a family of vector fields $\tilde{\boldsymbol{X}}(\delta t)$ such that

$$\boldsymbol{\Phi}_{\delta t,\tilde{\boldsymbol{X}}(\delta t)} \approx \boldsymbol{\Psi}_{\delta t}$$

or, equivalently,

$$\boldsymbol{\Phi}_{1,\boldsymbol{X}(\delta t)} \approx \boldsymbol{\Psi}_{\delta t}\,, \qquad \boldsymbol{X}(\delta t) \;:=\; \delta t\,\tilde{\boldsymbol{X}}(\delta t)\,,$$

for all δt sufficiently small. Here $\boldsymbol{\Phi}_{1,\boldsymbol{X}}$ denotes the time-one-flow map of the vector field $\boldsymbol{X}(\delta t)$. The family of *modified vector fields* $\boldsymbol{X}(\delta t)$, $\delta t \geq 0$, is formally defined in terms of an asymptotic expansion in the step-size δt, i.e.

$$\boldsymbol{X}(\delta t) = \delta t \boldsymbol{\Delta X}_1 + \delta t^2 \boldsymbol{\Delta X}_2 + \delta t^3 \boldsymbol{\Delta X}_3 + \ldots$$

The formally infinite sequence of vector fields $\{\boldsymbol{\Delta X}_i\}_{i=1,\ldots,\infty}$ can be obtained by Taylor series expansion of the one step method $\boldsymbol{\Psi}_{\delta t}$, i.e.

$$\boldsymbol{\Psi}_{\delta t} = \mathbf{id} + \delta t \boldsymbol{\Psi}_1 + \delta t^2 \boldsymbol{\Psi}_2 + \ldots\,,$$

$\mathbf{id}(\boldsymbol{x}) = \boldsymbol{x}$ the identity map, and comparison of this series with the expansion of the time-one-flow map $\boldsymbol{\Phi}_{1,\boldsymbol{X}(\delta t)}$ in terms of δt. The vector fields $\boldsymbol{\Delta X}_i$ are chosen such that these two series coincide term by term. This is the general approach followed by BENETTTIN & GIORGILLI [5] and HAIRER [16]. The two papers differ in the way the Taylor series expansions are written down. But they lead to exactly the same sequence of vector fields $\{\boldsymbol{\Delta X}_i\}_{i=1,\ldots,\infty}$. We obviously have $\boldsymbol{\Delta X}_1 = \boldsymbol{Z}$ and $\boldsymbol{\Delta X}_i = \boldsymbol{0}$, $i = 2, \ldots, p$, for a method of order p.

We now give a recursive definition of the modified vector field $\boldsymbol{X}(\delta t)$ that does not require an explicit Taylor series expansion. This recursion was introduced by the author in the unpublished report [26]. First we formally introduce the "truncated" expansions $\boldsymbol{X}_i(\delta t)$, $i = 1, 2, \ldots, \infty$, by means of

$$\boldsymbol{X}_i(\delta t) = \sum_{j=1}^{i} \delta t^j \boldsymbol{\Delta X}_j \,.$$

We obviously have

$$\boldsymbol{X}_{i+1}(\delta t) := \boldsymbol{X}_i(\delta t) + \delta t^{i+1} \boldsymbol{\Delta X}_{i+1} \,.$$

Let us assume that $\boldsymbol{X}_i(\delta t)$ has been chosen such that the difference between the time-one-flow map of $\boldsymbol{X}_i(\delta t)$ and the numerical one step method $\boldsymbol{\Psi}_{\delta t}$ is $\mathcal{O}(\delta t^{i+1})$. This suggests to consider the following recursion:

$$\boxed{\begin{aligned} \boldsymbol{X}_{i+1}(\delta t) &:= \boldsymbol{X}_i(\delta t) + \delta t^{i+1} \boldsymbol{\Delta X}_{i+1} \,, && (2.3)\\ \boldsymbol{\Delta X}_{i+1} &:= \lim_{\delta t \to 0} \frac{\boldsymbol{\Psi}_{\delta t} - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)}}{\delta t^{i+1}} \,. && (2.4) \end{aligned}}$$

Indeed, this definition of $\boldsymbol{\Delta X}_{i+1}$ implies that $\boldsymbol{X}_{i+1}(\delta t)$, defined by (2.3), generates a time-one-flow map that is $\mathcal{O}(\delta t^{i+2})$ away from the numerical method $\boldsymbol{\Psi}_{\delta t}$. This can be seen from

$$\begin{aligned} \boldsymbol{\Phi}_{1,\boldsymbol{X}_{i+1}(\delta t)} - \boldsymbol{\Psi}_{\delta t} &= \boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)} + \delta t^{i+1} \boldsymbol{\Delta X}_{i+1} - \boldsymbol{\Psi}_{\delta t} + \mathcal{O}(\delta t^{i+2}) \\ &= \delta t^{i+1} \boldsymbol{\Delta X}_{i+1} - \delta t^{i+1} \lim_{\delta t \to 0} \frac{\boldsymbol{\Psi}_{\delta t} - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)}}{\delta t^{i+1}} + \mathcal{O}(\delta t^{i+2}) \\ &= \mathcal{O}(\delta t^{i+2}) \,. \end{aligned}$$

Thus eqs. (2.3) and (2.4) recursively define the modified vector fields $\boldsymbol{X}_i(\delta t)$, $i = 1, \ldots, \infty$. The recursion is started with $\boldsymbol{X}_1(\delta t) = \delta t \boldsymbol{Z}$. The generated sequence $\{\boldsymbol{\Delta X}_i\}_{i=1,\ldots,\infty}$ is, of course, equivalent to the sequences obtained by using Taylor series expansions as described in [16, 5].

Throughout this paper, we will exclusively work with the recusion (2.3)–(2.4). In §3, it will be shown that this leads to a simple characterization of the geometric properties of the modified vector fields and, in §4, explicit estimates for the difference between the time-one-flow map of the modified vector field $\boldsymbol{X}_i(\delta t)$ and the numerical method will be given. We like point out that these results can also be (and have been [16, 5, 18, 20, 13]) derived using explicit Taylor series expansion of the flow map and the numerical method. However, we feel that the application of the recursion (2.3)–(2.4) leads to a simplification in the presentation of these results.

3. Geometric Properties of Backward Error Analysis. In this section, we consider differential equations (2.1) whose corresponding vector field $\boldsymbol{Z}$ belongs to a certain linear subspace $\mathfrak{g}$ of the infinite dimensional Lie algebra[4] of smooth vector fields on R^n [21],[1].

ASSUMPTION. *Given a linear subspace $\mathfrak{g}$ of the infinite dimensional Lie algebra of smooth vector fields on R^n, let us assume that there is a corresponding subset $\mathfrak{G}$ of the infinite dimensional Frechet manifold [21] of diffeomorphisms on R^n such that*

$$\mathfrak{g} = T_{\mathbf{id}} \mathfrak{G} \,.$$

[4]The algebraic operation is the Lie bracket $[\boldsymbol{X}, \boldsymbol{Y}]$ of two vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$ [3].

Here $T_{\mathbf{id}}\,\mathfrak{G}$ is defined as the set of all vector fields $\boldsymbol{X} := \partial_\tau\,[\boldsymbol{\Psi}_\tau]_{\tau=0}$ for which the one-parametric family of diffeomorphisms $\boldsymbol{\Psi}_\tau \in \mathfrak{G}$ is smooth in τ and $\boldsymbol{\Psi}_{\tau=0} = \mathbf{id}$.

For the linear space (Lie algebra) of Hamiltonian vector fields on R^n this is, for example, the subset of canonical transformations [1]. An important aspect of those differential equations is that the corresponding flow map $\boldsymbol{\Phi}_{t,\boldsymbol{Z}}$ forms a one-parametric subgroup in $\mathfrak{G}$ [21],[1]. Especially in the context of long term integration, it is desirable to discretize differential equations of this type in such a way that the corresponding iteration map $\boldsymbol{\Psi}_{\delta t}$ belongs to the same subset $\mathfrak{G}$ as $\boldsymbol{\Phi}_{t,\boldsymbol{Z}}$. We will call those integrators *geometric integrators.*

The following result concerning the backward error analysis of geometric integrators has been first stated in the unpublished report [26]:

THEOREM 3.1. *Let us assume that the vector field $\boldsymbol{Z}$ in*

$$\frac{d}{dt}\,\boldsymbol{x} = \boldsymbol{Z}(\boldsymbol{x})$$

belongs to a linear subspace $\mathfrak{g}$ of the Lie algebra of all smooth vector fields on R^n. Let us assume furthermore that

$$\boldsymbol{x}_{n+1} = \boldsymbol{\Psi}_{\delta t}\,(\boldsymbol{x}_n) \;=\; \boldsymbol{x}_n + \delta t\,\boldsymbol{\psi}(\boldsymbol{x}_n,\delta t)$$

is a geometric integrator for this subspace $\mathfrak{g}$, i.e., $\boldsymbol{\Psi}_{\delta t} \in \mathfrak{G}$ for all $\delta t \geq 0$ sufficiently small. Then the perturbed vector fields $\boldsymbol{X}_i(\delta t)$, $i = 1,\ldots,\infty$, defined through the recursion (2.3)–(2.4) belong to $\mathfrak{g}$, i.e.

$$\boldsymbol{X}_i(\delta t) \in \mathfrak{g}\,.$$

Proof. The statement is certainly true for $\boldsymbol{X}_1(\delta t) = \delta t\,\boldsymbol{Z}$. Let us assume that it also holds for $\boldsymbol{X}_i(\delta t)$, i.e., $\boldsymbol{X}_i(\delta t) \in \mathfrak{g}$ for all $\delta t \geq 0$ sufficiently small. Since

$$\boldsymbol{\Psi}_{\delta t}\,(\boldsymbol{x}) = \boldsymbol{x} + \delta t\,\boldsymbol{\psi}(\boldsymbol{x},\delta t) \;\in\; \mathfrak{G}$$

and

$$\boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)} \in \mathfrak{G}$$

for all $\delta t \geq 0$ sufficiently small as well as

$$\boldsymbol{\Psi}_{\delta t=0} = \boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t=0)} \;=\; \mathbf{id},$$

we have

$$\boldsymbol{\Delta X}_{i+1} = \lim_{\delta t\to 0}\,\frac{\boldsymbol{\Psi}_{\delta t} - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)}}{\delta t^{i+1}} \;\in\; T_{\mathbf{id}}\,\mathfrak{G}\,.$$

and $\boldsymbol{\Delta X}_{i+1} \in \mathfrak{g}$. This implies $\boldsymbol{X}_{i+1}(\delta t) \in \mathfrak{g}$ as required. □

REMARK. Often the linear subspace $\mathfrak{g}$ is, in fact, a subalgebra under the Lie bracket [3]

$$[\boldsymbol{X},\boldsymbol{Y}] := \frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{X}\cdot\boldsymbol{Y} - \frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{Y}\cdot\boldsymbol{X}\,, \tag{3.1}$$

i.e., $\boldsymbol{X},\boldsymbol{Y} \in \mathfrak{g}$ implies $[\boldsymbol{X},\boldsymbol{Y}] \in \mathfrak{g}$. But this property is not needed in Theorem 3.1. ◇

Let us discuss five examples:

EXAMPLE 1. Consider the subspace $\mathfrak{g}$ of all vector fields that preserve a particular first integral $F : \mathrm{R}^n \to \mathrm{R}$. In fact, this space is a subalgebra under the Lie bracket (3.1). In other words

$$\boldsymbol{\partial}_{\boldsymbol{x}} F \cdot \boldsymbol{X} = 0 \tag{3.2}$$

and

$$\boldsymbol{\partial}_{\boldsymbol{x}} F \cdot \boldsymbol{Y} = 0 \tag{3.3}$$

imply that

$$\boldsymbol{\partial}_{\boldsymbol{x}} F \cdot [\boldsymbol{X}, \boldsymbol{Y}] = 0\,. \tag{3.4}$$

To show this we differentiate (3.2) w.r.t. $\boldsymbol{x}$ which gives

$$\boldsymbol{X}^T \cdot \boldsymbol{\partial}_{\boldsymbol{x}\boldsymbol{x}} F + \boldsymbol{\partial}_{\boldsymbol{x}} F \cdot \boldsymbol{\partial}_{\boldsymbol{x}} \boldsymbol{X} = \boldsymbol{0}\,.$$

The same procedure is applied to (3.3). Using these identities and the definition (3.1) in (3.4) yield the desired result. The corresponding subset $\mathfrak{G}$ is given by the F-preserving diffeomorphisms $\boldsymbol{\Psi}$, i.e.

$$F \circ \boldsymbol{\Psi} = F\,.$$

In fact, let $\boldsymbol{\Psi}_\tau$ be a smooth family of F-preserving diffeomorphisms with $\boldsymbol{\Psi}_{\tau=0} = \mathbf{id}$, then $\boldsymbol{X} := \partial_\tau \left[\boldsymbol{\Psi}_\tau\right]_{\tau=0} \in \mathfrak{g}$ since

$$\partial_\tau \left[F \circ \boldsymbol{\Psi}_\tau\right]_{\tau=0} = \boldsymbol{\partial}_{\boldsymbol{x}} F \cdot \boldsymbol{X} = 0\,.$$

Thus, $T_{\mathbf{id}}\,\mathfrak{G} = \mathfrak{g}$ and we can apply Theorem 3.1. In particular, if a numerical method $\boldsymbol{\Psi}_{\delta t}$ satisfies

$$F \circ \boldsymbol{\Psi}_{\delta t} = F\,,$$

then the modified vector fields $\boldsymbol{X}_i(\delta t)$, $i = 1, \ldots, \infty$, possess F as a first integral. The same result was recently derived by GONZALEZ, HIGHAM & STUART [13] using a contradiction argument. $\diamond$

EXAMPLE 2. Consider the Lie subalgebra of all divergence-free vector fields $\boldsymbol{Z}$, i.e. $\operatorname{div} \boldsymbol{Z} = 0$. The corresponding subset $\mathfrak{G}$ are the volume preserving diffeomorphisms, i.e.

$$\det \left[\frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}} \boldsymbol{\Psi}(\boldsymbol{x})\right] = 1$$

Again we have $T_{\mathbf{id}}\,\mathfrak{G} = \mathfrak{g}$. Namely:

$$\begin{aligned} 0 &= \partial_\tau \det \left[\frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}} \boldsymbol{\Psi}_\tau(\boldsymbol{x})\right]_{\tau=0} \\ &= \operatorname{trace} \left[\boldsymbol{\partial}_{\boldsymbol{x}} \partial_\tau \boldsymbol{\Psi}_\tau(\boldsymbol{x})\right]_{\tau=0} \\ &= \operatorname{div} \boldsymbol{X}(\boldsymbol{x}), \end{aligned}$$

$\boldsymbol{X} := \partial_\tau \left[\boldsymbol{\Psi}_\tau\right]_{\tau=0}$. Thus, if the numerical method $\boldsymbol{\Psi}_{\delta t}$ is volume conserving, then the modified vector fields $\boldsymbol{X}_i(\delta t)$, $i = 1, \ldots, \infty$, are divergence-free. Again, the same result has been formulated by GONZALEZ, HIGHAM & STUART [13] using a contradiction argument. $\diamond$

EXAMPLE 3. Let an involution[5] $\boldsymbol{S} \in \mathrm{R}^{n\times n}$ be given and consider the subspace $\mathfrak{g}$ of vector fields $\boldsymbol{Z}$ on R^n that satisfy the time-reversal symmetry

$$-\boldsymbol{Z}(\boldsymbol{x}) = \boldsymbol{S}\boldsymbol{Z}(\boldsymbol{S}\boldsymbol{x}).$$

This subspace is *not* a subalgebra under the Lie bracket (3.1). The corresponding subset $\mathfrak{G}$ is given by the time-reversible diffeomorphisms $\boldsymbol{\Psi}$, i.e. $\boldsymbol{\Psi}^{-1}(\boldsymbol{x}) = \boldsymbol{S}\boldsymbol{\Psi}(\boldsymbol{S}\boldsymbol{x})$. Let $\boldsymbol{\Psi}_\tau \in \mathfrak{G}$ be smooth in τ with $\boldsymbol{\Psi}_{\tau=0} = \mathbf{id}$, then

$$\begin{aligned} \mathbf{0} &= \partial_\tau \left[\boldsymbol{S}\boldsymbol{\Psi}_\tau \circ \boldsymbol{S} - [\boldsymbol{\Psi}_\tau]^{-1}\right]_{\tau=0}, \\ &= \boldsymbol{S}\,\boldsymbol{X} \circ \boldsymbol{S} + \boldsymbol{X} \end{aligned}$$

which implies that $\boldsymbol{X} := \partial_\tau\,[\boldsymbol{\Psi}_\tau]_{\tau=0} \in \mathfrak{g}$. It follows that $T_{\mathbf{id}}\,\mathfrak{G} = \mathfrak{g}$ and we can apply Theorem 3.1. Thus, if a numerical method $\boldsymbol{\Psi}_{\delta t}$ satisfies the time-reversal symmetry, then the modified vector fields $\boldsymbol{X}_i(\delta t)$, $i = 1,\dots,\infty$, are time-reversible. This result has been first proven by HAIRER & STOFFER in [20] using ideas from [26]. $\diamond$

EXAMPLE 4. Let $\{.,.\}$ denote the Poisson bracket of a (linear) Poisson manifold $\mathcal{P} = \mathrm{R}^n$. Then the Lie algebra of Hamiltonian vector fields on $\mathcal{P}$ is given by

$$\frac{d}{dt}\boldsymbol{x} = \{\,\mathbf{id},\, H\,\}\,(\boldsymbol{x})$$

where $H : \mathcal{P} \to \mathrm{R}$ is a smooth function. The corresponding subset $\mathfrak{G}$ is given by the set of smooth diffeomorphisms on $\mathcal{P}$ that preserve the Poisson bracket $\{.,.\}$ [1]. Let $\boldsymbol{\Psi}_\tau$ be a family of maps in $\mathfrak{G}$ with $\boldsymbol{\Psi}_{\tau=0} = \mathbf{id}$. Then

$$\begin{aligned} 0 &= \partial_\tau\left[\{F \circ \boldsymbol{\Psi}_\tau, G \circ \boldsymbol{\Psi}_\tau\} - \{F, G\}\right]_{\tau=0} \\ &= \{F, \boldsymbol{\partial_x} G \cdot \boldsymbol{X}\} + \{\boldsymbol{\partial_x} F \cdot \boldsymbol{X}, G\} \end{aligned}$$

for all smooth functions $F, G : \mathcal{P} \to \mathrm{R}$, $\boldsymbol{X} := \partial_\tau\,[\boldsymbol{\Psi}_\tau]_{\tau=0}$. This is the condition for a vector field $\boldsymbol{X}$ to be locally Hamiltonian. Since $\mathcal{P}$ is simply connected, the vector field is also globally Hamiltonian [3].

If the discrete evolution (2.2) satisfies $\boldsymbol{\Psi}_{\delta t} \in \mathfrak{G}$ for all $\delta t > 0$, then $\boldsymbol{\Psi}_{\delta t}$ is called a *symplectic method* and it follows from Theorem 3.1 that the modified vector fields $\boldsymbol{X}_i(\delta t)$, $i = 1,\dots,\infty$, are Hamiltonian vector fields on $\mathcal{P}$. This result can also be found in [5, 16, 26].

If a symplectic method can be expanded as a P-series, then the vector fields $\boldsymbol{X}_i(\delta t)$ are globally Hamiltonian even if the phase space $\mathcal{P} \subset \mathrm{R}^n$ is not simply connected [16]. This result applies to all symplectic Runge-Kutta and partitioned Runge-Kutta methods. Furthermore, symplectic methods defined by a generating function of the third kind [32] are also always globally Hamiltonian [5]. The same statement is true for symplectic methods based on the composition of exact flow maps [40]. $\diamond$

EXAMPLE 5. Let us now consider differential equations on a matrix Lie group $\mathcal{G} \subset \mathrm{R}^{n\times n}$ [35]. In general, time independent differential equations on $\mathcal{G}$ can be written in the form

$$\frac{d}{dt}\boldsymbol{Y} = \boldsymbol{A}(\boldsymbol{Y})\,\boldsymbol{Y}$$

[5] An involution is a non-singular matrix that satisfies $\boldsymbol{S}^{-1} = \boldsymbol{S}$.

where $\boldsymbol{A} : \mathcal{G} \to \mathbf{g}$, $\mathbf{g} \subset \mathrm{R}^{n\times n}$ the Lie algebra of $\mathcal{G}$. Many recent papers (see [8] and references therein) have been devoted to methods that preserve the Lie group structure, i.e.

$$\boldsymbol{Y}_{n+1} = \boldsymbol{\Psi}_{\delta t}(\boldsymbol{Y}_n)$$

and $\boldsymbol{Y}_n \in \mathcal{G}$ implies $\boldsymbol{Y}_{n+1} \in \mathcal{G}$. Thus, $\boldsymbol{\Psi}_{\delta t}$ is a diffeomorphism defined on the submanifold $\mathcal{G} \subset \mathrm{R}^{n\times n}$. In fact, this submanifold can be characterized, at least locally, by a set of nonlinear equations which we denote by $\boldsymbol{F}(\boldsymbol{Y}) = \boldsymbol{0}$. Thus $\boldsymbol{\Psi}_{\delta t}$ is an $\boldsymbol{F}$ integral preserving map, i.e. $\boldsymbol{F} \circ \boldsymbol{\Psi}_{\delta t} = \boldsymbol{F}$ on $\mathcal{G}$. Following Example 1, we know then that the modified vector fields (as well as the given vector field) satisfy $\boldsymbol{\partial}_{\boldsymbol{Y}} \boldsymbol{F} \cdot \boldsymbol{X}_i(\delta t) = \boldsymbol{0}$. Hence the modified vector fields $\boldsymbol{X}_i(\delta t)$ are vector fields on $\mathcal{G}$ and give rise to modified differential equations of type

$$\frac{d}{dt}\boldsymbol{Y} = \tilde{\boldsymbol{A}}_i(\boldsymbol{Y};\delta t)\,\boldsymbol{Y}\,,$$

with

$$\boldsymbol{X}_i(\boldsymbol{Y};\delta t) = \delta t\,\tilde{\boldsymbol{A}}_i(\boldsymbol{Y};\delta t)\boldsymbol{Y}\,.$$

and $\tilde{\boldsymbol{A}}_i(\delta t) : \mathcal{G} \to \mathbf{g}$. See [28] for further results on backward error analysis for numerical methods on manifolds. $\diamond$

4. Truncation Error of Backward Error Analysis. We like to derive an explicit estimate for the norm of the vector fields $\boldsymbol{\Delta X}_{i+1}$ and the difference between the time-one-flow map $\boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)}$ and the numerical approximation $\boldsymbol{\Psi}_{\delta t}$, $i = 1,\ldots,\infty$. To do so we assume from now on that the vector field $\boldsymbol{Z}$ in (2.1) is real analytic. We also introduce the following notations: Let $\mathcal{B}_r(\boldsymbol{x}_0) \subset \mathrm{C}^n$ denote the complex ball of radius $r > 0$ around $\boldsymbol{x}_0 \in \mathrm{R}^n$ and define

$$||\,\boldsymbol{z}\,|| := \max_{i=1,\ldots,n} |z_i|\,, \qquad (\boldsymbol{z} \in \mathrm{C}^n)\,.$$

Let us consider a compact subset $\mathcal{K} \subset \mathrm{R}^n$ of phase space and a constant $r > 0$ such that a given real analytic vector field $\boldsymbol{Y}$ is bounded on $\mathcal{B}_r(\boldsymbol{x}_0)$ for all $\boldsymbol{x}_0 \in \mathcal{K}$. Then we define

$$||\,\boldsymbol{Y}\,||_r = \sup_{\boldsymbol{x}\in B_r\mathcal{K}} ||\,\boldsymbol{Y}(\boldsymbol{x})\,||$$

with

$$\mathcal{B}_r\mathcal{K} := \bigcup_{\boldsymbol{x}_0\in\mathcal{K}} \mathcal{B}_r(\boldsymbol{x}_0)\,.$$

We also define $\mathcal{B}_0\mathcal{K} = \mathcal{K}$ and

$$||\,\boldsymbol{Y}\,||_0 = \sup_{\boldsymbol{x}\in\mathcal{K}} ||\,\boldsymbol{Y}(\boldsymbol{x})\,||$$

To find an estimate for $\boldsymbol{\Delta X}_{i+1}$, as defined in (2.4), we need estimates for the mappings appearing on the right hand side of (2.4). We start with an estimate for the map $\boldsymbol{\Psi}_{\delta t}$.

LEMMA 4.1. *Let us assume that the vector field $\boldsymbol{Z}$ in (2.1) is real analytic and that there is a compact subset $\mathcal{K}$ of phase space and constants K, $R > 0$ such that*

$$|| \boldsymbol{Z} ||_R \leq K \,.$$

We also assume that the numerical method $\boldsymbol{\Psi}_{\delta t}$ is real analytic. Then there exists a constant $M \geq K$ such that

$$|| \boldsymbol{\Psi}_\tau - \mathbf{id} ||_{\alpha R} \leq |\tau| \, M \;\leq\; (1-\alpha) \, R \qquad \textit{for} \qquad |\tau| \;\leq\; \frac{(1-\alpha)\, R}{M} \,, \tag{4.1}$$

$\alpha \in [0,1)$.

Proof. Under the given assumptions, the flow map

$$\boldsymbol{\Phi}_{\tau,\boldsymbol{Z}}(\boldsymbol{x}) = \boldsymbol{x} + \int_0^\tau \boldsymbol{Z}(\boldsymbol{\Phi}_{t,\boldsymbol{Z}}(\boldsymbol{x}))\, dt$$

is defined for complex valued $\tau \in \mathrm{C}$ where the integral on the right hand side is independent of the path from zero to τ. The complexified flow map satisfies

$$\begin{aligned} || \boldsymbol{\Phi}_{\tau,\boldsymbol{Z}} - \mathbf{id} ||_{\alpha R} &\leq \sup_{\boldsymbol{x} \in B_{\alpha R}\mathcal{K}} \int_0^\tau ||\boldsymbol{Z}(\boldsymbol{\Phi}_{t,\boldsymbol{Z}}(\boldsymbol{x}))|| \, |dt| \\ &\leq |\tau| \, K \;\leq\; (1-\alpha)\, R \qquad \text{for} \qquad |\tau| \;\leq\; \frac{(1-\alpha)\, R}{K} \,, \end{aligned} \tag{4.2}$$

$\alpha \in [0,1)$. Consistency of the numerical method implies that there exists a constant $\Delta K > 0$ such that

$$|| \boldsymbol{\Psi}_\tau - \mathbf{id} ||_{\alpha R} \leq |\tau| \, (K + \Delta K) \;\leq\; (1-\alpha)\, R \,, \qquad |\tau| \;\leq\; \frac{(1-\alpha)\, R}{K + \Delta K} \,,$$

for the (complexified) map $\boldsymbol{\Psi}_\tau$. Take $M := K + \Delta K$. □

REMARK. Let us consider a s-stage Runge-Kutta method with coefficients $\{a_{ij}\}_{i,j=1,\ldots,s}$ and $\{b_i\}_{i=1,\ldots,s}$ [19] satisfying

$$\sum_{j=1}^s |a_{ij}| \;\leq\; d \qquad \text{and} \qquad \sum_{i=1}^s |b_i| \;\leq\; d$$

$d \geq 1$, and assume that the Runge-Kutta method uniquely[6] defines a real analytic map $\boldsymbol{\Psi}_\tau$ for all step-sizes $\tau \in \mathrm{C}$ with $|\tau| \leq R/K$. Then we have $M = d\, K$ in Lemma 4.1. This follows from the fact that, under the stated assumptions, all stage variables will be in $B_R\mathcal{K}$, where the vector field $\boldsymbol{Z}$ is bounded by the constant K. A similar statement holds for partitioned Runge-Kutta methods. $\diamond$

THEOREM 4.2. *Let the assumptions of Lemma 4.1 be satisfied. Then there exists a family of real analytic vector fields $\tilde{\boldsymbol{X}}(\delta t) : \mathcal{V} \subset R^n \to R^n$, $\mathcal{K} \subset \mathcal{V} \subset \mathcal{U}$, such that*

$$|| \boldsymbol{\Psi}_{\delta t} - \boldsymbol{\Phi}_{\delta t, \tilde{\boldsymbol{X}}(\delta t)} ||_0 \leq 8 \, \delta t \, b \, M \, e^{-p} \, e^{-\gamma/\delta t}$$

with $\gamma = R/(cMe)$, $b = 20$, $c = 300$, and $p \geq 1$ the order of the method. The family of modified vector fields $\tilde{\boldsymbol{X}}(\delta t)$ satisfies the estimate

$$|| \tilde{\boldsymbol{X}}(\delta t) - \boldsymbol{Z} ||_0 \leq 2 \, d_p \, b \, M \, \left(\frac{c \, \delta t \, M}{R} \right)^p$$

[6]For an implicit method, the solution can be obtained by fixpoint iteration if $|\tau|$ is sufficiently small.

with $d_p \geq 1$ a constant depending on the order p of the method. For example, $d_1 = 0.8$, $d_2 = 1.5$, $d_3 = 3.7$, and $d_4 = 12.8$.

Proof. We know that $\boldsymbol{X}_1(\delta t) = \delta t \boldsymbol{Z}$ and that $\boldsymbol{\Psi}_{\delta t}$ is a method of order $p \geq 1$. Thus $||\boldsymbol{X}_1(\tau)||_R \leq |\tau| M$ and $\boldsymbol{\Delta X}_i = \boldsymbol{0}$ for $i = 2, \ldots, p$. Next we find an estimate for the difference between the time-one-flow map $\boldsymbol{\Phi}_{1,\boldsymbol{X}_1(\tau)}$ and the map $\boldsymbol{\Psi}_\tau$. Using (4.1) and (4.2), we obtain

$$\begin{aligned} ||\, \boldsymbol{\Phi}_{1,\boldsymbol{X}_1(\tau)} - \boldsymbol{\Psi}_\tau \,||_{\alpha R} &= ||\, \boldsymbol{\Phi}_{1,\boldsymbol{X}_1(\tau)} - \mathbf{id} + \mathbf{id} - \boldsymbol{\Psi}_\tau \,||_{\alpha R} \\ &\leq ||\, \boldsymbol{\Phi}_{\tau,\boldsymbol{Z}} - \mathbf{id} \,||_{\alpha R} + ||\, \boldsymbol{\Psi}_\tau - \mathbf{id} \,||_{\alpha R} \\ &\leq 2\,(1-\alpha)\,R \end{aligned}$$

for $|\tau| \leq \tau_1 := (1-\alpha)R/M$. Since the mappings are real analytic and their difference is $\mathcal{O}(\delta t^{p+1})$, we obtain the estimate [5]

$$\begin{aligned} ||\, \boldsymbol{\Phi}_{1,\boldsymbol{X}_1(\delta t)} - \boldsymbol{\Psi}_{\delta t} \,||_{\alpha R} &\leq 2\,(1-\alpha)\,R\, \left(\frac{\delta t}{\tau_1}\right)^{p+1} \\ &\leq 2\,\delta t\, M\, \left(\frac{\delta t\, M}{(1-\alpha)\,R}\right)^p , \end{aligned}$$

$\alpha \in [0,1)$. Using this in (2.4) with $i = p$, we obtain

$$||\, \boldsymbol{\Delta X}_{p+1} \,||_{\alpha R} \leq 2\, M\, \left(\frac{M}{(1-\alpha)\,R}\right)^p . \tag{4.3}$$

Next we show that

$$||\, \boldsymbol{\Delta X}_i \,||_{\alpha R} \leq b\, M\, \left(\frac{c\,(i-p)\,M}{(1-\alpha)\,R}\right)^{i-1} \tag{4.4}$$

for $i \geq p+1$ with $b = 20$, $c = 300$, and $\alpha \in [0,1)$. The estimate is true for $i = p+1$ (compare (4.3)). We proceed by induction. First note that, for $j \geq p+2$,

$$\begin{aligned} ||\, \boldsymbol{X}_j(\tau) \,||_{\alpha R} &\leq |\tau|\, ||\, \boldsymbol{Z} \,||_{\alpha R} \;+ \sum_{i=p+1}^{j} |\tau|^i\, ||\, \boldsymbol{\Delta X}_i \,||_{\alpha R} \\ &\leq |\tau| M \left[1 + 2\left(\frac{|\tau| M}{(1-\alpha)R}\right)^p + \sum_{i=p+2}^{j} b\, \left(\frac{c(i-p)|\tau| M}{(1-\alpha)R}\right)^{i-1} \right], \end{aligned} \tag{4.5}$$

$\alpha \in [0,1)$. We replace the parameter $\alpha \in [0,1)$ in this formula by $\alpha + \delta_j(1-\alpha) \in [\delta_j, 1)$ where

$$\delta_j := \frac{b-1}{(j-p+1)\,c} \;=\; \frac{0.067}{j-p+1} .$$

This yields

$$||\, \boldsymbol{X}_j(\tau) \,||_{(\alpha+\delta_j(1-\alpha))R} \leq (b-1)\,\tau_j\, M \;=\; \delta_j\,(1-\alpha)\,R ,$$

for all $\alpha \in [0,1)$ and all $\tau \in \mathrm{C}$ with

$$|\tau| \leq \frac{(1-\alpha)\,R}{(j-p+1)\,c\,M} \;=:\; \tau_j .$$

Here we have used that

$$\sum_{i=p+2}^{j} \left(\frac{i-p}{(1-\delta_j)\,(j-p+1)} \right)^{i-1} \leq 0.891 \tag{4.6}$$

for $p \geq 1$ and all $j \geq p+2$. In particular, substitute τ_j for $|\tau|$ and $\alpha + \delta_j(1-\alpha)$ for α in (4.5). Then use the identity $1-\alpha-\delta_j(1-\alpha) = (1-\delta_j)(1-\alpha)$ and the inequality (4.6) to derive

$$\tau_j\, M \left[1 + 2\left(\frac{1}{c\,\beta_j}\right)^p + b \sum_{i=p+2}^{j} \left(\frac{i-p}{\beta_j}\right)^{i-1} \right] \leq (b-1)\,\tau_j\, M,$$

$\beta_j := (1-\delta_j)(j-p+1)$. Next we introduce the vector-valued, real analytic function

$$\begin{aligned} \boldsymbol{f}_\tau(\boldsymbol{x}) &:= \boldsymbol{\Phi}_{1,\boldsymbol{X}_j(\tau)}(\boldsymbol{x}) - \boldsymbol{x}\,, \\ &= \int_0^1 \boldsymbol{X}_j\,(\boldsymbol{\Phi}_{t,\boldsymbol{X}_j(\tau)}(\boldsymbol{x});\delta t)\,dt\,, \end{aligned}$$

and observe that

$$||\,\boldsymbol{f}_\tau\,||_{\alpha R} \leq (b-1)\,\tau_j\,M \;=\; \delta\,(1-\alpha)\,R \tag{4.7}$$

for $|\tau| \leq \tau_j$ and $\alpha \in [0,1)$. Here we have used that $\boldsymbol{x} \in \mathcal{B}_{\alpha R}\mathcal{K}$ implies $\boldsymbol{\Phi}_{t,\boldsymbol{X}_j(\tau)}(\boldsymbol{x}) \in \mathcal{B}_{(\alpha+\delta_j(1-\alpha))R}\mathcal{K}$ for all $0 \leq |t| \leq 1$ and any $|\tau| \leq \tau_j$. Now we can find an estimate for the difference between the time-one-flow map $\boldsymbol{\Phi}_{1,\boldsymbol{X}_j(\tau)}$ and the map $\boldsymbol{\Psi}_\tau$. Using (4.1) and (4.7), we obtain

$$\begin{aligned} ||\,\boldsymbol{\Phi}_{1,\boldsymbol{X}_j(\tau)} - \boldsymbol{\Psi}_\tau\,||_{\alpha R} &\leq ||\,\boldsymbol{f}_\tau\,||_{\alpha R} + ||\,\boldsymbol{\Psi}_\tau - \mathbf{id}\,||_{\alpha R} \\ &\leq (b-1)\,\tau_j\,M + \tau_j\,M \\ &\leq b\,\tau_j\,M \end{aligned}$$

for $|\tau| \leq \tau_j$. Since the mappings are real analytic and their difference is $\mathcal{O}(\delta t^{j+1})$, we obtain the estimate [5]

$$\begin{aligned} ||\,\boldsymbol{\Phi}_{1,\boldsymbol{X}_j(\delta t)} - \boldsymbol{\Psi}_{\delta t}\,||_{\alpha R} &\leq b\,\tau_j\,M\,\left(\frac{\delta t}{\tau_j}\right)^{j+1} \\ &\leq b\,\delta t\,M\,\left(\frac{c\,(j-p+1)\,\delta t\,M}{(1-\alpha)\,R}\right)^j\,, \end{aligned} \tag{4.8}$$

$\alpha \in [0,1)$. Using this in (2.4) with $i = j$, we finally obtain

$$||\,\boldsymbol{\Delta X}_{j+1}\,||_{\alpha R} \leq b\,M\left(\frac{c\,(j-p+1)\,M}{(1-\alpha)\,R}\right)^j$$

which verifies (4.4) for $i = j+1$.

Next we need an estimate for the difference between the time-one-flow map $\boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)}$ and the map $\boldsymbol{\Psi}_{\delta t}$ on the compact set $\mathcal{K}$. Using (4.8) with $\alpha = 0$ and $i = j$, we immediately have

$$||\,\boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\delta t)} - \boldsymbol{\Psi}_{\delta t}\,||_0 \leq \delta t\,b\,M\,\left(\frac{c\,\delta t\,(i-p+1)\,M}{R}\right)^i .$$

The family of vector fields $\tilde{\boldsymbol{X}}(\delta t)$ is now defined by taking an optimal number $i_*(\delta t)$ of iterations. We take $i_*(\delta t)$ as the integer part of

$$i_o(\delta t) := \frac{R}{c\,\delta t\,M\,e} + p - 1\,.$$

Thus

$$\begin{aligned} ||\,\boldsymbol{\Phi}_{\tau,\boldsymbol{X}_{i_*}(\delta t)} - \boldsymbol{\Psi}_{\delta t}\,||_0 &\leq \delta t\,b\,M\,e^{-i_*} \\ &\leq \delta t\,b\,M\,e^{-i_o+1} \\ &\leq 8\,\delta t\,b\,M\,e^{-p}\,e^{-\gamma/\delta t}\,, \end{aligned}$$

$\gamma = R/(cMe)$. We define $\tilde{\boldsymbol{X}}(\delta t) := \delta t^{-1}\boldsymbol{X}_{i_*}(\delta t)$. This completes the first part of the proof.

According to (4.5), the difference between the modified vector fields $\tilde{\boldsymbol{X}}(\delta t)$ and $\boldsymbol{Z}$ is given by

$$||\,\tilde{\boldsymbol{X}}(\delta t) - \boldsymbol{Z}\,||_0 \leq M\,\left(\frac{c\,\delta t\,M}{R}\right)^p \left[\frac{2}{c^p} + \sum_{i=p+2}^{i_*} b\,(i-p)^p\,\left(\frac{c\,(i-p)\,\delta t\,M}{R}\right)^{i-p-1}\right]\,.$$

Next we use

$$\delta t \leq \frac{R}{c\,(i_* - p + 1)\,M\,e}$$

to obtain

$$\begin{aligned} ||\,\tilde{\boldsymbol{X}}(\delta t) - \boldsymbol{Z}\,||_0 &\leq M\,\left(\frac{c\,\delta t\,M}{R}\right)^p \left[0.0067 + b\,\sum_{i=p+2}^{i_*} \frac{(i-p)^p}{e^{i-p-1}}\,\left(\frac{i-p}{i_* - p + 1}\right)^{i-p-1}\right]\,, \\ &\leq M\,\left(\frac{c\,\delta t\,M}{R}\right)^p [0.0067 + b\,d_p\,1.38]\,, \\ &\leq 2\,d_p\,b\,M\,\left(\frac{c\,\delta t\,M}{R}\right)^p\,. \end{aligned}$$

Here $d_p \geq 1$ is chosen such that

$$d_p \geq \frac{j^p}{e^{j-1}}$$

for all $j \geq 2$. □

REMARK. The proof of Theorem 4.2 is similar in spirit to the one given by BENETTIN & GIORGILLI [5] on the exponentially small difference between an optimal interpolating vector field and a near-to-the-identity map. However, there are a couple of important differences: (i) We explicitly take the order of a method into account. (ii) We directly derive estimates on the difference between the flow maps $\boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\tau)}$ and $\boldsymbol{\Psi}_{\delta t}$ instead of using Taylor series expansions of $\boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\tau)}$ and $\boldsymbol{\Psi}_{\delta t}$ and corresponding estimates for the elements in the series. We believe that this simplifies the proof of Theorem 4.2. (iii) By introducing the parameter $\alpha \in [0,1)$, we do not have to shrink the domain of definition of the vector fields $\boldsymbol{X}_i(\tau)$ as the iteration index i increases. Again we feel that this simplifies the proof. (iv) As in [18], we work directly with an estimate for the given vector field $\boldsymbol{Z}$ instead of making assumptions on the map $\boldsymbol{\Psi}_{\delta t}$. The rather pessimistic constants $c = 300$ and $b = 20$ entering the estimates seem the main disadvantage of our approach. ◇

A more elaborate version of the proof of Theorem 4.2 can be found in [30].

5. An Application: Ergodic Hamiltonian Systems. Let us consider a (real analytic) Hamiltonian system

$$\frac{d}{dt}\boldsymbol{q} = \boldsymbol{M}^{-1}\boldsymbol{p}\,, \tag{5.1}$$

$$\frac{d}{dt}\boldsymbol{p} = -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q})\,, \tag{5.2}$$

$\boldsymbol{q}, \boldsymbol{p} \in \mathrm{R}^n$, together with a smooth function $A : \mathrm{R}^{2n} \to \mathrm{R}$. We are interested in evaluating the time-average of A along a trajectory $(\boldsymbol{q}(t), \boldsymbol{p}(t))$ of the Hamiltonian system (5.1)-(5.2), i.e

$$\langle A \rangle_T := \frac{1}{T}\int_0^T A(\boldsymbol{q}(t), \boldsymbol{p}(t))\, dt\,, \qquad T \gg 1\,.$$

We assume that

$$\langle A \rangle_\infty := \lim_{T\to\infty} \langle A \rangle_T$$

exists and is equal to the micro-canonical ensemble average corresponding to the Hamiltonian

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

i.e., we assume that the system (5.1)-(5.2) is ergodic[7] (or even mixing) [37]. Thus

$$\langle A \rangle_\infty = \frac{\int A(\boldsymbol{q},\boldsymbol{p})\, \delta(E - H(\boldsymbol{q},\boldsymbol{p}))\, \boldsymbol{dqdp}}{\int \delta(E - H(\boldsymbol{q},\boldsymbol{p}))\, \boldsymbol{dqdp}} \;=:\; \frac{1}{C}\langle A, \delta(E-H)\rangle$$

with $E = H(\boldsymbol{q}(0), \boldsymbol{p}(0))$, $\delta(x)$ Dirac's delta distribution,

$$C := \int \delta(E - H(\boldsymbol{q},\boldsymbol{p}))\, \boldsymbol{dqdp}\,,$$

and

$$\langle A, \delta(E-H)\rangle := \int A(\boldsymbol{q},\boldsymbol{p})\, \delta(E - H(\boldsymbol{q},\boldsymbol{p}))\, \boldsymbol{dqdp}$$

the inner product of A and $\delta(E-H)$.

Let us write the equations (5.1)-(5.2) in more compact form as

$$\frac{d}{dt}\boldsymbol{x} = \boldsymbol{J}\boldsymbol{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x}) \;=\; \{\mathbf{id}, H\}(\boldsymbol{x})\,,$$

$\boldsymbol{x} := (\boldsymbol{q}^T, \boldsymbol{p}^T)^T \in \mathrm{R}^{2n}$. The Hamiltonian H is preserved under the flow map $\boldsymbol{\Phi}_{t,H}$. Let us assume that the hypersurface $\mathcal{M}_0$ of constant energy $H = 0$,

$$\mathcal{M}_0 := \{\boldsymbol{x} \in \mathrm{R}^{2n} : H(\boldsymbol{x}) = 0\}\,,$$

[7] To be more precise: Ergodicity of a system implies that the time average is equivalent to the ensemble average except for, at most, a set of initial conditions of measure zero.

is a compact subset of R^{2n}. We also assume that there is a constant $\gamma_1 > 0$ such that $||\boldsymbol{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x})|| > \gamma_1$ for all $\boldsymbol{x} \in \mathcal{M}_0$. This implies that $\mathcal{M}_0$ is a smooth $(2n-1)$ dimensional compact submanifold. Furthermore, the family of hypersurfaces

$$\mathcal{M}_E = \{\boldsymbol{x} \in \mathrm{R}^{2n} : H(\boldsymbol{x}) = E\}, \qquad E \in (-\Delta E, +\Delta E),$$

$\Delta E > 0$ sufficiently small, are smooth and compact as well (in fact diffeomorphic to $\mathcal{M}_0$). We define the open subset $\mathcal{U}$ of phase space by

$$\mathcal{U} := \bigcup_{E \in (-\Delta E, +\Delta E)} \mathcal{M}_E .$$

So far we have made fairly generic assumptions. In the sequel, we become more specific to ensure that the Hamiltonian system (5.1)-(5.2) is ergodic/mixing.

In a first step we construct a Poincaré return map [14]. Let $\psi : \mathcal{U} \to \mathrm{R}$ be a smooth function and $\gamma_2 > 0$ a positive constant such that $|\{\psi, H\}(\boldsymbol{x})| > \gamma_2$ on the level sets

$$\mathcal{S}_s := \{\boldsymbol{x} \in \mathcal{U} : \psi(\boldsymbol{x}) = s\}, \qquad s \in (-\Delta s, +\Delta s),$$

$\Delta s > 0$ sufficiently small. Let us assume that $\mathcal{S}_s$ defines a Poincaré section for each $s \in (-\Delta s, +\Delta s)$ in the following way: For all $\boldsymbol{x} \in \mathcal{S}_s$, there is a positive number $t_p(\boldsymbol{x}) > 0$ such that the solution $\boldsymbol{x}(t)$, $t \geq 0$, with initial condition $\boldsymbol{x}(0) = \boldsymbol{x}$ satisfies $\boldsymbol{x}(t_p) \in \mathcal{S}_s$ and there is no $0 < t_p' < t_p$ such that $\boldsymbol{x}(t_p') \in \mathcal{S}_s$. The positive number $t_p(\boldsymbol{x})$ is called the Poincaré return time of the point $\boldsymbol{x} \in \mathcal{S}_s$. Knowing the Poincaré return time for each $\boldsymbol{x} \in \mathcal{S}_s$, we define the "global" Poincaré map $\boldsymbol{\Pi} : \mathcal{V} \to \mathcal{V}$ by

$$\boldsymbol{\Pi}(\boldsymbol{x}) := \boldsymbol{\Phi}_{t_p(\boldsymbol{x}),H}(\boldsymbol{x})$$

and

$$\boldsymbol{x} \in \mathcal{V} := \bigcup_{s \in (-\Delta s, +\Delta s)} \mathcal{S}_s .$$

We assume that the Poincaré return times $t_p(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{V}$, are bounded by some constant $K > 0$.

We are interested in the solutions on a particular level set of constant energy. For simplicity, we take the level set $\mathcal{M}_0$. Then it is sufficient to consider the "restricted" Poincaré map $\boldsymbol{\Pi}_0$ which is defined as the restriction of $\boldsymbol{\Pi}$ to

$$\mathcal{D} := \mathcal{S}_0 \cap \mathcal{M}_0 .$$

Thus we have reduced the study of the dynamical properties of the Hamiltonian system (5.1)-(5.2) on the energy shell $\mathcal{M}_0$ to the study of the properties of the Poincaré map $\boldsymbol{\Pi}_0$. If $\boldsymbol{\Pi}_0$ is an ergodic (mixing) map, then the Hamiltonian system is ergodic (mixing) on $\mathcal{M}_0$. Note that $\boldsymbol{\Pi}_0$ is volume preserving, i.e. $\det \boldsymbol{\partial}_{\boldsymbol{x}} \boldsymbol{\Pi}_0(\boldsymbol{x}) = 1$.

From now on we assume that $\boldsymbol{\Pi}_0$ is a uniformly hyperbolic map, i.e., for each $\boldsymbol{x} \in \mathcal{D}$, the linearization $\boldsymbol{\partial}_{\boldsymbol{x}} \boldsymbol{\Pi}_0(\boldsymbol{x})$ at $\boldsymbol{x}$ possesses strictly expanding and contracting directions only [14],[36]. The "stochastic" behavior of such a (deterministic) map has been investigated, for example, in [36]. Here we only point out the four main results:

- There is a unique invariant density μ_0 on $\mathcal{D}$ that is invariant under $\boldsymbol{\Pi}_0$. Furthermore, μ_0 is given by the Lebesgue measure on $\mathcal{D}$.

- The autocorrelation function $\langle A \circ [\mathbf{\Pi}_0]^n, A\rangle$ of a Hölder continuous function $A : \mathcal{U} \to \mathrm{R}$ decays exponentially fast, i.e.

$$|\langle A \circ [\mathbf{\Pi}_0]^n, A\rangle - (\langle A, \mu_0\rangle)^2| \le C\,\Lambda^n\,, \qquad 0 < \Lambda < 1\,,$$

  $C > 0$ an appropriate constant.
- The time-averages

$$\langle A\rangle_N = \frac{1}{N}\sum_{i=1}^{N} A(\boldsymbol{x}_i)$$

  of A along trajectories $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$ of $\mathbf{\Pi}_0$ satisfy a central limit theorem.
- The time-average $\langle A\rangle_N$ of A along trajectories of $\mathbf{\Pi}_0$ with initial value $\boldsymbol{x}_0 \in \mathcal{D}$ satisfy a large deviation theorem. To be more specific [39]: Given any $c > 0$ there is a $h(c) > 0$ such that

$$\mu_0\left(\{\boldsymbol{x}_0 \in \mathcal{D} : |\langle A\rangle_N - \langle A, \mu_0\rangle| > c\}\right) \le e^{-Nh(c)} \tag{5.3}$$

  for all large $N \ge 1$.

These results can be proven (see, for example, [36]) by carefully studying the properties of the corresponding Frobenius-Perron operator $\boldsymbol{P}_0 : L^1(\mathcal{D}) \to L^1(\mathcal{D})$ defined by

$$\boldsymbol{P}_0\,\mu := \mu \circ [\mathbf{\Pi}_0]^{-1}$$

$\mu \in L^1(\mathcal{D})$.

Definition 5.1. *We call a Hamiltonian system (5.1)-(5.2) with the above introduced properties* Poincaré hyperbolic. *In particular, we assume (i) that the level sets $\mathcal{M}_E$, $E \in (-\Delta E, +\Delta E)$, of constant energy are compact submanifolds, (ii) that there is a constant $\gamma_1 > 0$ such that $||\mathbf{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x})|| > \gamma_1$ for all $\boldsymbol{x} \in \mathcal{U}$, (iii) that a global Poincaré map $\mathbf{\Pi}$ can be defined on*

$$\mathcal{V} = \bigcup_{s \in (-\Delta s, +\Delta s)} \mathcal{S}_s$$

which is uniformly hyperbolic as a map restricted to $\mathcal{D} = \mathcal{S}_0 \cap \mathcal{M}_0$, (iv) that the Poincaré return times $t_p(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{V}$, are bounded by some constant $K > 0$, and (v) that there is a constant $\gamma_2 > 0$ such that $|\{\psi, H\}(\boldsymbol{x})| > \gamma_2$ on $\mathcal{V}$.

Let $\widetilde{H}$ be a perturbation of H such that

$$|H(\boldsymbol{x}) - \widetilde{H}(\boldsymbol{x})| + ||\mathbf{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x}) - \mathbf{\nabla}_{\boldsymbol{x}} \widetilde{H}(\boldsymbol{x})|| \le \epsilon$$

for all $\boldsymbol{x} \in \mathcal{U}$ and some $\epsilon > 0$. Then we call $\widetilde{H}$ an ϵ-perturbation of H.

Lemma 5.2. *The property of being Poincaré hyperbolic is stable under ϵ-perturbations of the Hamiltonian H provided ϵ is sufficiently small.*

Proof. The assumption $||\mathbf{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x})|| > \gamma_1$ on the level sets $\mathcal{M}_E$ implies that these sets are persistent under small perturbations. Furthermore, there exists a constant $\tilde{\gamma}_2 > 0$ such that $|\{\psi, \tilde{H}\}(\boldsymbol{x})| > \tilde{\gamma}_2$ for a perturbed Hamiltonian $\tilde{H}$ and $\boldsymbol{x} \in \mathcal{V}$. Thus a Poincaré map is also defined for the perturbed Hamiltonian $\tilde{H}$. Uniform hyperbolicity is also stable under small perturbations of the Poincaré map [2]. □

Let us discretize (5.1)-(5.2) by a symplectic (real analytic) integrator $\mathbf{\Psi}_{\delta t}$ of order $p \ge 1$.

ASSUMPTION. *We assume that backward error analysis can be applied on a compact subset* $\mathcal{K}$ *with* $\mathcal{U} \subset \mathcal{K}$. *The corresponding perturbed Hamiltonian is denoted by* $\widetilde{H}(\delta t)$, *i.e., for all* $\boldsymbol{x} \in \mathcal{K}$,

$$|| \, \boldsymbol{\Phi}_{\delta t, \widetilde{H}}(\boldsymbol{x}) - \boldsymbol{\Psi}_{\delta t}(\boldsymbol{x}) \, || \leq \delta t \, d_1 \, e^{-p} \, e^{-d_2/\delta t} \,,$$

$d_1, d_2 > 0$ *appropriate constants. Let the step-size* δt *be sufficiently small such that the perturbed Hamiltonian system is also Poincaré hyperbolic. For simplicity, we shift the modified Hamiltonian* $\widetilde{H}(\delta t)$ *such that* $H(\boldsymbol{x}_0) = \widetilde{H}(\boldsymbol{x}_0; \delta t) = 0$.

Let us introduce a couple of notations for the perturbed system. As for the unperturbed system, we define the compact level sets $\widetilde{\mathcal{M}}_E$ and the open set $\widetilde{\mathcal{U}}$ (replacing H by $\widetilde{H}$ in the definition). Without loss of generality, we can assume that $\widetilde{\mathcal{U}} \subset \mathcal{K}$. Furthermore,

$$\widetilde{\mathcal{S}}_s := \{ \boldsymbol{x} \in \widetilde{\mathcal{U}} : \psi(\boldsymbol{x}) = s \} \,,$$

$s \in (-\Delta s, +\Delta s)$. The corresponding sets $\widetilde{\mathcal{V}}$ and $\widetilde{\mathcal{D}}$ are now defined in the obvious way. Finally, the global Poincaré map $\widetilde{\boldsymbol{\Pi}}$ and the reduced Poincaré map $\widetilde{\boldsymbol{\Pi}}_0$ are introduced as for the unperturbed system. Again, without loss of the generality, we can assume that $\mathcal{D} \subset \widetilde{\mathcal{V}}$.

We extend the discrete time map $\boldsymbol{\Psi}_{\delta t}$ to a map $\boldsymbol{\Psi}_t$, $t \in [0, \delta t]$, by using the exact flow map $\boldsymbol{\Phi}_{t,\widetilde{H}}$ of the modified problem as an interpolation for $t \in [0, \delta t)$. The map is then extended to $t \geq \delta t$ in the obvious way[8] as the composition of k steps with $\boldsymbol{\Psi}_{\delta t}$ and one step with $\boldsymbol{\Phi}_{dt,\widetilde{H}}$ where $t = k\delta t + dt$, $\delta t > dt \geq 0$. Thus, in correspondence with the definition of the global Poincaré map

$$\widetilde{\boldsymbol{\Pi}}(\boldsymbol{x}) := \boldsymbol{\Phi}_{\widetilde{t}_p(\boldsymbol{x}), \widetilde{H}}(\boldsymbol{x}) \,,$$

we define

$$\widehat{\boldsymbol{\Pi}}(\boldsymbol{x}) := \boldsymbol{\Psi}_{\widetilde{t}_p(\boldsymbol{x})}(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \widetilde{\mathcal{V}}$. Here the Poincaré return times $\widetilde{t}_p(\boldsymbol{x})$, $\boldsymbol{x} \in \widetilde{\mathcal{V}}$, are the same as in the definition of $\widetilde{\boldsymbol{\Pi}}$. Lemma 5.2 implies that there is a constant $\widetilde{K} > 0$ such that

$$\sup_{\boldsymbol{x} \in \widetilde{\mathcal{V}}} \widetilde{t}_p(\boldsymbol{x}) \leq \widetilde{K} \,.$$

It follows from backward and forward error analysis [18] that there is a constant $d_3 > 0$ such that

$$|| \, \widetilde{\boldsymbol{\Pi}}(\boldsymbol{x}) - \widehat{\boldsymbol{\Pi}}(\boldsymbol{x}) \, || \leq d_3 \, e^{-p} \, e^{-d_2/\delta t}$$

for all $\boldsymbol{x} \in \widetilde{\mathcal{V}}$ and for all δt sufficiently small. More importantly, let $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$ be a "numerically" computed sequence of points with $\boldsymbol{x}_{i+1} = \widehat{\boldsymbol{\Pi}}(\boldsymbol{x}_i)$ and let $\{\tilde{\boldsymbol{x}}_i\}_{i=1,\dots,N}$ be the corresponding sequence under the map $\widetilde{\boldsymbol{\Pi}}$ with $\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0 \in \widetilde{\mathcal{D}}$. Each sequence $\{\boldsymbol{x}_i\}$, $\{\tilde{\boldsymbol{x}}_i\}$ respectively, generates two sequences of real numbers $\{E_i\}$ and $\{s_i\}$, $\{\widetilde{E}_i\}$ and $\{\widetilde{s}_i\}$ respectively, which are defined by $E_i = \widetilde{H}(\boldsymbol{x}_i)$ and $s_i = \psi(\boldsymbol{x}_i)$, by $\widetilde{E}_i = \widetilde{H}(\tilde{\boldsymbol{x}}_i)$

[8] The resulting map is discontinuous at multiples of the step-size δt. A smooth interpolation could be defined. But this is not needed in the context of this paper.

and $\widetilde{s}_i = \psi(\tilde{\boldsymbol{x}}_i)$ respectively. We obviously have $\widetilde{E}_i = 0$ and $\widetilde{s}_i = 0$ while The "drift" in the values of E_i and s_i per step away from zero is exponentially small and sums up linearly with the number of steps. This energy conserving property of a symplectic method has been discussed by Benettin & Giorgilli [5] and Hairer & Lubich [18]. The same exponentially slow drift follows for the sequence $\{s_i\}$ from

$$\begin{aligned} |\psi(\boldsymbol{x}_N) - \psi(\boldsymbol{x}_0)| &\leq \sum_{i=1}^{N} |\psi(\boldsymbol{x}_i) - \psi(\boldsymbol{x}_{i-1})| \\ &\leq \sum_{i=1}^{N} |\psi(\widehat{\boldsymbol{\Pi}}(\boldsymbol{x}_{i-1})) - \psi(\widetilde{\boldsymbol{\Pi}}(\boldsymbol{x}_{i-1}))| \\ &\leq N\,\lambda\, d_3\, e^{-p}\, e^{-d_2/\delta t}\,, \end{aligned}$$

$\lambda > 0$ the Lipschitz constant of ψ on $\widetilde{\mathcal{V}}$.

In other words, if we start initially on $\widetilde{\mathcal{D}}$, then the points computed "numerically" with the Poincaré map $\widehat{\boldsymbol{\Pi}}$ will stay in an exponentially small neighborhood of $\widetilde{\mathcal{D}}$ over exponentially many iterates of $\widehat{\boldsymbol{\Pi}}$. Now, since our numerical method is of order $p \geq 1$, the compact manifolds $\widetilde{\mathcal{M}}_E$ and $\mathcal{M}_E$ are $\mathcal{O}(\delta t^p)$ away from each other, i.e., the modified Hamiltonian $\widetilde{H}$ is an ϵ-perturbation of the Hamiltonian H with $\epsilon \sim \delta t^p$. Thus the sequence $\{\boldsymbol{x}_i\}$ will also stay in a $\mathcal{O}(\delta t^p)$ neighborhood of $\mathcal{D}$ as long as the number of iterates N satisfies

$$N \leq d_4\, e^{d_2/(2\delta t)}\,, \tag{5.4}$$

$d_4 > 0$ an appropriate constant.

Now the *Shadowing Lemma* [33] is applied to the sequence $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$.

Proposition 5.3. *There exists an exact trajectory $\{\hat{\boldsymbol{x}}_i\}_{i=1,\dots,N}$ of the Poincaré map $\boldsymbol{\Pi}_0$ on $\mathcal{D}$ such that the "numerically" computed sequence $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$ stays in a $\mathcal{O}(\delta t^p)$ neighborhood of the (shadowing) exact trajectory provided the number of iterates N satisfies (5.4).*

Proof. We first project the sequence $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$ down onto $\mathcal{D}$ using a "search direction" orthogonal to the manifold $\mathcal{D}$. Denote the result by $\{\bar{\boldsymbol{x}}_i\}$. The projected sequence $\{\bar{\boldsymbol{x}}_i\}$ and the sequence $\{\boldsymbol{x}_i\}$ are $\mathcal{O}(\delta t^p)$ close to each other provided N satisfies (5.4). The "local" error per step between the "exact" Poincaré map $\boldsymbol{\Pi}$ and the "numerical" Poincaré map $\widehat{\boldsymbol{\Pi}}$ is also of order p in the step-size δt. This follows from standard forward error analysis. Thus the Shadowing Lemma [33] for uniformly hyperbolic maps can be applied to the Poincaré map $\boldsymbol{\Pi}_0 : \mathcal{D} \to \mathcal{D}$ and the projected sequence $\{\bar{\boldsymbol{x}}_i\}$ on $\mathcal{D}$. The shadowing distance is $\mathcal{O}(\delta t^p)$. This shadowing result also applies to the sequence $\{\boldsymbol{x}_i\}$. □

Let us now assume that we want to compute the ensemble average of a smooth function $A : \mathrm{R}^{2n} \to \mathrm{R}$ up to a certain accuracy $c > 0$. The large deviation theorem (5.3) for hyperbolic maps tells us that the probability to obtain the ensemble average in the desired accuracy as the time average along a single trajectory goes to one exponentially fast as the length N of the trajectory is increased. If we numerically compute an approximative trajectory for the system (5.1)-(5.2), then we know from Proposition 5.3 that this trajectory is $\mathcal{O}(\delta t^p)$ close to *some* exact trajectory over exponentially many integration steps N. Let us denote the time average of A along this exact trajectory by $\langle A\rangle_N^e$ and the numerically computed time average by $\langle A\rangle_N$, then

$$\langle A\rangle_N - \langle A\rangle_N^e = \mathcal{O}(\delta t^p) \tag{5.5}$$

for all N satisfying a bound of type (4.4). Thus we obtain the following:

PROPOSITION 5.4. *Let (5.1)-(5.2) be a Poincaré hyperbolic (real-analytic) system which we discretize by a symplectic method of order $p \geq 1$ in the step-size δt. Then the time-average $\langle A \rangle_N$ of an observable A along a "numerically" computed trajectory $\{\boldsymbol{x}_n\}_{n=1,\dots,N}$,*

$$\boldsymbol{x}_{n+1} = \boldsymbol{\Psi}_{\delta t}(\boldsymbol{x}_n)\,,$$

satisfies (5.5) where $\langle A \rangle_N^e$ is the time-average along some exact trajectory and the number of steps N satisfies a bound of type (5.4). Furthermore, assume we want to compute the ensemble average of A within a given accuracy $c > 0$. We assume, for simplicity, that the constant c is larger than the difference between the time averages (5.5) which is always true for sufficiently small step-sizes δt. Then the probability to obtain the average in the desired accuracy as the time average along a numerically computed trajectory goes to one exponentially fast as the number of integration steps N is increased. Taking the maximum number (5.4) of steps, the probability can be made double exponentially close to one in (5.3) as $\delta t \to 0$.

EXAMPLE. As a numerical example, we look at the following planar anisotropic Kepler problem [15]:

$$\frac{d}{dt}\boldsymbol{q} = \boldsymbol{M}^{-1}\boldsymbol{p}\,,$$
$$\frac{d}{dt}\boldsymbol{p} = -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q})\,,$$

$\boldsymbol{q} = (q_x, q_y)^T, \boldsymbol{p} = (p_x, p_y)^T \in \mathrm{R}^2$,

$$V(q_x, q_y) = \frac{-1}{\sqrt{(q_x)^2 + (q_y)^2}}\,,$$

and

$$\boldsymbol{M} = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}.$$

The initial conditions are chosen such that

$$H = \frac{\boldsymbol{p}^T \boldsymbol{M}^{-1}\boldsymbol{p}}{2} + V(\boldsymbol{q}) \;\; = \;\; -\frac{1}{2} \tag{5.6}$$

and $L = p_y q_x - p_x q_y \neq 0$. Note that angular momentum L is not conserved.

We define the Poincaré section $\mathcal{S}_0$ by $\psi = q_y = 0$ and record the sequence of points (q_x, p_x). Conservation of energy implies that the thus defined sequence is restricted to the subset

$$|\, q_x \,| < \frac{2}{1 + (p_x)^2/10}$$

where $-\infty < p_x < +\infty$. This subset has an awkward shape. But it can be transformed into a rectangle by means of the area preserving transformation

$$X_1 := q_x\,(1 + (p_x)^2/10)\,, \tag{5.7}$$
$$X_2 := \sqrt{10}\,\mathrm{arctg}\,(p_x/\sqrt{10}) \tag{5.8}$$

where $|X_1| \leq 2$ and $|X_2| \leq \sqrt{10}\pi/2$. The corresponding Poincaré map is hyperbolic, i.e, stable and unstable manifolds intersect transversally, and the dynamics can be encoded in a binary Bernoulli shift [15].

The main computational difficulty consists in the existence of a weak singularity at $\boldsymbol{q} = \boldsymbol{0}$. To remove this singularity, we have to scale the equations of motion by introducing the time transformation

$$dt = \rho(\boldsymbol{q})\, d\tau\,, \qquad \rho(\boldsymbol{q}) := (q_x)^2 + (q_y)^2\,,$$

which implies that, in the new time τ, the norm of the vector field remains bounded at $\boldsymbol{q} = \boldsymbol{0}$. The time transformation has to be introduced such that the transformed equations of motion are still Hamiltonian. A constant step-size symplectic method can then be used to integrate the transformed system. Let us describe the general approach: Assume that a Hamiltonian function $H(\boldsymbol{q},\boldsymbol{p}) = \boldsymbol{p}^T \boldsymbol{M}^{-1}\boldsymbol{p}/2 + V(\boldsymbol{q})$ and a scaling function $\rho(\boldsymbol{q})$ are given. Following Zare & Szebehely [41], we introduce the modified Hamiltonian function

$$E\left(\boldsymbol{q},\boldsymbol{p},t,e\right) := \rho(\boldsymbol{q})\left[H(\boldsymbol{q},\boldsymbol{p}) - e\right]$$

with corresponding Hamiltonian equations of motion

$$\frac{d}{d\tau}\,\boldsymbol{q} = \rho(\boldsymbol{q})\,\boldsymbol{M}^{-1}\boldsymbol{p}\,, \tag{5.9}$$

$$\frac{d}{d\tau}\,\boldsymbol{p} = -\rho(\boldsymbol{q})\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,V(\boldsymbol{q}) - \left[H(\boldsymbol{q},\boldsymbol{p}) - e\right]\boldsymbol{\nabla}_{\boldsymbol{q}}\,\rho(\boldsymbol{q})\,, \tag{5.10}$$

$$\frac{d}{d\tau}\,t = \rho(\boldsymbol{q})\,, \tag{5.11}$$

$$\frac{d}{d\tau}\,e = 0$$

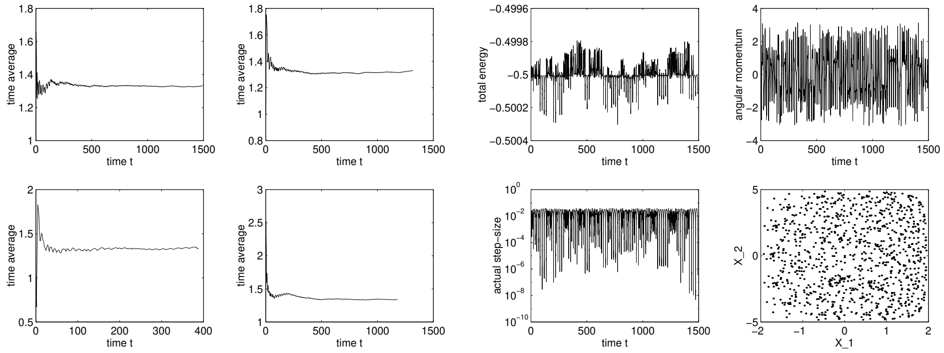
in extended phase space $\mathrm{R}^{2n} \times \mathrm{R}^2$. In particular, let us consider the case $e = H(\boldsymbol{q}(0),\boldsymbol{p}(0))$. Then (5.9)–(5.11) can be simplified to

$$\begin{aligned}
\frac{d}{d\tau}\,\boldsymbol{q} &= \rho(\boldsymbol{q})\,\boldsymbol{M}^{-1}\boldsymbol{p}\,,\\
\frac{d}{d\tau}\,\boldsymbol{p} &= -\rho(\boldsymbol{q})\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,V(\boldsymbol{q})\,,\\
\frac{d}{d\tau}\,t &= \rho\left(\boldsymbol{q}\right)
\end{aligned}$$

on the hypersurface of constant energy $E = 0$. This is a scaled vector field as desired but which is not Hamiltonian anymore. Therefore, as suggested by the author in [27] and independently by Hairer [17], the Hamiltonian equations (5.9)–(5.11) are discretized by a symplectic method and $e = H(\boldsymbol{q}_0,\boldsymbol{p}_0)$. For example, the equations can be discretized by the symplectic Euler method, i.e.

$$\begin{aligned}
\boldsymbol{q}_{n+1} &= \boldsymbol{q}_n + \delta\tau\,\rho(\boldsymbol{q}_n)\,\boldsymbol{M}^{-1}\boldsymbol{p}_{n+1}\,,\\
\boldsymbol{p}_{n+1} &= \boldsymbol{p}_n - \delta\tau\,\rho(\boldsymbol{q}_n)\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,V(\boldsymbol{q}_n) - \delta\tau\left(H(\boldsymbol{q}_n,\boldsymbol{p}_{n+1}) - e\right)\boldsymbol{\nabla}_{\boldsymbol{q}}\,\rho(\boldsymbol{q}_n)\,,\\
t_{n+1} &= t_n + \delta\tau\,\rho\left(\boldsymbol{q}_n\right).
\end{aligned}$$

The method is explicit in the variable $\boldsymbol{q}$. Unfortunately this implies that the method is only first order in $\delta\tau$. However, the method is symplectic and, therefore, the



(a) (b)

FIG. 5.1. *(a) The time evolution of the average $\langle r\rangle_n$ (mean distance) is shown for four different initial conditions with equal initial energy $e = -1/2$. (b) Time evolution of the error in energy, angular momentum, and the actual step-size. The bottom-right figure shows the intersections of the trajectory with the Poincaré section in the (X_1, X_2) coordinates. One thousand intersections are plotted.*

Hamiltonian $E = [H(\boldsymbol{q},\boldsymbol{p}) - e]\rho(\boldsymbol{q})$ is conserved to $\mathcal{O}(\delta\tau)$ over exponentially long periods of time. A second-order symplectic discretization can be obtained by using the second-order Lobatto IIIa-b partitioned Runge-Kutta formula [34], i.e.

$$\boldsymbol{p}_{n+1/2} = \boldsymbol{p}_n - \frac{\delta\tau}{2}\Big[\rho(\boldsymbol{q}_n)\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}_n) - [H(\boldsymbol{q}_n,\boldsymbol{p}_{n+1/2}) - e]\boldsymbol{\nabla}_{\boldsymbol{q}}\rho(\boldsymbol{q}_n)\Big]\,, \tag{5.12}$$

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n + \frac{\delta\tau}{2}[\rho(\boldsymbol{q}_{n+1}) + \rho(\boldsymbol{q}_n)]\,\boldsymbol{M}^{-1}\boldsymbol{p}_{n+1/2}\,, \tag{5.13}$$

$$\boldsymbol{p}_{n+1} = \boldsymbol{p}_{n+1/2} - \frac{\delta\tau}{2}\Big[\rho(\boldsymbol{q}_{n+1})\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}_{n+1}) - [H(\boldsymbol{q}_{n+1},\boldsymbol{p}_{n+1/2}) - e]\boldsymbol{\nabla}_{\boldsymbol{q}}\rho(\boldsymbol{q}_{n+1})\Big]\,, \tag{5.14}$$

$$t_{n+1} = t_n + \frac{\delta\tau}{2}[\rho(\boldsymbol{q}_n) + \rho(\boldsymbol{q}_{n+1})]\,. \tag{5.15}$$

The resulting scheme is implicit in $\rho(\boldsymbol{q})$.

This approach is applied to the anisotropic Kepler problem with a scaling function $\rho(\boldsymbol{q}) = (q_x)^2 + (q_y)^2$. We chose initial values such that $e = H = -1/2$ and $L \neq 0$. The equations of motion are integrated using the second order symplectic method (5.12)-(5.15) with $\delta\tau = 0.05$. The time average of an observable $A(\boldsymbol{q})$ along a trajectory $\{\boldsymbol{q}_n\}_{n=1,\ldots,M}$ is computed according to the recursive formula

$$\langle A\rangle_n = \frac{1}{t_n}\left[t_{n-1}\langle A\rangle_{n-1} + \delta t_n A(\boldsymbol{q}_n)\right]\,.$$

The time average of $r = \sqrt{(q_x)^2 + (q_y)^2}$ (mean distance) was computed for four different initial conditions and the evolution of the corresponding time averages $\langle r\rangle_n$ can be found in Fig. 5.1(a). The different lengths of the time intervals are due to the fact that the same number of steps with step-size $\delta\tau$ were taken which leads to different actual step-sizes δt_n. Within a tolerance of $c = 0.04$, these averages converge to the same value ≈ 1.33. In Fig. 5.1(b), the total energy H, the angular momentum

L, and the variation in the actual step-size $\delta t = r^2\,\delta\tau$ can be found for a particular trajectory. We also plotted the intersections of the trajectory with the (q_x, p_x) plane in the (X_1, X_2) representation (5.7)–(5.8). Theoretically, these points should fill the rectangle in a uniform way (the invariant meassure is the Lebesgue meassure). With one thousand points plotted, the uniform distribution is satisfied quite well. $\diamond$

Acknowledgements. I like to thank Ernst Hairer, Chus Sanz-Serna, Andrew Stuart, and Claudia Wulff for comments on an earlier version of this paper and the referees for making valuable suggestions.

REFERENCES

[1] M. Adams, T. Ratiu, and R. Schmid, *The Lie group structure of diffeomorphism groups and invertible Fourier integrals operators with Applications*, in Infinite Dimensional Groups with Applications, V. Kac, ed., Springer-Verlag, New York, 1985.

[2] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd edition, Springer-Verlag, New York, 1987.

[3] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd edition, Springer-Verlag, New York, 1988.

[4] S. P. Auerbach, and A. Friedman, *Long-time behaviour of numerically computed orbits: small and intermediate time-step analysis of one-dimensional systems*, J. Comput. Phys., 93 (1991), pp. 189–223.

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

[6] W.-J. Beyn, *Numerical methods for dynamical systems*, in Advances in Numerical Analysis, Vol. I, Clarendon Press, Oxford, 1991.

[7] M. P. Calvo, A. Murua, and J. M. Sanz-Serna, *Modified Equations for ODEs*, Contemporary Mathematics, 172 (1994), pp. 63–74.

[8] M. P. Calvo, A. Iserles, and A. Zanna, *Runge-Kutta methods on manifolds*, in Numerical Analysis: A.R. Mitchell's 75th Birthday Volume, G. A. Watson and D. F. Griffiths, eds., World Science, Singapore, 1997, pp. 57–70.

[9] T. Eirola, *Aspects of backward error analysis of numerical ODEs*, J. Comput. Appl. Math., 45 (1993), pp. 65–73.

[10] K. Feng, *Formal power series and numerical algorithms for dynamical systems*, in Proceedings of International Conference on Scientific Computation, Tony Chan and Zhong-Ci Shi, eds., Series on Appl. Math, 1 (1991), pp. 28–35.

[11] B. Fiedler and J. Scheurle, *Discretization of homoclinic orbits and invisible chaos*, Memoirs Amer. Math. Soc., 570 (1996).

[12] D. F. Griffiths and J. M. Sanz-Serna, *On the scope of the modified equations*, J. Sci. Statist. Comput., 7 (1986), pp. 994–1008.

[13] O. Gonzalez, D. J. Higham, and A. Stuart, *Qualitative properties of modified equations*, IMA J. Numer. Anal., to appear.

[14] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[15] M. C. Gutzwiller, *Chaos in Classical and Quantum Mechanics*, Springer-Verlag, New-York, 1990.

[16] E. Hairer, *Backward analysis of numerical integrators and symplectic methods*, Annals of Numerical Mathematics, 1 (1994), pp. 107–132.

[17] E. Hairer, *Variable time step integration with symplectic methods*, Appl. Numer. Math., 25 (1997), pp. 219–227.

[18] E. Hairer and Ch. Lubich, *The life-span of backward error analysis for numerical integrators*, Numer. Math., 76 (1997), pp. 441–462.

[19] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations, Vol. I*, 2nd edition, Springer-Verlag, 1993.

[20] E. Hairer and D. Stoffer, *Reversible long-term integration with variable step sizes*, SIAM J. Sci. Comput., 18 (1997), pp. 257–269.

[21] R. Hamilton, *The inverse function theorem of Nash and Moser*, Bull. Am. Math. Soc., 7 (1982), pp. 65-222.

[22] K. Jänich, *Funktionentheorie*, 3rd edition, Springer-Verlag, 1993.

[23] F. M. Lasagni, *Canonical Runge-Kutta methods*, ZAMP, 39 (1988), pp. 952–953.

[24] J. Moser, *Lectures on Hamiltonian systems*, Mem. Am. Math. Soc. 81 (1968), pp. 1–60.

[25] A. I. Neishtadt, *The separation of motions in systems with rapidly rotating phase*, J. Appl. Math. Mech., 48 (1984), pp. 133–139.

[26] S. Reich, *Numerical integration of generalized Euler equations*, preprint, 1993.

[27] S. Reich, *Backward error analysis for numerical integrators*, Preprint SC 96-21, Konrad-Zuse-Zentrum Berlin, 1996.

[28] S. Reich, *On higher order semi-explicit symplectic partitioned Runge-Kutta methods for constrained Hamiltonian systems*, Numer. Math. 76 (1997), pp. 231–247.

[29] S. Reich, *Preservation of adiabatic invariants under symplectic discretization*, Appl. Numer. Math., to appear.

[30] S. Reich, *Dynamical Systems, Exponentially Small Estiamtes, and Numerical Integration*, Habilitationsschrift, FU Berlin, 1998.

[31] J. M. Sanz-Serna, *Symplectic integrators for Hamiltonian problems: an overview*, Acta Numerica, 1 (1992), pp. 243–286.

[32] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman & Hall, London, 1994.

[33] T. Sauer and J. A. York, *Rigorous verification of trajectories for the computer simulation of dynamical systems*, Nonlinearity, 4 (1994), pp. 961–979.

[34] G. Sun, *Symplectic partitioned Runge-Kutta methods*, J. Comput. Math., 11 (1993), pp. 365–372.

[35] V. S. Varadarajan, *Lie Groups, Lie Algebras, and Their Representation*, Prentice-Hall, Englewood Cliffs, 1974.

[36] M. Viana, *Stochastic Dynamics of Deterministic Systems*, Instituto de Matemática Pura e Applicada (IMPA), Rio de Janeiro, 1997.

[37] P. Walters, *Introduction to Ergodicity Theory*, 2nd edition, Springer-Verlag, New York, 1985.

[38] R. F. Warming and B. J. Hyett, *The modified equation approach to the stability and accuracy of finite-difference methods*, J. Comp. Phys., 14 (1974), pp. 159–179.

[39] L.-S. Young, *Large deviations in dynamical systems*, Trans. Am. Math. Soc., 318 (1990), pp. 525–543.

[40] H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A, 150 (1990), pp. 262–268.

[41] K. Zare and V. Szebehely, *Time transformations for the extended phase space*, Celestial Mechanics, 11 (1975), pp. 469–482.