Konrad-Zuse-Zentrum
für Informationstechnik Berlin

Takustraße 7
D-14195 Berlin-Dahlem
Germany

SEBASTIAN REICH

Dynamical Systems, Numerical Integration, and Exponentially Small Estimates

Preprint SC 98-19 (July 1998)

Preface

Overview

Various problems in science and engineering lead to a differential equation containing a small parameter ϵ, i.e.

$$\frac{d}{dt}\boldsymbol{x} \quad = \quad \boldsymbol{A}(\boldsymbol{x}) + \epsilon\,\boldsymbol{B}(\boldsymbol{x};\epsilon)\,,$$

$\boldsymbol{x} \in \mathrm{R}^n$. Typically the solution to the problem is known when the parameter is equal to zero and one would like to find an approximative solution for non-zero parameter values. To be more precise, one is looking for a coordinate transformation from $\boldsymbol{x} \in \mathrm{R}^n$ to new coordinates $\bar{\boldsymbol{x}} \in \mathrm{R}^n$ such that for the corresponding transformed system

$$\frac{d}{dt}\bar{\boldsymbol{x}} \quad = \quad \boldsymbol{A}(\bar{\boldsymbol{x}}) + \epsilon\,\bar{\boldsymbol{B}}(\bar{\boldsymbol{x}};\epsilon)$$

the two vector fields $\boldsymbol{A}$ and $\bar{\boldsymbol{B}}$ commute. This implies that the solutions of the modified problem are given by integrating

$$\frac{d}{dt}\bar{\boldsymbol{x}} \quad = \quad \boldsymbol{A}(\bar{\boldsymbol{x}})$$

and

$$\frac{d}{dt}\bar{\boldsymbol{x}} \quad = \quad \epsilon\,\bar{\boldsymbol{B}}(\bar{\boldsymbol{x}};\epsilon) \tag{0.1}$$

separately. The transformed system in *normal form* is usually obtained in form of an *asymptotic expansion* in the small parameter [60]. It is known that these expansions diverge in general. However, starting with a paper by NEKHOROSHEV [89], one became aware that the truncation error can be made exponentially small provided that the asymptotic expansion is truncated at an optimal index related to the size of the small parameter. Since then, the idea of optimal truncation of asymptotic expansions has been applied in various areas (see, for example, [88],[13],[14],[15],[93],[91],[16]. The importance of normal form transformations for numerical analysis rests upon the fact that the system (0.1) is much easier to integrate numerically than the original problem formulation. Asymptotic expansions play an important role in other fields as well. For example, the approximation of maps by flows of vector fields. For these problems estimates on the truncated asymptotic expansion are of equal importance.

One of the main objective of this work is to provide a new proof for the exponential smallness of the truncation error for the following two classes of problems:

(A) Numerical integrators (one step methods)

$$\begin{aligned} \boldsymbol{x}_{n+1} &= \boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}_n)\,, \\ t_{n+1} &= t_n + \Delta t\,, \end{aligned}$$

with the step-size Δt as the small parameter,

(B) Differential equations with two different time scales and the reciprocal of the separation in the time scales as the small parameter.

For class (A), we look for a modified vector field $\tilde{\boldsymbol{X}}$ such that its time-Δt-flow $\boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}$ is equivalent to the numerical Δt-approximation $\boldsymbol{\Psi}_{\Delta t}$ to the exact time-Δt-flow of the given vector field $\boldsymbol{Z}$. This approach is called *backward error analysis*. It has been shown by NEISHTADT [88] that the difference between the flow of the modified differential equation and the numerical approximation can be made exponentially small, i.e.

$$\|\, \boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}) - \boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}(\boldsymbol{x}) \,\| \quad \leq \quad c_1\, \Delta t\, e^{-c_2/\Delta t}\,,$$

$c_1, c_2 > 0$ appropriate constants. This result is of utmost importance for understanding the behavior of symplectic methods for Hamiltonian differential equations and their energy conserving property. Further papers on exponential estimates in the context of backward error analysis include BENETTIN & GIORGILLI [16] and HAIRER & LUBICH [55]. Here we give a new proof that seems simpler than the existing ones and also allows one to take the order of a method into account. This is achieved by defining a new recursion for the asymptotic expansion of the modified vector field $\tilde{\boldsymbol{X}}(\Delta t)$ and by providing a new approach to estimate the difference between the numerical one step method $\boldsymbol{\Psi}_{\Delta t}$ and the flow map $\boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}$. The assumptions are that (i) the given vector field $\boldsymbol{Z}$ is real analytic and that (ii) the vector field is bounded by a constant on an appropriate subset of $\boldsymbol{C}^n$. Note that our approach is not restricted to a particular representation of the one step method $\boldsymbol{\Psi}_{\Delta t}$ as required in [55] and works directly with an estimate for the vector field $\boldsymbol{Z}$ instead of assuming certain estimates for the Taylor series representation of the one step method $\boldsymbol{\Psi}_{\Delta t}$ as in [16].

Backward error analysis is used to explain the preservation of the adiabatic invariants associated with Hamiltonian systems with two different time scales under symplectic discretization and to discuss the symplectic integration of ergodic Hamiltonian systems. In particular, we show that adiabatic invariants are preserved over exponentially long periods of time and that, for ergodic systems, the time averages computed along numerically generated trajectories converge to the exact time averages[1]. While the first result involves backward error analysis and normal form theory, the second result requires backward and forward error analysis as well as the concept of *shadowing* for hyperbolic mappings. Both results seem of great importance for a better understanding of symplectic integration methods for Hamiltonian systems with a complex solution behavior.

[1] To be more precise, we show this results for systems with a hyperbolic structure.

We also show how variable step-size symplectic integration can be achieved by scaling the given Hamiltonian function and by an appropriate discretization of the resulting equations of motion. This opens up a possibility to overcome the constant step-size restriction of standard symplectic methods.

In case of class (B), we look for a coordinate transformation that decouples the slow and fast motion. For many problems, this transformation to *normal form* can be achieved up to terms exponentially small. Corresponding estimates have been derived for various special types of differential equations. For example, NEKHOROSHEV [89], PÖSCHEL [93], PERRY & WIGGINS [91] looked at perturbed integrable Hamiltonian systems, NEISHTADT [88] considered systems with a single fast degree of motion, and BENETTTIN, GALGANI & GIORGILLI [14],[15] discuss Hamiltonian systems with linear oscillators as fast degrees of motion coupled to a slow motion in some other variables. Here we consider the *normal form* problem from a general point of view and derive an exponentially small estimate for a rather broad class of problems. Using a different normal form recursion, a similar result has been stated by FASSÒ in [36]. Our approach is based on the observation that the coordinate transformation to normal form can be defined as the time-one-flow map of an appropriate autonomous, i.e. time independent, vector field $\boldsymbol{W}(\epsilon)$ which leads to a particularly simple normal form recursion.

We apply our estimate to various special problems such as systems of coupled linear oscillators with a slowly varying parameter, nonlinear systems with a single fast degree of motion, and mechanical systems with highly oscillatory internal degree(s) of motion. In particular, the normal form recursion and the exponential estimate for general mechanical systems with a single fast degree of motion are novel and provide much sharper results than those obtained previously by RUBIN & UNGAR [105] and, more recently, by BORNEMANN & SCHÜTTE [24] via *homogenization* techniques.

We show how normal form theory can be used to derive new algorithms and provide a deeper understanding of existing ones. In particular, we like to mention (i) the preservation of adiabatic invariants under symplectic discretization, (ii) elimination of highly oscillatory internal degrees of freedom and explicit symplectic integration of systems of rigid bodies, (iii) the idea of soft constraints to model flexible water molecules without having to include the high-frequency internal degrees of freedom, and (iv) an improved multiple-time-stepping integrator that overcomes the resonance induced instabilities of standard multiple-time-stepping applied to highly oscillatory mechanical systems[2].

[2]This improvement was inspired by reading the paper by GARCIA-ARCHILLA, SANZ-SERNA & SKEEL [42]

Acknowledgements

It is a pleasure to thank all the individuals who made this work possible and provided valuable input.

Among those I like to thank in particular

Peter Deuflhard for his interest in this work and his generous support over the last three years,

Ben Leimkuhler as a loyal friend and source of inspiration,

Uri Ascher, Ernst Hairer, Christian Lubich, Chus Sanz-Serna, Jörg Schmeling, Bob Skeel, Andrew Stuart, and Claudia Wulff for various hints during the preparation of this work,

my colleagues Folkmar Bornemann and Christof Schütte for discussions on the elimination of high frequency degrees of motion,

the lunch-gang at ZUSE for sharing more than lunch,

two young ladies from Prenzlauer Berg for suggesting a major improvement and for being a great company,

and, last but certainly not least, my parents who made it all possible in the first place.

Notations

Symbol	Meaning
R	real numbers
C	complex numbers
Z	integers
I	non-negative integers
T^n	n-torus $\mathrm{T}^n = \mathrm{R}^n/2\pi\mathrm{Z}^n$
$\boldsymbol{x}$	phase space variable, $\boldsymbol{x} \in \mathrm{R}^n$ or $\boldsymbol{x} \in \mathrm{C}^n$
$\boldsymbol{q}$	vector of position coordinates
$\boldsymbol{p}$	vector of conjugate momenta
$\mathfrak{i}$	imaginary unit
$\boldsymbol{\Phi}_{t,\boldsymbol{X}}$	time-t-flow map of a vector field $\boldsymbol{X}$
$\boldsymbol{\Phi}_{t,H}$	time-t-flow map corresponding to a Hamiltonian H
Δt	time step-size
$\boldsymbol{\Psi}_{\Delta t}$	numerical time-Δt-step map (one step method)
$\boldsymbol{\nabla}_{\boldsymbol{x}}$	gradient with respect to vector $\boldsymbol{x}$
$\boldsymbol{\partial}_{\boldsymbol{x}}$	partial derivative with respect to vector $\boldsymbol{x}$
$\|\|\boldsymbol{x}\|\|$	max-norm of vector $\boldsymbol{x} \in \mathrm{R}^n$, $\boldsymbol{x} \in \mathrm{C}^n$ respectively
$\mathcal{B}_r(\boldsymbol{x})$	complex ball of radius $r > 0$ around a point $\boldsymbol{x}$
$\mathcal{K},\mathcal{U},\mathcal{V}$	subsets of phase space R^n
$\mathcal{B}_r\mathcal{K}$	union of all complex balls of radius r around each $\boldsymbol{x} \in \mathcal{K}$
$\|\|\,\boldsymbol{G}\,\|\|_r$	sup-norm of function $\boldsymbol{G}$ on complex domain $\mathcal{B}_r(\boldsymbol{x})$
$\|\,\boldsymbol{G}\,\|_r$	norm of function $\boldsymbol{G}$ on complex domain $\mathcal{B}_r\mathcal{K}$
$[\boldsymbol{X},\boldsymbol{Y}]$	Lie bracket of two vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$
$\{G,F\}$	Poisson bracket of two functions F and G
$\boldsymbol{\Phi}^*_{1,\boldsymbol{W}}\boldsymbol{Y}$	pull-back of vector field $\boldsymbol{Y}$ under $\boldsymbol{\Phi}_{1,\boldsymbol{W}}$
$\boldsymbol{f} \circ \boldsymbol{g}$	composition of map $\boldsymbol{f}$ with map $\boldsymbol{g}$
$[\boldsymbol{g}]^n$	n-fold composition of a mapping $\boldsymbol{g}$
$\boldsymbol{X} \cdot \boldsymbol{Y}$	vector product of two vector fields, i.e. $\boldsymbol{X} \cdot \boldsymbol{Y}\,(\boldsymbol{x}) = \boldsymbol{X}(\boldsymbol{x}) \cdot \boldsymbol{Y}(\boldsymbol{x})$
$\delta(x)$	Dirac's delta distribution
$\langle f,g\rangle$	L^2 inner product of two functions f and g
$\langle\mathcal{A}\rangle_T$	time average of an observable $\mathcal{A}$
$SO(3)$	Lie group of orthogonal 3×3 matrices
$\mathfrak{so}(3)$	Lie algebra of skew-symmetric 3×3 matrices
$\mathcal{M}, T\mathcal{M}, T^*\mathcal{M}$	manifold, tangent manifold, cotangent manifold
$\mathfrak{G}$	submanifold of the Frechet manifold of smooth diffeomorphisms
$\mathfrak{g}$	linear subspace in the Lie algebra of smooth vector fields
$\mathbf{id}$	identity map, i.e. $\mathbf{id}(\boldsymbol{x}) = \boldsymbol{x}$
$\boldsymbol{A}^{-1}$	inverse of matrix $\boldsymbol{A}$
$\boldsymbol{A}^T$	transpose of matrix $\boldsymbol{A}$

Contents

1

Mappings Near the Identity

In this chapter, we discuss the approximation of mappings (diffeomorphisms) near the identity by flows of autonomous vector fields. The approximating vector fields will be given in terms of an asymptotic expansion which is defined by a simple recursion. Optimal truncation of this series yields an exponentially small remainder. The proof of this result is fundamental in the sense that similar techniques will be used in the chapters on backward error analysis and normal form theory. The question on how to interpolate a (symplectic) map close to the identity by an autonomous vector field was apparently first posed by Moser [85]. The first "direct" proof for the exponential smallness of the difference between the map and the best interpolating vector field was given by Benettin & Giorgilli [16].

1.1 The Asymptotic Expansion

Let $\boldsymbol{G} : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$ be a real analytic map on an open subset $\mathcal{U}$ of R^n. We assume that

$$||\,\boldsymbol{G}(\boldsymbol{x}) - \boldsymbol{x}\,|| \quad < \quad \epsilon \tag{1.1}$$

for all $\boldsymbol{x} \in \mathcal{K}$, $\mathcal{K} \subset \mathcal{U}$ a compact subset of $\mathcal{U}$; $||\,.\,||$ the l^∞-norm on R^n. Here $\epsilon > 0$ is a small number. In other words, $\boldsymbol{G}$ is a real analytic map ϵ close to the identity map on $\mathcal{K} \subset \mathcal{U}$. Our aim is to find a real analytic vector field $\hat{\boldsymbol{X}} : \mathcal{V} \subset \mathrm{R}^n \to \mathrm{R}^n$ on an appropriate open subset $\mathcal{V}$ of R^n such that the corresponding time-t-flow map $\boldsymbol{\Phi}_{t,\hat{\boldsymbol{X}}} : \mathcal{V} \to \mathrm{R}^n$ satisfies

$$\boldsymbol{\Phi}_{t=1,\hat{\boldsymbol{X}}}(\boldsymbol{x}) \quad \approx \quad \boldsymbol{G}(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathcal{K}$. For that reason, let us consider the recursion

$$\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} \quad := \quad \boldsymbol{G} - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i}\,, \tag{1.2}$$

$$\hat{\boldsymbol{X}}_{i+1} \quad := \quad \hat{\boldsymbol{X}}_i + \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} \tag{1.3}$$

with $\hat{\boldsymbol{X}}_0 = 0$ and $i = 0, 1, \ldots, s$.

Remark 1.1. Note that (1.2)-(1.3) can be considered as a simplified Newton method applied to the "nonlinear equation"

$$0 = \boldsymbol{G} - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}} \tag{1.4}$$

in the "unknown" $\hat{\boldsymbol{X}}$. The exact Newton method would lead to the equation (see, for example, [32])

$$\boldsymbol{G}(\boldsymbol{x}_0) - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i}(\boldsymbol{x}_0) = \int_0^1 \boldsymbol{W}(1,s;\boldsymbol{x}_0)\, \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}(\boldsymbol{x}(s))\, ds\,, \quad (x_0 \in \mathcal{U})\,. \tag{1.5}$$

Here $\boldsymbol{x}(t)$ denotes the solution of the differential equation

$$\frac{d}{dt}\boldsymbol{x} = \hat{\boldsymbol{X}}_i(\boldsymbol{x})$$

with initial value $\boldsymbol{x}(0) = \boldsymbol{x}_0$ and $\boldsymbol{W}(t,s;\boldsymbol{x}_0)$ is the Wronskian matrix of the variational equation

$$\frac{d}{dt}\boldsymbol{u} = \left[\frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\hat{\boldsymbol{X}}_i(\boldsymbol{x}(t))\right]\boldsymbol{u}\,, \qquad \boldsymbol{u}(s) = \boldsymbol{I}_n\,.$$

Note that (1.5) is, in general, not solvable for $\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}$ [58]. However, it would be certainly of interest to identify cases for which (1.4) has a solution. In some cases, this question is related to Kolmogorov's method of proving KAM theory [29]. □

In the sequel, we will use the vector fields

$$\boldsymbol{Y}(\xi) := \xi\,\boldsymbol{Y}_0 \quad \text{and} \quad \boldsymbol{Y}_0 := (\boldsymbol{G} - \mathbf{id})/\epsilon\,,$$

$||\boldsymbol{Y}_0(\boldsymbol{x})|| < 1$ on $\mathcal{K}$. This allows us to introduce the one parametric family of mappings

$$\boldsymbol{G}_\xi := \mathbf{id} + \boldsymbol{Y}(\xi) = \mathbf{id} + \xi\,\boldsymbol{Y}_0\,, \qquad \xi \geq 0\,,$$

with

$$\boldsymbol{G} = \boldsymbol{G}_{\xi=\epsilon}\,.$$

Thus, we can formally consider the vector fields $\hat{\boldsymbol{X}}_i$ and $\boldsymbol{\Delta}\hat{\boldsymbol{X}}_i$, $1,\ldots,s$, as functions of ξ. Obviously, we have

$$\boldsymbol{\Delta}\hat{\boldsymbol{X}}_1(\xi) = \xi\,\boldsymbol{Y}_0 \tag{1.6}$$

and, using Taylor series representation of the flow map $\boldsymbol{\Phi}_{t,\hat{\boldsymbol{X}}_1}$ with respect to time t, i.e.,

$$\boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_1} = \mathbf{id} + \xi\,\boldsymbol{Y}_0 + \frac{\xi^2}{2}\boldsymbol{L}_{\boldsymbol{Y}_0}\boldsymbol{Y}_0 + \mathcal{O}(\xi^3) = \mathbf{id} + \boldsymbol{Y} + \frac{1}{2}\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{Y} + \mathcal{O}(\xi^3)\,,$$

we obtain

$$\boldsymbol{\Delta}\hat{\boldsymbol{X}}_2(\xi) = -\frac{\xi^2}{2}\boldsymbol{L}_{\boldsymbol{Y}_0}\boldsymbol{Y}_0 + \mathcal{O}(\xi^3) = -\frac{1}{2}\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{Y} + \mathcal{O}(\xi^3)\,.$$

Here $\boldsymbol{L}_{\boldsymbol{Y}_0}\boldsymbol{Y}_0$ denotes the Lie derivative of $\boldsymbol{Y}_0$ with respect to $\boldsymbol{Y}_0$ [119], i.e.,

$$\boldsymbol{L}_{\boldsymbol{Y}_0}\boldsymbol{Y}_0(\boldsymbol{x}) = \frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{Y}_0(\boldsymbol{x})\cdot\boldsymbol{Y}_0(\boldsymbol{x}).$$

Continuing this process, we obtain

Lemma 1.1. The vector fields $\hat{\boldsymbol{X}}_i(\xi)$, $i=1,2,\ldots,s$, satisfy

$$\boldsymbol{G}_\xi - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i} = \mathcal{O}(\xi^{i+1}).$$

□

Proof. We have to show that, if

$$\boldsymbol{G}_\xi - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i} = \mathcal{O}(\xi^{i+1}),$$

then

$$\boldsymbol{G}_\xi - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_{i+1}} = \mathcal{O}(\xi^{i+2}).$$

Now, with $\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} = \mathcal{O}(\xi^{i+1})$,

$$\begin{aligned}
\boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_{i+1}} &= \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i+\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}} \\
&= \boldsymbol{\Phi}_{1,\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}} \circ \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i} + \mathcal{O}(\xi^{i+2}) \\
&= (\mathbf{id} + \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}) \circ \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i} + \mathcal{O}(\xi^{i+2}) \\
&= \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i} + \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} + \mathcal{O}(\xi^{i+2})
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{G}_\xi - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_{i+1}} &= \boldsymbol{G}_\xi - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i} - \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} + \mathcal{O}(\xi^{i+2}) \\
&= (\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} - \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}) + \mathcal{O}(\xi^{i+2}) \\
&= \mathcal{O}(\xi^{i+2}).
\end{aligned}$$

□

From Lemma 1.1 it follows that

$$\boldsymbol{\Delta}\hat{\boldsymbol{X}}_i = -\frac{\xi^i}{i!}\left[\frac{\partial^i}{\partial\xi^i}\boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_{i-1}}\right]_{\xi=0} + \mathcal{O}(\xi^{i+1}).$$

From now on we will drop the higher order ξ terms in $\boldsymbol{\Delta}\hat{\boldsymbol{X}}_i$ and use the modified recursion

$$\boldsymbol{\Delta X}_i := -\frac{\xi^i}{i!}\left[\frac{\partial^i}{\partial\xi^i}\boldsymbol{\Phi}_{1,\boldsymbol{X}_{i-1}}\right]_{\xi=0}, \tag{1.7}$$

$$\boldsymbol{X}_{i+1} := \boldsymbol{X}_i + \boldsymbol{\Delta X}_{i+1} \tag{1.8}$$

with $\boldsymbol{X}_1 = \xi\,\boldsymbol{Y}_0$ and $i=1,2,\ldots,s$.

instead of (1.2)-(1.3). The result of Lemma 1.1 also applies to the modified recursion. All we need is that

$$\boldsymbol{G}_\xi - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i} \quad = \quad \boldsymbol{\Delta X}_{i+1} + \mathcal{O}(\xi^{i+1})$$

which follows from standard Taylor series expansion.

The sequence $\{\boldsymbol{\Delta X}_i(\xi = \epsilon)\}_{i\geq 1}$ does not, in general, converge to zero. Thus we are looking for the integer i_* such that

$$||\, \boldsymbol{G} - \boldsymbol{\Phi}_{1,\boldsymbol{X}_{i_*}} ||_\infty \quad = \quad \text{Min!}$$

where $||\,.\,||_\infty$ denotes the supremum norm on $\mathcal{K}$, i.e.,

$$||\, \boldsymbol{G} - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i} ||_\infty \quad := \quad \sup_{\boldsymbol{x}\in\mathcal{K}} ||\, \boldsymbol{G}\,(\boldsymbol{x}) - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i}(\boldsymbol{x})\, ||$$

and

$$\boldsymbol{X}_i(\epsilon) \quad := \quad \sum_{j=1}^{i} \boldsymbol{\Delta X}_j(\xi = \epsilon)\,.$$

1.2 The Theorem on the Exponential Estimate

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

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

Under the assumption that the real analytic vector field

$$\boldsymbol{Y}_0 \quad := \quad (\boldsymbol{G} - \mathbf{id})/\epsilon$$

is bounded by one on a complex ball of radius $R > 0$ around each $\boldsymbol{x}_0 \in \mathcal{K} \subset \mathrm{R}^n$, i.e.,

$$||\, \boldsymbol{Y}_0\, ||_R \quad = \quad \sup_{\boldsymbol{x}\in B_R(\boldsymbol{x}_0)} ||\, \boldsymbol{Y}_0(\boldsymbol{x})\, || \;\leq\; 1\,, \qquad (\boldsymbol{x}_0 \in \mathcal{K})\,,$$

one can prove the following result:

Theorem 1.1. Let $\boldsymbol{G} : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$ be a real analytic map ϵ close to the identity on a compact set $\mathcal{K} \subset \mathcal{U}$, i.e.,

$$||\, \boldsymbol{G}(\boldsymbol{x}) - \boldsymbol{x}\, || \quad < \quad \epsilon\,, \qquad \text{for all } \boldsymbol{x} \in \mathcal{K}\,.$$

Then there exists a real analytic vector field $\boldsymbol{X} : \mathcal{V} \subset \mathrm{R}^n \to \mathrm{R}^n$ such that

$$||\, \boldsymbol{G}(\boldsymbol{x}) - \boldsymbol{\Phi}_{1,\boldsymbol{X}}\,(\boldsymbol{x})\, || \quad \leq \quad 6\,\epsilon\, b\, e^{-\gamma/\epsilon}\,, \qquad \text{for all } \boldsymbol{x} \in \mathcal{K}\,. \tag{2.9}$$

Here $\gamma = R/(2ce)$, $b = 10$, $c = 150$, and $R > 0$ such that, for all $\boldsymbol{x}_0 \in \mathcal{K}$,

$$||\, \boldsymbol{G}\,(\boldsymbol{x}) - \boldsymbol{x}\, || \quad \leq \quad \epsilon$$

on the complex ball of radius R around $\boldsymbol{x}_0 \in \mathcal{K}$. □

1.3 Proof of the Theorem

Let us recall the following result for analytic functions: Let $f : \mathrm{R} \to \mathrm{R}$ be a real analytic function on a complex ball of radius $r > 0$ around $0 \in \mathrm{R}$. Cauchy's inequality [64] yields then, under the assumption that

$$| f(y) | \quad \le \quad m$$

for all $|y| \le y_0$, $0 < y_0 \le r$, the estimate

$$| \partial_y^j f(y = 0) | \quad \le \quad j!\, m\, y_0^{-j} \, .$$

for the jth derivative of f at $y = 0$. We will also use the following notation: Let $\boldsymbol{X}$ be a real analytic vector field on a complex ball of radius $R > 0$ around a point $\boldsymbol{x}_0 \in \mathrm{R}^n$. Then one denotes

$$|| \boldsymbol{X} ||_r \quad = \quad \sup_{\boldsymbol{x} \in \mathcal{B}_r(\boldsymbol{x}_0)} || \boldsymbol{X}(\boldsymbol{x}) ||$$

where $\mathcal{B}_r(\boldsymbol{x}_0) \subset \mathrm{C}^n$ is the complex ball of radius r, $R \ge r \ge 0$, around $\boldsymbol{x}_0 \in \mathrm{R}^n$.

The real analytic vector field

$$\boldsymbol{Y}(\xi) \quad := \quad \xi \, \boldsymbol{Y}_0 \quad \text{with} \quad \boldsymbol{Y}_0 \; := \; (\boldsymbol{G} - \mathbf{id})/\epsilon$$

satisfies

$$|| \boldsymbol{Y}(\xi) ||_R \quad \le \quad \xi \tag{3.10}$$

with $R > 0$ appropriately chosen and $\boldsymbol{x}_0 \in \mathcal{K}$. We also define $\boldsymbol{G}_\xi = \mathbf{id} + \boldsymbol{Y}(\xi)$ which allows us to formally consider the vector fields $\boldsymbol{X}_i$ and $\boldsymbol{\Delta X}_i$, $i = 1, \ldots , s$, as functions of ξ.

Lemma 1.2. The Lie derivative $\boldsymbol{L_Y Y}$ satisfies the estimate

$$\frac{1}{2} || \boldsymbol{L_Y Y} ||_{\alpha R} \quad \le \quad \xi \left(\frac{\xi}{(1-\alpha)R} \right)$$

for $\alpha \in [0, 1)$. □

Proof. Since

$$|| \boldsymbol{\Phi}_{t,\boldsymbol{Y}}(\boldsymbol{x}) - \boldsymbol{x} || \quad \le \quad \int_0^{|t|} || \boldsymbol{Y}(\boldsymbol{\Phi}_{\tau,\boldsymbol{Y}}(\boldsymbol{x})) || \, |d\tau| \, ,$$

and (3.10), the flow map $\boldsymbol{\Phi}_{t,\boldsymbol{Y}}$ certainly satisfies

$$\boldsymbol{\Phi}_{t,\boldsymbol{Y}}(\boldsymbol{x}) \quad \in \quad \mathcal{B}_R(\boldsymbol{x}_0)$$

for all

$$|t| \quad \le \quad \frac{(1-\alpha)R}{\xi} \quad =: \; t_0$$

and all $\boldsymbol{x} \in \mathcal{B}_{\alpha R}(\boldsymbol{x}_0)$. For $\boldsymbol{x} \in \mathcal{B}_{\alpha R}(\boldsymbol{x}_0)$, define

$$\boldsymbol{f}\,(t, \boldsymbol{x}) \quad := \quad \boldsymbol{\Phi}_{t,\boldsymbol{Y}}(\boldsymbol{x}) - \boldsymbol{x}\,.$$

Since $\boldsymbol{f}$ is real analytic in t,

$$||\,\boldsymbol{f}\,(t, \boldsymbol{x})\,|| \;=\; ||\,\boldsymbol{\Phi}_{t,\boldsymbol{Y}}(\boldsymbol{x}) - \boldsymbol{x}\,|| \quad \leq \quad (1-\alpha)R$$

for $|t| \leq t_0$, as well as

$$\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{Y}\,(\boldsymbol{x}) \quad = \quad \frac{\partial^2}{\partial t^2}\,\boldsymbol{\Phi}_{t=0,\boldsymbol{Y}}\,(\boldsymbol{x}) \quad = \quad \frac{\partial^2}{\partial t^2}\boldsymbol{f}\,(t=0, \boldsymbol{x})\,,$$

it follows from Cauchy's estimate that

$$\begin{aligned} \frac{1}{2}||\,\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{Y}\,(\boldsymbol{x})\,|| \quad &\leq \quad (1-\alpha)\,R\,t_0^{-2}\,, \\ &\leq \quad \xi\,\left(\frac{\xi}{(1-\alpha)R}\right) \end{aligned}$$

for all $\boldsymbol{x} \in \mathcal{B}_{\alpha R}(\boldsymbol{x}_0)$. □

Remark 1.2. A slightly better estimate than the one given in Lemma 1.2 can be obtained by applying Cauchy's estimate to $\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{Y}(\boldsymbol{x}) = \partial_t \boldsymbol{f}(t=0, \boldsymbol{x})$, $\boldsymbol{f}(t, \boldsymbol{x}) := \boldsymbol{Y}(\boldsymbol{x} + t\boldsymbol{Y}(\boldsymbol{x}))$ [16]. However, the "*flow map technique*" introduced in the proof of Lemma 1.2 is essential for the proof of Lemma 1.3 below. □

Next we have to derive an estimate for $||\boldsymbol{\Delta X}_i(\xi)||$ (1.7), $i = 1, \ldots, s$. According to (1.6), (1.7), and Lemma 2, we have

$$||\,\boldsymbol{\Delta X}_1(\xi)\,||_{\alpha R} \;=\; ||\,\boldsymbol{Y}\,||_{\alpha R} \quad \leq \quad \xi$$

and

$$||\,\boldsymbol{\Delta X}_2\,||_{\alpha R} \;=\; \frac{1}{2}||\,\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{Y}\,||_{\alpha R} \quad \leq \quad \xi\,\left(\frac{\xi}{(1-\alpha)R}\right)$$

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

Lemma 1.3. The vector fields $\boldsymbol{\Delta X}_i(\xi)$ (1.7) satisfy

$$||\,\boldsymbol{\Delta X}_i\,||_{\alpha R} \quad \leq \quad b\,\xi\,\left(\frac{c\,(i-1)\,\xi}{(1-\alpha)R}\right)^{i-1} \tag{3.11}$$

for $i \geq 3$ and $\alpha \in [0, 1)$. The constants b and c can be chosen as

$$b \;=\; 10 \qquad \text{and} \qquad c = 150\,.$$

□

Proof. We know that $\boldsymbol{\Delta X}_i(\xi)$ and $\boldsymbol{X}_i(\xi)$, $i = 1, \ldots, s$, are analytic functions of ξ. Let us assume that (3.11) holds for $i = 3, \ldots, j$. Then

$$\begin{aligned} ||\,\boldsymbol{X}_j\,(\xi)\,||_{\alpha R} &\leq \sum_{i=1}^{j} ||\,\boldsymbol{\Delta X}_i\,(\xi)\,||_{\alpha R} \\ &\leq \xi\left[1 + \frac{\xi}{(1-\alpha)R} + \sum_{i=3}^{j} b\left(\frac{c\,(i-1)\,\xi}{(1-\alpha)R}\right)^{i-1}\right] \qquad (3.12) \end{aligned}$$

which implies

$$||\,\boldsymbol{X}_j\,(\xi)\,||_{\alpha R} \;\leq\; b\,\xi_0 \;=\; \delta(1-\alpha)R$$

for

$$\xi \;\leq\; \frac{(1-\alpha)R}{c\,j} \;=:\; \xi_0\,,$$

$b \geq 9$, $c \geq 1$, and

$$\delta \;:=\; \frac{b}{c\,j}\,.$$

Here we have used that

$$\sum_{i=3}^{j}\left(\frac{i-1}{j}\right)^{i-1} \;\leq\; 0.85$$

for $j \geq 3$ which implies that

$$\begin{aligned} \left[1 + \frac{\xi_0}{(1-\alpha)R} + \sum_{i=3}^{j} b\left(\frac{c\,(i-1)\,\xi_0}{(1-\alpha)R}\right)^{i-1}\right] &= 1 + \frac{1}{c\,j} + b\left[\sum_{i=3}^{j}\left(\frac{i-1}{j}\right)^{i-1}\right] \\ &\leq b \end{aligned}$$

for $j \geq 3$, $b \geq 9$, and $c \geq 1$. Next we chose $b \geq 9$ and $c \geq 1$ large enough (for example, $b = 10$ and $c = 150$) such that

$$||\,\boldsymbol{X}_j(\xi)\,||_{(\alpha+\delta(1-\alpha))R} \;\leq\; b\,\xi_0 \;=\; \delta(1-\alpha)R \qquad (3.13)$$

for $\xi \leq \xi_0$ as well. In other words, we chose b and c such that

$$1 + \frac{1}{(1-\delta)\,c\,j} + b\sum_{i=3}^{j}\left(\frac{i-1}{(1-\delta)\,j}\right)^{i-1} \;\leq\; b$$

where we have used (3.12) with α replaced by $\alpha + \delta(1-\alpha)$ and

$$1 - (\alpha + (1-\alpha)\,\delta) \;=\; (1-\alpha)\,(1-\delta)\,.$$

In particular, for $b = 10$ and $c = 150$, we obtain $\delta < 0.067/j$ and

$$\begin{aligned} \sum_{i=3}^{j} \left(\frac{i-1}{(1-\delta)\, j} \right)^{i-1} &\leq \sum_{i=3}^{j} \left(\frac{i-1}{j-0.067} \right)^{i-1}, \\ &\leq 0.891 \end{aligned}$$

for $j \geq 3$. Let us now consider the real analytic function

$$\boldsymbol{f}\,(\xi, \boldsymbol{x}) \quad := \quad \boldsymbol{\Phi}_{1,\boldsymbol{X}_j}(\boldsymbol{x}) - \boldsymbol{x}$$

for $\boldsymbol{x} \in \mathcal{B}_{\alpha R}(\boldsymbol{x}_0)$. Since

$$||\, \boldsymbol{\Phi}_{1,\boldsymbol{X}_j}(\boldsymbol{x}) - \boldsymbol{x}\,|| \quad \leq \quad \int_0^1 ||\, \boldsymbol{X}_j\,(\boldsymbol{\Phi}_{\tau,\boldsymbol{X}_j}(\boldsymbol{x}))\,||\,|d\tau|$$

and (3.13), we have

$$\boldsymbol{\Phi}_{1,\boldsymbol{X}_j}(\boldsymbol{x}) \quad \in \quad \mathcal{B}_{\delta(1-\alpha)R}\,(\boldsymbol{x})$$

and, therefore,

$$||\, \boldsymbol{f}\,(\xi, \boldsymbol{x})\,|| \quad \leq \quad \delta\,(1-\alpha)\,R \;=\; b\,\xi_0 \tag{3.14}$$

for $|\xi| \leq \xi_0$ and $\boldsymbol{x} \in \mathcal{B}_{\alpha R}(\boldsymbol{x}_0)$. The function $\boldsymbol{f}$ is real analytic in ξ. Thus, by Cauchy's estimate, we obtain

$$\begin{aligned} ||\, \boldsymbol{\Delta X}_{j+1}(\xi)\,(\boldsymbol{x})\,|| &= \frac{\xi^{j+1}}{(j+1)!}\,||\,\partial_\xi^{j+1} \boldsymbol{f}\,(\xi = 0, \boldsymbol{x})\,||\,, \\ &\leq b\,\xi_0 \left(\frac{\xi}{\xi_0} \right)^{j+1}, \qquad (3.15) \\ &\leq b\,\xi \left(\frac{c\,j\,\xi}{(1-\alpha)R} \right)^{j}, \qquad (3.16) \end{aligned}$$

$\boldsymbol{x} \in B_{\alpha R}(\boldsymbol{x}_0)$, and the desired estimate for $||\boldsymbol{\Delta X}_{j+1}||_{\alpha R}$ follows. □

Next we need an estimate for the difference between $\boldsymbol{G}(\boldsymbol{x}_0) = \boldsymbol{G}_{\xi=\epsilon}(\boldsymbol{x}_0)$ and the flow map $\boldsymbol{\Phi}_{1,\boldsymbol{X}_i}(\boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathcal{K}$, for $\xi = \epsilon$. This is the subject of the following

Lemma 1.4. Whenever the constant ϵ in (1.1) satisfies

$$\epsilon \quad \leq \quad \frac{R}{2\,c\,i}\,,$$

then

$$||\, \boldsymbol{G}(\boldsymbol{x}_0) - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i}(\boldsymbol{x}_0)\,|| \quad \leq \quad 2\,\epsilon\,b \left(\frac{2\,c\,i\,\epsilon}{R} \right)^{i}, \qquad \boldsymbol{x}_0 \in \mathcal{K}\,,$$

with $b = 10$ and $c = 150$. □

Proof. According to standard Taylor series expansion, we have

$$|| \boldsymbol{G}_\epsilon (\boldsymbol{x}_0) - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i} (\boldsymbol{x}_0) || \quad \leq \quad \frac{\epsilon^{i+1}}{(i+1)!} \sup_{0 \leq \hat{\xi} \leq \epsilon} || \partial_\xi^{i+1} \boldsymbol{f} (\xi = \hat{\xi}, \boldsymbol{x}_0) ||$$

with, as in the proof of Lemma 1.3,

$$\boldsymbol{f} (\xi, \boldsymbol{x}_0) \quad := \quad \boldsymbol{\Phi}_{1,\boldsymbol{X}_i}(\boldsymbol{x}_0) - \boldsymbol{x}_0 \, .$$

This requires an estimate for $||\partial_\xi^{i+1} \boldsymbol{f}(\xi = \hat{\xi}, \boldsymbol{x}_0)||$, $0 \leq \hat{\xi} \leq \epsilon$. Following the proof of Lemma 1.3, i.e., taking $\alpha = 0$ in (3.14), we know that

$$|| \boldsymbol{f}(\xi, \boldsymbol{x}_0) || \quad \leq \quad b \, \xi_0$$

for $|\xi| \leq \xi_0 := R/(ci)$. Assume that $\epsilon \leq \xi_0/2$. Then

$$|| \boldsymbol{f}(\xi, \boldsymbol{x}_0) || \quad \leq \quad b \, \xi_0$$

for $|\xi - \hat{\xi}| \leq \xi_0/2$ and $\hat{\xi} = \epsilon \leq \xi_0/2$. Thus, Cauchy's estimate implies

$$\begin{aligned} \frac{1}{(i+1)!} || \partial_\xi^{i+1} \boldsymbol{f} (\xi = \hat{\xi}, \boldsymbol{x}_0) || \quad &\leq \quad b \, \xi_0 \left(\frac{2}{\xi_0} \right)^{i+1} , \\ &\leq \quad 2 \, b \left(\frac{2 \, c \, i}{R} \right)^i \end{aligned}$$

and, for $\epsilon \leq R/(2ci)$, the desired estimate

$$|| \boldsymbol{G}(\boldsymbol{x}_0) - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i} (\boldsymbol{x}_0) || \quad \leq \quad 2 \, \epsilon \, b \left(\frac{2 \, c \, i \, \epsilon}{R} \right)^i$$

follows. □

Starting with $i = 1$, Lemma 1.4 yields now

$$|| \boldsymbol{G}(\boldsymbol{x}_0) - \boldsymbol{\Phi}_{1,\boldsymbol{X}_i}(\boldsymbol{x}_0) || \quad \leq \quad 2 \, \epsilon \, b \left(\frac{2 \, i \, c \, \epsilon}{R} \right)^i , \qquad (\boldsymbol{x}_0 \in \mathcal{K}) \, ,$$

provided $\epsilon \leq R/(2ic)$ for all $i = 1, \ldots, s$. Let $i_*(\epsilon)$ be the integer part of

$$i_o(\epsilon) \quad := \quad \frac{R}{2 \, c \, \epsilon \, e} \, .$$

Note that this choice of $i_*(\epsilon)$ certainly implies

$$\epsilon \quad \leq \quad \frac{R}{2 \, c \, i_*} \, .$$

Then, for all $\boldsymbol{x}_0 \in \mathcal{K}$,

$$\begin{aligned} ||\boldsymbol{G}(\boldsymbol{x}_0) - \boldsymbol{\Phi}_{1,\boldsymbol{X}_{i_*}}(\boldsymbol{x}_0) || \quad &\leq \quad 2 \, \epsilon \, b \, e^{-i_*} , \\ &\leq \quad 2 \, \epsilon \, b \, e^{-i_o + 1} \\ &\leq \quad 6 \, \epsilon \, b \, e^{-\gamma/\epsilon} \end{aligned}$$

where $\gamma = i_o\epsilon = R/(2ce)$. Thus we have proved Theorem 1.1.

Remark 1.3. It was not our intention to provide an optimal estimate for the constants b and c. Instead we tried to keep the proof as simple as possible. A better estimate can, for example, be obtained by replacing the estimate (3.11) in Lemma 1.3 by

$$||\boldsymbol{\Delta X}_i||_{\alpha R} \quad \leq \quad b_i\,\xi \left(\frac{c_i\,(i-1)\,\xi}{(1-\alpha)\,R}\right)^{i-1} .$$

Here c_i and b_i are appropriate constants. For example, one can chose $c_3 = 1.5$, $c_4 = 2.0$, $c_5 = 2.5$, $c_6 = 3.0$, $c_7 = 3.5$, and

$$c_i \quad = \quad 3.5 + 0.3\left[e^{-(i-8)/10}\right]^{1/4} , \qquad \text{for } i \geq 8$$

and the constants b_i such that

$$1 + \frac{1}{c_{j+1}\,(j - \frac{b_{j+1}}{c_{j+1}})} + \sum_{i=3}^{j} b_i \left(\frac{c_i\,(i-1)}{c_{j+1}\,(j-\frac{b_{j+1}}{c_{j+1}})}\right)^{i-1} \quad \leq \quad b_{j+1} .$$

The above choice implies $c_i < 16$ and $b_i < 4.5$ for all $i \geq 3$. Of course, this is still not optimal. Further improvements could be obtained by slightly decreasing $\alpha \in [0,1)$ in $||\boldsymbol{X}_i||_{\alpha R}$ in each step. □

1.4 A First Application

Let us consider the following "rapidly forced" differential equation

$$\begin{aligned} \frac{d}{dt}\boldsymbol{x} &= \epsilon\,\boldsymbol{Z}(\boldsymbol{x}, s)\,, \\ \frac{d}{dt}s &= 1\,, \end{aligned}$$

where $\boldsymbol{Z}$ is 1-periodic in s and $||\boldsymbol{Z}(\boldsymbol{x}, s)|| < 1$ for $s \in [0,1]$ and all $\boldsymbol{x} \in \mathcal{U} \subset \mathrm{R}^n$. Let $\boldsymbol{G}$ denote the time-1-flow map of the differential equation. Then the map $\boldsymbol{G}$ certainly satisfied $||\boldsymbol{G}(\boldsymbol{x}) - \boldsymbol{x}|| < \epsilon$ for $\boldsymbol{x} \in \mathcal{K} \subset \mathcal{U}$ and Theorem 1.1 can be applied provided $\boldsymbol{Z}$ is analytic. Thus the time-1-flow of the rapidly forced differential equation is equivalent to the time-1-flow of an autonomous differential equation up to terms exponentially small in ϵ. This result was first stated by NEISHTADT in [88]. The first term in the asymptotic expansion of the autonomous (interpolating) vector field is given by

$$\boldsymbol{X}_1(\boldsymbol{x}) \quad = \quad \epsilon \int_0^1 \boldsymbol{Z}(\boldsymbol{x}, s)\,ds\,.$$

2

Backward Error Analysis

In this chapter, we consider the relationship between solutions to a given system of ordinary differential equations

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

numerical approximations

$$\boldsymbol{x}_{n+1} \quad = \quad \boldsymbol{\Psi}_{\Delta t}\,(\boldsymbol{x}_n) \tag{0.1}$$

to them, and solutions to associated modified equations

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

The vector fields $\tilde{\boldsymbol{X}}_i(\Delta t)$ are formulated in terms of an asymptotic expansion in 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 [123], Griffiths & Sanz-Serna [48], Beyn [18], Feng [37], Fiedler & Scheurle [38], and Sanz-Serna [108].

More recently, general formulas for the computation of the modified vector fields $\tilde{\boldsymbol{X}}_i$ have been derived by Hairer [53], Calvo, Murua & Sanz-Serna [28], Benettin & Giorgilli [16], and Reich [94]. In papers by Neishtadt [88], Benettin & Giorgilli [16], and Hairer & Lubich [55], the question of closeness of the numerical approximations and the solutions of the modified equations has been addressed. It has also been shown by Hairer [53], Calvo, Murua & Sanz-Serna [28], Reich [94], and Benettin & Giorgilli [16] that for symplectic discretizations, the modified vector fields $\tilde{\boldsymbol{X}}_i$ are Hamiltonian. For special cases see also the papers by Auerbach & Friedman [9] and Yoshida [124].

2.1 Perturbed Vector Fields for Numerical Integration

Let us now consider a real analytic vector field

$$\frac{d}{dt}\boldsymbol{x} \quad = \quad \boldsymbol{Z}\,(\boldsymbol{x}), \tag{1.2}$$

$\boldsymbol{Z} : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$ and its discretization by a one step method [56],[32]

$$\boldsymbol{x}_{n+1} \quad = \quad \boldsymbol{\Psi}_{\Delta t}\,(\boldsymbol{x}_n) \;=\; \boldsymbol{x}_n + \Delta t\,\boldsymbol{\psi}(\boldsymbol{x}_n, \Delta t)\,. \tag{1.3}$$

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

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

$\boldsymbol{z} \in \mathcal{U}$, and that there exists an appropriate constant $M > 0$ such that

$$||\,\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}) - \boldsymbol{x}\,|| \quad \leq \quad \Delta t\, M$$

for all $\boldsymbol{x}$ in a compact subset $\mathcal{K}$ of $\mathcal{U}$ and all $\Delta t > 0$ sufficiently small. As in the previous chapter, we look for a vector field $\boldsymbol{X}$ such that

$$\boldsymbol{\Phi}_{1,\boldsymbol{X}} \quad \approx \quad \boldsymbol{\Psi}_{\Delta t}$$

or, equivalently,

$$\boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}} \quad \approx \quad \boldsymbol{\Psi}_{\Delta t} \quad \text{with} \quad \tilde{\boldsymbol{X}} \;:=\; \frac{1}{\Delta t}\boldsymbol{X}\,.$$

Upon defining

$$\boldsymbol{G}_\xi \quad = \quad \mathbf{id} + \xi\,\boldsymbol{Y}_0\,, \qquad \xi \in [0,\epsilon]\,,$$

with

$$\boldsymbol{Y}_0(\boldsymbol{x}) \quad := \quad \frac{1}{M}\boldsymbol{\psi}(\boldsymbol{x},\Delta t) \;\leq\; 1 \quad \text{and} \quad \epsilon \;:=\; \Delta t\, M\,,$$

we can apply the recursion (1.7)-(1.8) from Chapter 1. Thus, with $\boldsymbol{G} = \boldsymbol{\Psi}_{\Delta t}$ and $\epsilon = \Delta t M$, we can also apply Theorem 1.1. Specifically:

Corollary 2.1. Let $\boldsymbol{\Psi}_{\Delta t} : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$ be a real analytic map close to the identity on a compact set $\mathcal{K} \subset \mathcal{U}$, i.e.,

$$||\,\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}) - \boldsymbol{x}\,|| \quad < \quad \Delta t\, M \qquad \text{for all } \boldsymbol{x} \in \mathcal{K}$$

and all $\Delta t > 0$ sufficiently small. Then there exists a family of real analytic vector fields $\tilde{\boldsymbol{X}}(\Delta t) : \mathcal{V} \subset \mathrm{R}^n \to \mathrm{R}^n$ such that

$$||\,\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}) - \boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}(\boldsymbol{x})\,|| \quad \leq \quad 6\,\Delta t\, b\, M\, e^{-\gamma/\Delta t}\,, \qquad \text{for all } \boldsymbol{x} \in \mathcal{K}\,.$$

Here $\gamma = R/(2cMe)$, $b = 10$, $c = 150$, and $R > 0$ such that, for all $\boldsymbol{x}_0 \in \mathcal{K}$ and all $\Delta t > 0$ sufficiently small,

$$||\,\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}) - \boldsymbol{x}\,|| \quad \leq \quad \Delta t\, M \tag{1.4}$$

on the complex ball of radius R around $\boldsymbol{x}_0 \in \mathcal{K}$. □

Remark 2.1. Theorem 1.1 yields the exponential estimate for the time one flow map of an appropriate vector field $\boldsymbol{X}$. This is equivalent to the time-Δt-flow of the vector field $\tilde{\boldsymbol{X}} := \boldsymbol{X}/\Delta t$. Furthermore, for a method $\boldsymbol{G}_{\Delta t}$ of order $p \geq 1$, we have

$$\boldsymbol{Z}(\boldsymbol{x}) - \tilde{\boldsymbol{X}}(\boldsymbol{x}, \Delta t) \quad = \quad \mathcal{O}(\Delta t^p)\,.$$

□

Remark 2.2. The discrete evolution (1.3) can now be considered as the discretization of the modified vector field $\tilde{\boldsymbol{X}}$ (as long as the numerical solution does not leave the compact set $\mathcal{K}$). According to Corollary 2.1 and standard results in numerical analysis [56],[32], the global error

$$\boldsymbol{e}_n\,(\boldsymbol{x}) \quad := \quad \boldsymbol{\Phi}_{n\Delta t, \tilde{\boldsymbol{X}}}(\boldsymbol{x}) - [\,\boldsymbol{\Psi}_{\Delta t}\,]^n\,(\boldsymbol{x})$$

after n steps with step-size Δt is bounded by

$$||\,\boldsymbol{e}_n\,(x)\,|| \quad \leq \quad \frac{2\,b\,M}{\tilde{L}}\,\left(e^{n\Delta t \tilde{L}} - 1\right)\,e^{-\gamma/\Delta t}$$

where $\tilde{L} \geq 0$ is the Lipschitz constant of the modified vector field $\tilde{\boldsymbol{X}}$ on $\mathcal{K}$. Thus the global error $\boldsymbol{e}_n$ remains exponentially small over a time interval $T = n\Delta t < \gamma/(2\Delta t \tilde{L})$ [16]. □

Remark 2.3. In [38], Fiedler & Scheurle show that $\boldsymbol{\Psi}_{\Delta t}$ is equivalent to the time-Δt-flow of a non-autonomous differential equation

$$\frac{d}{dt}\boldsymbol{z} \quad = \quad \boldsymbol{Z}(\boldsymbol{x}) + \boldsymbol{F}(\boldsymbol{x}, t; \Delta t) \tag{1.5}$$

with $\boldsymbol{F}(\Delta t)$ a vector field Δt-periodic in t and

$$||\,\boldsymbol{F}(\boldsymbol{x}, t; \Delta t)\,|| \quad = \quad \mathcal{O}(\Delta t^p)\,,$$

$p \geq 1$ the order of $\boldsymbol{\Psi}_{\Delta t}$. In view of Corollary 2.1, we can use the same construction to show that $\boldsymbol{\Psi}_{\Delta t}$ is equivalent to the time-Δt-flow of the non-autonomous differential equation

$$\frac{d}{dt}\boldsymbol{x} \quad = \quad \tilde{\boldsymbol{X}}(\boldsymbol{x}) + e^{-\gamma/\Delta t}\,\tilde{\boldsymbol{F}}(\boldsymbol{x}, t; \Delta t)$$

where $\tilde{\boldsymbol{X}}$ is the modified vector field of Corollary 2.1 and $\tilde{\boldsymbol{F}}$ is a vector field Δt-periodic in t. Furthermore, because of Corollary 2.1, there exists a constant $c > 0$ such that

$$||\,\tilde{\boldsymbol{F}}(\boldsymbol{x}, t; \Delta t)\,|| \quad \leq \quad c$$

for $\boldsymbol{x} \in \mathcal{K}$, $t \in [0, \Delta t]$, and Δt sufficiently small. In fact, equation (1.5) was taken as the starting point by Neishtadt in his paper [88] and averaging in time was used to show that the non-autonomous perturbation can be made exponentially small in Δt. To the knowledge of the author, this was the first mentioning of a result as stated in

Corollary 2.1. □

Remark 2.4. In some cases stronger results than the one presented in Corollary 2.1 can be derived. We like to mention the behavior of a numerical discretization near a hyperbolic fixed point [35], the "numerical" flow box theorem [40], and the discretization of flows on domains of attraction [41]. □

Let us now discuss the Taylor series expansion of the modified vector field $\tilde{\boldsymbol{X}}(\Delta t)$ in terms of Δt in more detail. This will be useful in Section 2.2 when we consider geometric properties of $\tilde{\boldsymbol{X}}$. In this context, it is more appropriate to use $\boldsymbol{G} = \boldsymbol{\Psi}_{\Delta t}$ and $\hat{\boldsymbol{X}}_0 = \Delta t\, \boldsymbol{Z}$ in the recursion (1.2)-(1.3) from Section 2 which immediately implies $\boldsymbol{\Delta}\hat{\boldsymbol{X}}_1 = \mathcal{O}(\Delta t^{p+1})$. More generally, similar to the proof of Lemma 1.1, it is easy to show that we have

$$\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} \quad := \quad \boldsymbol{\Psi}_{\Delta t} - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i} \;=\; \mathcal{O}(\Delta t^{i+p+1})\,, \qquad (i \geq 0)\,,$$

where $\hat{\boldsymbol{X}}_{i+1} := \hat{\boldsymbol{X}}_i + \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}$ as before. Thus we consider the limit

$$\lim_{\tau\to 0} \frac{1}{\tau^{i+p+1}}\, \boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}(\tau) \quad = \quad \lim_{\tau\to 0} \frac{\boldsymbol{\Psi}_\tau - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i(\tau)}}{\tau^{i+p+1}}$$

where

$$\hat{\boldsymbol{X}}_i(\tau) \quad := \quad \hat{\boldsymbol{X}}_i(\Delta t = \tau)$$

and

$$\boldsymbol{\Psi}_\tau\,(\boldsymbol{x}) \quad := \quad \boldsymbol{\Psi}_{\Delta t=\tau} \;=\; \boldsymbol{x} + \tau\, \boldsymbol{\psi}(\boldsymbol{x},\tau)\,.$$

This yields

$$\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1}(\Delta t) \quad := \quad \Delta t^{p+i+1} \lim_{\tau\to 0} \frac{\boldsymbol{\Psi}_\tau - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i(\tau)}}{\tau^{i+p+1}} + \mathcal{O}(\Delta t^{p+i+2})\,.$$

Since

$$\lim_{\tau\to 0} \frac{\boldsymbol{\Psi}_\tau - \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i(\tau)}}{\tau^{i+p+1}} \quad = \quad \frac{1}{(i+p+1)!} \left[\frac{\partial^{i+p+1}}{\partial \tau^{i+p+1}}\, \boldsymbol{\Psi}_\tau - \frac{\partial^{i+p+1}}{\partial \tau^{i+p+1}}\, \boldsymbol{\Phi}_{1,\hat{\boldsymbol{X}}_i(\tau)} \right]_{\tau=0},$$

we are led to the modified recursion [94]:

$$\boxed{\begin{aligned} \boldsymbol{\Delta X}_{i+1} \quad &:= \quad \frac{\Delta t^l}{l!} \left[\frac{\partial^l}{\partial \tau^l} \boldsymbol{\Psi}_\tau - \frac{\partial^l}{\partial \tau^l} \boldsymbol{\Phi}_{1,\boldsymbol{X}_i(\tau)} \right]_{\tau=0},\; l = p+i+1, && (1.6)\\ \boldsymbol{X}_{i+1} \quad &:= \quad \boldsymbol{X}_i + \boldsymbol{\Delta X}_{i+1} && (1.7)\\ &\text{with } \boldsymbol{X}_0 := \Delta t\, \boldsymbol{Z}. \end{aligned}}$$

The modified vector fields $\tilde{\boldsymbol{X}}_i$ are defined by

$$\tilde{\boldsymbol{X}}_i \quad := \quad \frac{1}{\Delta t} \boldsymbol{X}_i$$

and we have

$$\boldsymbol{\Psi}_{\Delta t} - \boldsymbol{\Phi}_{\Delta t, \tilde{\boldsymbol{X}}_i} \quad = \quad \mathcal{O}(\Delta t^{p+i+1}), \qquad i \geq 0 .$$

Using the same notations as in Chapter 1, the proof of Theorem 1.1 can be adjusted to also cover the recursion (1.6)-(1.7). The major difference is that we also have to provide an estimate for the derivatives $\partial_\tau^i \boldsymbol{\Psi}_\tau$, $i \geq 1$, which is the subject of the following lemma.

Lemma 2.1. Let us assume that the vector field $\boldsymbol{Z}$ in (1.2) is real analytic and that there is a compact subset $\mathcal{K}$ of phase space and constants K, $R > 0$ such that, for all $\boldsymbol{x}_0 \in \mathcal{K}$,

$$||\, \boldsymbol{Z}(\boldsymbol{x}) \,|| \quad \leq \quad K$$

on the complex ball of radius R around $\boldsymbol{x}_0 \in \mathcal{K}$. We also assume that the numerical method $\boldsymbol{\Psi}_{\Delta t}$ is real analytic. Then there exists a constant $M \geq K$ such that (i)

$$||\, \boldsymbol{\Psi}_{\Delta t} - \mathbf{id} \,||_{\alpha R} \quad \leq \quad \Delta t \, M \qquad \text{for } |\Delta t| \, \leq \, \frac{(1-\alpha)\, R}{M} \, ,$$

$\alpha \in [0, 1)$, and (ii)

$$\frac{\Delta t^{i+1}}{(i+1)!} \, || \, \frac{\partial^{i+1}}{\partial \tau^{i+1}} \boldsymbol{\Psi}_{\tau=0} \, ||_{\alpha R} \quad \leq \quad \Delta t \, M \left(\frac{\Delta t \, M}{(1-\alpha)\, R} \right)^i . \tag{1.8}$$

□

Proof. Under the given assumptions, the flow map $\boldsymbol{\Phi}_{t,\boldsymbol{Z}}$ satisfies

$$||\, \boldsymbol{\Phi}_{\Delta t, \boldsymbol{Z}} - \mathbf{id} \,||_{\alpha R} \quad \leq \quad \Delta t \, K \qquad \text{for } |\Delta t| \, \leq \, \frac{(1-\alpha)\, R}{K} \, ,$$

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

$$||\, \boldsymbol{\Psi}_{\Delta t} - \mathbf{id} \,||_{\alpha R} \quad \leq \quad \Delta t \,(K + \Delta K) \qquad \text{for } |\Delta t| \, \leq \, \frac{(1-\alpha)\, R}{K + \Delta K} \, .$$

The estimate (ii) follows from Cauchy's estimate. □

Remark 2.5. Let us consider a Runge-Kutta method with coefficients $\{a_{ij}\}$ and $\{b_i\}$ [56] satisfying

$$\sum_j |a_{ij}| \; \leq \; d \quad \text{and} \quad \sum_i |b_i| \; \leq \; d \, ,$$

$d \geq 1$, and assume that the Runge-Kutta method uniquely[1] defines a real analytic map $\boldsymbol{\Psi}_{\Delta t}$ for all step-sizes $\Delta t \leq R/K$. Then we have $M = d\,K$ in Lemma 2.1. This follows from the fact that, under the stated assumptions, all the stage variables will be in $B_R(\boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathcal{K}$, where the vector field $\boldsymbol{Z}$ is bounded by the constant K. □

Theorem 2.1. Let the assumption of Lemma 2.1 be satisfied. Then there exists a family of real analytic vector fields $\tilde{\boldsymbol{X}}(\Delta t) : \mathcal{V} \subset \mathrm{R}^n \to \mathrm{R}^n$, $\mathcal{K} \subset \mathcal{V} \subset \mathcal{U}$, such that, for all $\boldsymbol{x}_0 \in \mathcal{K}$,

$$|| \, \boldsymbol{\Psi}_{\Delta t} - \boldsymbol{\Phi}_{\Delta t, \tilde{\boldsymbol{X}}} \, ||_{R/2} \quad \leq \quad 6 \, \Delta t \, b \, M \, e^{-p} \, e^{-\gamma/\Delta t}$$

with $\gamma = R/(4cMe)$, $b = 12$, $c = 40$, and $p \geq 1$ the order of the method. The modified vector fields $\tilde{\boldsymbol{X}}(\Delta t)$ satisfy the estimate

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

with $d_p \geq 1$ a constant depending on the order p of the method. For example, $d_1 = 1.0$, $d_2 = 1.0$, $d_3 = 1.0$, and $d_4 = 1.3$. □

Proof. With $\epsilon := \Delta t M$, the estimates for the recursion (1.6)-(1.7) are similar to the ones given in the proof of Theorem 1.1. Thus we only point out the differences. We know that $\boldsymbol{X}_0 = \Delta t \boldsymbol{Z}$. Thus $||\boldsymbol{X}_0||_R \leq \Delta t M$. For $\boldsymbol{\Delta X}_1$ we obtain the estimate

$$\begin{aligned} || \, \boldsymbol{\Delta X}_1 \, ||_{\alpha R} \quad &\leq \quad \frac{\Delta t^{p+1}}{(p+1)!} \left[|| \, \frac{\partial^{p+1}}{\partial \tau^{p+1}} \boldsymbol{\Phi}_{1, \boldsymbol{X}_0(\tau=0)} \, ||_{\alpha R} + || \, \frac{\partial^{p+1}}{\partial \tau^{p+1}} \boldsymbol{\Psi}_{\tau=0} \, ||_{\alpha R} \right] , \\ &\leq \quad 2 \, \Delta t \, M \left(\frac{\Delta t \, M}{(1-\alpha) \, R} \right)^p . \end{aligned}$$

Here we have used (1.8) with $i = p$ and (compare Lemma 1.2)

$$\frac{\Delta t^{p+1}}{(p+1)!} \, || \, \frac{\partial^{p+1}}{\partial \tau^{p+1}} \boldsymbol{\Phi}_{1, \boldsymbol{X}_0(\tau=0)} \, ||_{\alpha R} \quad \leq \quad \Delta t \, M \left(\frac{\Delta t \, M}{(1-\alpha) \, R} \right)^p .$$

The estimate of Lemma 1.3 is now replaced by the estimate

$$|| \, \boldsymbol{\Delta X}_i \, ||_{\alpha R} \quad \leq \quad \Delta t \, b \, M \left(\frac{c \, (i-1) \, \Delta t \, M}{(1-\alpha) \, R} \right)^{p+i-1} . \tag{1.9}$$

Thus

$$\begin{aligned} || \, \boldsymbol{X}_j(\Delta t) \, ||_{\alpha R} \quad &\leq \quad \sum_{i=0}^{j} || \, \boldsymbol{\Delta X}_i(\Delta t) \, ||_{\alpha R} \, , \\ &\leq \quad \Delta t \, M \left[1 + 2 \left(\frac{\Delta t \, M}{(1-\alpha) \, R} \right)^p + \sum_{i=2}^{j} b \left(\frac{c \, (i-1) \, \Delta t \, M}{(1-\alpha) \, R} \right)^{p+i-1} \right] . \end{aligned}$$

[1]For an implicit method, the solution can be obtained by fixpoint iteration for Δt sufficiently small.

Following the proof of Lemma 1.3, we obtain

$$|| \boldsymbol{X}_j(\Delta t) ||_{(\alpha+\delta(1-\alpha))R} \;\leq\; (b-1)\,\Delta t_0\, M \;\leq\; \delta\,(1-\alpha)\,R\,, \qquad \delta \;:=\; \frac{b-1}{j\,c}\,,$$

for all

$$\Delta t \;\leq\; \frac{(1-\alpha)\,R}{c\,j\,M} \;=:\; \Delta t_0\,,$$

$b \geq 12$, and $c \geq 40$. Here we have used that

$$\sum_{i=2}^{j} \left(\frac{i-1}{(1-\delta)\,j} \right)^{i} \;\leq\; 0.82$$

for $\delta \leq 0.3/j$ and all $j \geq 2$. This implies that

$$1 + 2\left(\frac{1}{(1-\delta)\,c\,j} \right)^{p} + b \sum_{i=2}^{j} \left(\frac{i-1}{(1-\delta)\,j} \right)^{p+i-1} \;\leq\; b-1$$

for all $p \geq 1$, $j \geq 2$, $b \geq 12$, and $c \geq 40$. Next we introduce the vector-valued function

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

and observe that

$$|| \boldsymbol{f}(\tau) ||_{\alpha R} \;\leq\; \delta\,(1-\alpha)\,R$$

for $|\tau| \leq \Delta t_0$. Following the proof of Lemma 1.3 and taking into account that $\boldsymbol{\Delta X}_{j+1}$ is defined by (1.6), we obtain

$$\begin{aligned} || \boldsymbol{\Delta X}_{j+1} ||_{\alpha R} \;&\leq\; \frac{\Delta t^{p+j+1}}{(p+j+1)!} \,||\, \frac{\partial^{p+j+1}}{\partial \tau^{p+j+1}} \boldsymbol{\Phi}_{1,\boldsymbol{X}_j(\tau=0)} ||_{\alpha R} \,+ \\ &\qquad\qquad + \frac{\Delta t^{p+j+1}}{(p+j+1)!} \,||\, \frac{\partial^{p+j+1}}{\partial \tau^{p+j+1}} \boldsymbol{\Psi}_{\tau=0} ||_{\alpha R}\,, \\ &\leq\; \frac{\Delta t^{p+j+1}}{(p+j+1)!} \left[||\, \frac{\partial^{p+j+1}}{\partial \tau^{p+j+1}} \boldsymbol{f}_{\tau=0} ||_{\alpha R} + ||\, \frac{\partial^{p+j+1}}{\partial \tau^{p+j+1}} \boldsymbol{\Psi}_{\tau=0} ||_{\alpha R} \right], \\ &\leq\; (b-1)\,\Delta t\, M \left(\frac{c\,j\,\Delta t\, M}{(1-\alpha)\,R} \right)^{p+j} + \Delta t\, M \left(\frac{\Delta t\, M}{(1-\alpha)\,R} \right)^{p+j}, \\ &\leq\; b\,\Delta t\, M \left(\frac{c\,j\,\Delta t\, M}{(1-\alpha)\,R} \right)^{p+j} \end{aligned}$$

which verifies our claim. Lemma 1.4 and the choice of the "optimal" number of iterations $i_*(\Delta t)$ carry over to Theorem 2.1 and

$$\begin{aligned} || \boldsymbol{\Psi}_{\Delta t} - \boldsymbol{\Phi}_{1,\boldsymbol{X}_{i_*}} ||_{R/2} \;&\leq\; 2\,\Delta t\, b\, M\, e^{-p}\, e^{-i_*}\,, \\ &\leq\; 2\,\Delta t\, b\, M\, e^{-p}\, e^{-i_\circ+1}\,, \\ &\leq\; 6\,\Delta t\, b\, M\, e^{-p}\, e^{-i_\circ} \end{aligned}$$

where $i_*(\Delta t)$ is the integer part of

$$i_o(\Delta t) \quad := \quad \frac{R}{4\,c\,\Delta t\,M\,e}\,.$$

We define

$$\tilde{\boldsymbol{X}}(\Delta t) \quad := \quad \frac{1}{\Delta t}\,\boldsymbol{X}_{i_*}(\Delta t)$$

which completes the first part of the proof.

The difference between the modified vector fields $\tilde{\boldsymbol{X}}(\Delta t)$ and $\boldsymbol{Z}$ is given by

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

Next we use

$$\Delta t \quad \leq \quad \frac{R}{4\,c\,i_*\,M\,e}$$

to obtain

$$\begin{aligned} ||\,\tilde{\boldsymbol{X}} - \boldsymbol{Z}\,||_{R/2} \quad &\leq \quad M\,\left(\frac{2\,c\,\Delta t\,M}{R}\right)^p \left[2 + b\sum_{i=2}^{i_*}\frac{(i-1)^p}{(2e)^{i-1}}\left(\frac{i-1}{i_*}\right)^{i-1}\right], \\ &\leq \quad M\,\left(\frac{2\,c\,\Delta t\,M}{R}\right)^p [2 + b\,d_p\,0.89]\,, \\ &\leq \quad d_p\,b\,M\,\left(\frac{2\,c\,\Delta t\,M}{R}\right)^p. \end{aligned}$$

Here $d_p \geq 1$ is chosen such that

$$d_p \quad \geq \quad \frac{i^p}{(2e)^i}$$

for all $i \geq 1$. □

Remark 2.6. Again it was not our intention to provide an optimal estimate for the constants b and c. A better estimate can, for example, be obtained by replacing the estimate (1.9) in the proof of Theorem 2.1 by

$$||\,\boldsymbol{\Delta X}_i\,||_{\alpha R} \quad \leq \quad b_i\,\xi\,\left(\frac{c_i\,(i-1)\,\xi}{(1-\alpha)\,R}\right)^{p+i-1}.$$

Here c_i and b_i are appropriate constants. Similar to Remark 1.3, one can, for example, choose these constants such that $c_i \leq 16$, $b_i \leq 4.1$ for $i \geq 2$ and

$$1 + 2\left(\frac{1}{c_{j+1}(j - \frac{b_{j+1}}{c_{j+1}})}\right)^p + \sum_{i=2}^{j} b_i\,\left(\frac{c_i(i-1)}{c_{j+1}(j - \frac{b_{j+1}}{c_{j+1}})}\right)^{p+i-1} \quad \leq \quad b_{j+1} - \left(\frac{1}{c_{j+1}j}\right)^{p+j}$$

for all $j \geq 2$. Thus Theorem 2.1 is also valid with $c = 16$ and $b = 4.1$. □

2.2 Geometric Properties of Backward Error Analysis

In this section, we consider differential equations (1.2) whose corresponding vector field $\boldsymbol{Z}$ belongs to a certain linear subspace $\mathfrak{g}$ of the infinite dimensional Lie algebra[2] of smooth vector fields on R^n [58],[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 [58] of diffeomorphisms on R^n such that

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

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

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 $\mathfrak{G}$ [58],[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 can be derived [94]:

Theorem 2.2. Let us assume that the vector field $\boldsymbol{Z}$ in

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

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

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

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 $\boldsymbol{X}_i$, $i = 1, \ldots, s$, defined through the recursion (1.6)-(1.7) belong to $\mathfrak{g}$, i.e.

$$\boldsymbol{X}_i \quad \in \quad \mathfrak{g}\,,$$

and the vector field $\tilde{\boldsymbol{X}}$ in Theorem 2.1 satisfies $\tilde{\boldsymbol{X}} \in \mathfrak{g}$. $\square$

Proof. The statement is certainly true for $\boldsymbol{X}_0 = \Delta t\,\boldsymbol{Z}$. Let us assume that it also holds for $\boldsymbol{X}_i$, i.e., $\boldsymbol{X}_i(\Delta t) \in \mathfrak{g}$ for all $\Delta t \geq 0$ sufficiently small. Since

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

[2] The algebraic operation is the Lie bracket $[\boldsymbol{X},\boldsymbol{Y}]$ of two vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$ [6].

and

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

as well as

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

we have

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

and $\boldsymbol{\Delta X}_{i+1}(\Delta t) \in \mathfrak{g}$ for all $\Delta t \geq 0$ sufficiently small. This implies $\boldsymbol{X}_{i+1}(\Delta t) \in \mathfrak{g}$ as required. □

Remark 2.7. Often the linear subspace $\mathfrak{g}$ is, in fact, a subalgebra under the Lie bracket [6]

$$[\boldsymbol{X},\boldsymbol{Y}] := \frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{X}\cdot\boldsymbol{Y} - \frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{Y}\cdot\boldsymbol{X}\,, \tag{2.10}$$

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 2.2. □.

Let us discuss four examples:

Example 2.1. 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 (2.10). In other words

$$\boldsymbol{\partial_x} F \cdot \boldsymbol{X} = 0 \tag{2.11}$$

and

$$\boldsymbol{\partial_x} F \cdot \boldsymbol{Y} = 0 \tag{2.12}$$

imply that

$$\boldsymbol{\partial_x} F \cdot [\boldsymbol{X},\boldsymbol{Y}] = 0\,. \tag{2.13}$$

To show this we differentiate (2.11) w.r.t. $\boldsymbol{x}$ which gives

$$\boldsymbol{X}^T \cdot \boldsymbol{\partial_{xx}} F + \boldsymbol{\partial_x} F \cdot \boldsymbol{\partial_x X} = \boldsymbol{0}\,.$$

The same procedure is applied to (2.12). Using these identities and the definition (2.10) in (2.13) yields 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 $\boldsymbol{X} := \partial_\tau \boldsymbol{\Psi}_{\tau=0} \in \mathfrak{g}$ since

$$\partial_\tau F \circ \boldsymbol{\Psi}_{\tau=0} = \boldsymbol{\partial_x} F \cdot \boldsymbol{X} = 0\,.$$

Thus, $T_{\mathbf{id}}\,\mathfrak{G} = \mathfrak{g}$ and we can apply Theorem 2.2. In particular, if a numerical method $\boldsymbol{\Psi}_{\Delta t}$ satisfies

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

then the modified vector field $\tilde{\boldsymbol{X}}$ possesses F as a first integral. The same result was recently derived by Gonzales & Stuart [49] by a contradiction argument. □

Example 2.2. 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:

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

$\boldsymbol{X} := \partial_\tau \boldsymbol{\Psi}_{\tau=0}$. Thus, if the numerical method $\boldsymbol{\Psi}_{\Delta t}$ is volume conserving, then the modified vector field $\tilde{\boldsymbol{X}}$ is divergence-free. □

Example 2.3. Let an involution[3] $\boldsymbol{S}$ 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 (2.10). 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

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

which implies that $\boldsymbol{X} := \partial_\tau \boldsymbol{\Psi}_{\tau=0} \in \mathfrak{g}$. It follows that $T_{\mathbf{id}}\,\mathfrak{G} = \mathfrak{g}$ and we can apply Theorem 2.2. Thus, if a numerical method $\boldsymbol{\Psi}_{\Delta t}$ satisfies the time-reversal symmetry, then the modified vector field $\tilde{\boldsymbol{X}}$ is time-reversible. The same result has been first stated by Hairer & Stoffer in [57]. □.

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

Example 2.4. 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 $F, G : \mathcal{P} \to \mathrm{R}$, $\boldsymbol{X} := \partial_\tau \boldsymbol{\Psi}_{\tau=0}$. 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 [6].

If the discrete evolution (1.3) 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 2.2 that the modified vector fields $\tilde{\boldsymbol{X}}_i$ are Hamiltonian vector fields on $\mathcal{P}$. If we assume furthermore, that the Hamiltonian H is analytic and the discrete evolution $\boldsymbol{\Psi}_{\Delta t}$ satisfies the conditions of Theorem 2.1, then one has (i)

$$|\,\tilde{H}(\boldsymbol{x}) - H(\boldsymbol{x})\,| \;=\; \mathcal{O}(\Delta t^p)\,,$$

$\tilde{H}$ the Hamiltonian of the modified vector field $\tilde{\boldsymbol{X}}$, i.e. $\tilde{\boldsymbol{X}} = \{\mathbf{id}, \tilde{H}\}$, $p \geq 1$ the order of the method, and (ii)

$$|\,\tilde{H}(\boldsymbol{x}_n) - \tilde{H}(\boldsymbol{x}_0)\,| \;\leq\; c\,e^{-\gamma/(2\Delta t)}\,, \quad \boldsymbol{x}_n = [\boldsymbol{\Psi}_{\Delta t}]^n(\boldsymbol{x}_0)\,, \tag{2.14}$$

γ, $c > 0$ appropriate constants, over time intervals

$$T \;=\; \Delta t\, n \;\leq\; e^{\gamma/(2\Delta t)}\,.$$

The estimate (2.14) follows from the fact that the global error in $\tilde{H}(\boldsymbol{x}_n)$ grows only linearly with $n \geq 1$ [16],[55] and that after one step

$$\begin{aligned} |\,\tilde{H}(\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x})) - \tilde{H}(\boldsymbol{x})\,| &= |\,\tilde{H}(\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x})) - \tilde{H}(\boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}(\boldsymbol{x}))\,|\,, \\ &= \mathcal{O}(\Delta t\, e^{-\gamma/\Delta t})\,. \end{aligned}$$

Thus

$$|\,\tilde{H}(\boldsymbol{x}_n) - \tilde{H}(\boldsymbol{x}_0)\,| \;=\; \mathcal{O}(T\, e^{-\gamma/\Delta t})$$

and the desired estimates follow. □

2.3 An Application: Adiabatic Invariants

Let us consider a time-dependent Hamiltonian system on R^2 with real analytic Hamiltonian $H(q, p, t)$, $q, p \in \mathrm{R}$. Using the extended Hamiltonian

$$E(q, p, s, e) \;:=\; H(q, p, s) - e\,,$$

the corresponding equations of motion

$$\begin{aligned}
\frac{d}{dt}q &= +\nabla_p\, E(q,p,s,e) &&= +\nabla_p\, H(q,p,s)\,,\\
\frac{d}{dt}p &= -\nabla_q\, E(q,p,s,e) &&= -\nabla_q\, H(q,p,s)\,,\\
\frac{d}{dt}e &= +\nabla_s\, E(q,p,s,e) &&= +\nabla_s\, H(,q,p,s)\,,\\
\frac{d}{dt}s &= -\nabla_e\, E(q,p,s,e) &&= 1
\end{aligned}$$

are Hamiltonian in the extended phase space R^4. We assume that the Hamiltonian H is of the form

$$H(q,p,s) \;=\; \frac{1}{2}\,p^2 + V(q,s)\,.$$

For example,

$$V(q,s) \;=\; \frac{1}{2}\,\omega(s)^2\,q^2\,. \tag{3.15}$$

Then the equations of motion can be discretized by a generalization of the well-known Verlet method [56], i.e.,

$$\begin{aligned}
q_{n+1} &= q_n + \Delta t\, p_{n+1/2}\,,\\
p_{n+1/2} &= p_n - \frac{\Delta t}{2}\,\nabla_q\, V(q_n, s_n)\,,\\
p_{n+1} &= p_{n+1/2} - \frac{\Delta t}{2}\,\nabla_q\, V(q_{n+1}, s_{n+1})\,,\\
e_{n+1} &= e_n + \frac{\Delta t}{2}\,[\nabla_s\, V(q_n, s_n) + \nabla_s\, V(q_{n+1}, s_{n+1})]\,,\\
s_{n+1} &= s_n + \Delta t\,.
\end{aligned}$$

This discretization is symplectic and, therefore, according to Theorem 2.2, there exists a modified Hamiltonian vector field $\tilde{\boldsymbol{X}}$ with modified Hamiltonian $\tilde{E}$, i.e. $\tilde{\boldsymbol{X}} = \{\mathbf{id}, \tilde{E}\}$, such that its time-one-flow map is exponentially close to the discrete evolution $\boldsymbol{\Psi}_{\Delta t}$ given by the Verlet discretization. Furthermore, because the equation of motion in the variable s is integrated exactly, the modified Hamiltonian $\tilde{E}$ is again of the form

$$\tilde{E}(q,p,s,e) \;=\; \tilde{H}(q,p,s;\Delta t) + e\,,$$

$\tilde{H}(q,p,s)$ an appropriate function. Since the Verlet method is second order, we also have

$$\tilde{H}(q,p,s;\Delta t) - H(q,p,s) = \mathcal{O}(\Delta t^2)\,.$$

Let us assume now that, for fixed s, the Hamiltonian $H(q,p,s)$ has periodic solutions with period $T(s)$ and that $H(q,p,s)$ varies slowly in time compared to the periodic motion in (q,p), i.e.,

$$T(s)\,|\,\frac{\partial}{\partial s}H(q,p,s)\,| \quad \leq \quad \epsilon \;\ll\; 1$$

for all s. Then the corresponding equations of motion possess an *adiabatic invariant* J which is defined as the area enclosed by the periodic motion $(q(t),p(t))$, $t\in[0,T(s)]$, for fixed s [7]. The adiabatic invariant J remains almost constant over an exponentially long period of time [88] (see also Section 4.6.1), i.e.,

$$|\,J(q(t),p(t),t) - J(q(0),p(0),0)\,| \quad \leq \quad c_1\,t\,e^{-c_2/\epsilon} + \mathcal{O}(\epsilon)\,, \tag{3.16}$$

$c_1,c_2 > 0$ appropriate constants. Now, for fixed s, the perturbed Hamiltonian $\tilde{H}(q,p,s;\Delta t)$ will also possess periodic solutions with period $\tilde{T}(s)$ and $\tilde{H}(q,p,s;\Delta t)$ varies slowly in time compared to the periodic motion in (q,p), i.e.,

$$\tilde{T}(s)\,|\,\frac{\partial}{\partial s}\tilde{H}(q,p,s)\,| \quad \leq \quad \tilde{\epsilon}$$

with $|\tilde{\epsilon}-\epsilon| = \mathcal{O}(\Delta t^2)$ since the Verlet method is second order. Thus the perturbed Hamiltonian equations of motion has a modified adiabatic invariant $\tilde{J}$ (the area enclosed by the periodic motion of the modified Hamiltonian $\tilde{H}$). Application of *normal form* theory implies [88] (see also Section 4.6.1) that there exists a slightly perturbed $\hat{J}$ such that $\hat{J}$ is preserved along solution curves $(q(t),p(t))$ of $\tilde{\boldsymbol{X}}$ up to exponentially small terms, i.e.

$$|\,\hat{J}(q(t),p(t),t) - \hat{J}(q(0),p(0),0)\,| \quad \leq \quad \tilde{c}_1\,t\,e^{-\tilde{c}_2/\tilde{\epsilon}}\,,$$

$\tilde{c}_1,\tilde{c}_2 > 0$ appropriate constants, and $\bar{J}(q,p,t) - \tilde{J}(q,p,t) = \mathcal{O}(\tilde{\epsilon})$. Let us write $\boldsymbol{x} = (q,p,t)^T$ and $\boldsymbol{x}_n = [\boldsymbol{\Psi}_{\Delta t}]^n(\boldsymbol{x}_0)$. Then, along numerically computed solutions $\boldsymbol{x}_n = (q_n,p_n,t_n)^T$,

$$\begin{aligned}
|\hat{J}(\boldsymbol{x}_n) - \hat{J}(\boldsymbol{x}_0)| \quad &\leq \quad \sum_{j=0}^{n-1} |\hat{J}(\boldsymbol{x}_{j+1}) - \bar{J}(\boldsymbol{x}_j)| \\
&\leq \quad \sum_{j=0}^{n-1} |\hat{J}(\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}_j)) - \hat{J}(\boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}(\boldsymbol{x}_j)) + \hat{J}(\boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}(\boldsymbol{x}_j)) - \hat{J}(\boldsymbol{x}_j)| \\
&\leq \quad n\left[\lambda\,c_3\,\Delta t\,e^{-c_4/\Delta t} + \tilde{c}_1\,\Delta t\,e^{-\tilde{c}_2/\tilde{\epsilon}}\right],
\end{aligned}$$

with $c_3,c_4 > 0$ appropriate constants and $\lambda > 0$ the Lipschitz constant of $\hat{J}$. This implies

$$|J(\boldsymbol{x}_n) - J(\boldsymbol{x}_0)| \quad \leq \quad n\left[\lambda\,c_3\,\Delta t\,e^{-c_4/\Delta t} + \tilde{c}_1\,\Delta t\,e^{-\tilde{c}_2/\tilde{\epsilon}}\right] + \mathcal{O}(\epsilon) + \mathcal{O}(\Delta t^2)$$

for the adiabatic invariant J which is to be compared to (3.16). Thus, for Δt sufficiently small, one can conclude that symplectic integrators do not only approximately conserve total energy over exponentially long periods of time but adiabatic invariants are approximately conserved over exponentially long periods of time as well. Compare the numerical experiments conducted by SHIMADA & YOSHIDA [112] and Example 2.5 below. For further results on the preservation of adiabatic invariants under a symplectic discretization, see Section 4.7.

Example 2.5. Let us consider a one-dimensional harmonic oscillator with a slowly varying frequency. The Hamiltonian is

$$H(q,p,\epsilon t) \quad = \quad \frac{1}{2}\,p^2 + \frac{1}{2}\,\omega(\epsilon t)^2\,q^2$$

where

$$\omega\,(\epsilon t) \quad = \quad 1 + \delta\,\sin\,(\epsilon t)$$

with $\delta = 0.1$ and $\epsilon \leq 0.1$ [112]. The adiabatic invariant is

$$J(q(t),p(t),t) \quad = \quad \frac{H(q(t),p(t),\epsilon t)}{\omega\,(\epsilon t)}\,.$$

We integrated the corresponding equations of motion by the symplectic implicit midpoint rule with step-sizes $\Delta t \leq 16.0$. Note that the period of the "fast" oscillations in (q,p) is $T \approx 2\pi$. For a step-size $\Delta t = 1.0$, the fast oscillations are accurately reproduced and one obtains the typical ϵ-dependence in the variation of the adiabatic invariant (Fig. 2.1). For step-sizes $\Delta t > 1$, the fast oscillations are no longer correctly resolved. However, the implicit midpoint rule is stable for arbitrary step-sizes when applied to an unperturbed harmonic oscillator and one could expect that one can also use larger step-sizes for the harmonic oscillator with slowly varying frequency. However, as our numerical results indicate, one has to be very cautious with this statement (Fig. 2.2). This has been explained by ASCHER & REICH [8] as follows:

First rescale time t to $\tau = \epsilon t$. Then the equations of motion are

$$\begin{aligned} \frac{d}{d\tau}q &= \epsilon^{-1}p \\ \frac{d}{d\tau}p &= -\epsilon^{-1}\omega(\tau)^2 q\,. \end{aligned}$$

Define

$$\alpha := \frac{\Delta\tau^2}{4\epsilon}$$

and consider the midpoint equations

$$\begin{aligned} (q_n - q_{n-1})/\Delta\tau &= \epsilon^{-1}(p_n + p_{n-1})/2 \\ (p_n - p_{n-1})/\Delta\tau &= -\epsilon^{-1}\omega(\tau_{n-1/2})^2(q_n + q_{n-1})/2 \end{aligned}$$

Let $u_n := (-1)^n q_n$, $v_n := (-1)^{n+1} p_n$. Note that the Hamiltonian H, and therefore also the adiabatic invariant J, satisfy

$$H(q_n,p_n,\tau_n) = H(u_n,v_n,\tau_n), \quad J(q_n,p_n,\tau_n) = J(u_n,v_n,\tau_n)$$

For the new variables we get

$$\begin{aligned}(u_n + u_{n-1})/\Delta\tau &= -\epsilon^{-1}(v_n - v_{n-1})/2 \\ (v_n + v_{n-1})/\Delta\tau &= \epsilon^{-1}\omega(\tau_{n-1/2})^2(u_n - u_{n-1})/2\end{aligned}$$

Multiplying both equations by $\Delta\tau/2$ and rearranging we get

$$\begin{aligned}(u_n + u_{n-1})/2 &= -\alpha(v_n - v_{n-1})/\Delta\tau \\ (v_n + v_{n-1})/2 &= \alpha\omega(\tau_{n-1/2})^2(u_n - u_{n-1})/\Delta\tau\end{aligned}$$

Now observe what approximation we get as $\Delta\tau \to 0$ for a fixed α. In fact, this is again the midpoint scheme(!) applied to the "ghost ODE"

$$\begin{aligned}-\alpha\frac{d}{d\tau}v &= u \\ \alpha\omega(\tau)^2\frac{d}{d\tau}u &= v\end{aligned}$$

or,

$$\frac{d^2}{d\tau^2}v = -(\omega(\tau)\alpha)^{-2}v \tag{3.17}$$

The features of u and v therefore depend in the limit only on α (and of course ω). Whenever $\alpha \ll 1$, the modified system (3.17) describes again a highly oscillatory system with slowly varying frequency. Its frequency is

$$\tilde{\omega}(\tau) = (\omega(\tau)\alpha)^{-1}$$

and its Hamiltonian is

$$\tilde{H}(u, v, \tau) = (\omega^2(\tau)\alpha)^{-1}v^2/2 + \alpha^{-1}u^2/2\,.$$

Thus the corresponding adiabatic invariant is

$$\tilde{J}(u, v, \tau) = \omega^{-1}(\tau)v^2/2 + \omega(\tau)u^2/2$$

and so, similarly to the original problem formulation, we obtain

$$\tilde{J}(u(\tau), v(\tau), \tau) - \tilde{J}(u(0), v(0), 0) = O(\alpha)$$

over a time-interval

$$\tilde{T} = \tilde{c}_1 e^{\tilde{c}_2/\alpha}\,,$$

$\tilde{c}_1, \tilde{c}_2 > 0$ appropriate constants, provided that α is small enough. Finally note that

$$\tilde{J}(u_n, v_n, \tau_n) = \tilde{J}(q_n, p_n, \tau_n) = J(q_n, p_n, \tau_n)\,.$$

Thus, for step-sizes large compared to ϵ, the implicit midpoint rule is equivalent to the exact solution of a harmonic oscillator with lower frequency and the adiabatic invariance condition [112]

$$\alpha \ll \frac{1}{2\,\pi\,\delta} \approx 1.6$$

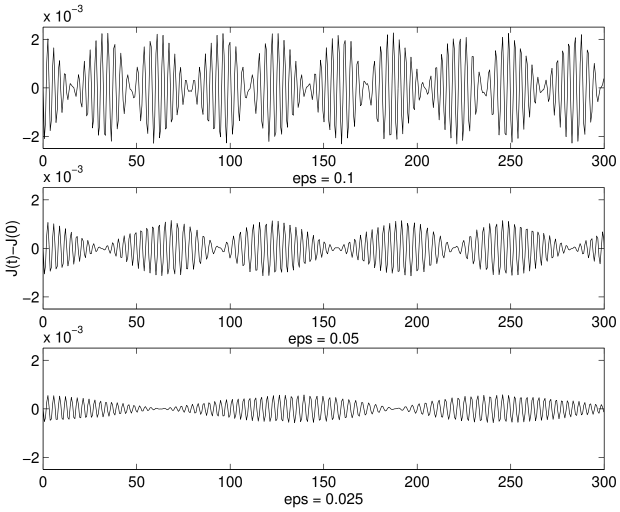


Figure 2.1: Variation in the adiabatic invariant J for different values of ϵ and (small) constant step-size $\Delta t = 1.0$.

is not necessarily satisfied anymore. Thus the quantity $J(\tau)$ can start to drift. For the results presented in Fig. 2.2, the corresponding α's are $\alpha = 0.2$, 0.8, 3.2 and the $\mathcal{O}(\alpha)$ dependence of $J(t) - J(0)$ is approximately satisfied. (Note that the calculations were carried out in slow time t and that α is then given by $\alpha = \epsilon \Delta t^2/4!$) For a more detailed numerical study see [8]. □

2.4 Symplectic Variable Step-Size Integration

According to a result by Stoffer & Nipp [114], classical variable step-size methods asymptotically reduce to a sequence of mappings

$$\boldsymbol{x}_{n+1} = \boldsymbol{\Psi}_{\Delta t(\boldsymbol{x}_n)}(\boldsymbol{x}_n)\,, \tag{4.18}$$

$$t_{n+1} = t_n + \Delta t\,(\boldsymbol{x}_n) \tag{4.19}$$

with $\Delta t(\boldsymbol{x})$ an appropriate function determined by the step-size selection criteria. Typically, we have

$$\Delta t\,(\boldsymbol{x}) = \delta\, s(\boldsymbol{x}, \delta)$$

with $\delta = \mathrm{TOL}^{1/p}$, $\mathrm{TOL} \ll 1$ a given parameter and p the order of the method $\boldsymbol{\Psi}_{\Delta t}$. Let us now take a different point of view: The variable step-size method (4.18)-(4.19) can be viewed as a constant step-size discretization with step-size δ applied to the scaled differential equation

$$\frac{d}{d\tau}\boldsymbol{x} = \rho(\boldsymbol{x})\,\boldsymbol{Z}(\boldsymbol{x})\,, \tag{4.20}$$

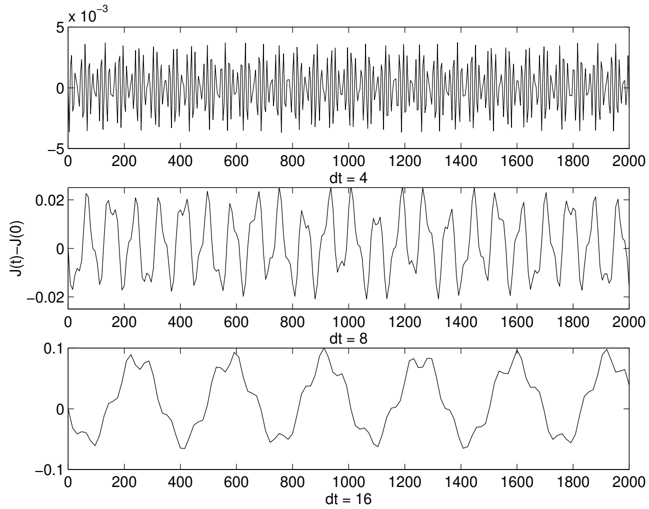


Figure 2.2: Variation in the adiabatic invariant J for constant $\epsilon = 0.05$ and different (large) step-sizes Δt.

$\rho(\boldsymbol{x}) \approx s(\boldsymbol{x}, \delta)$. As advocated by Huang & Leimkuhler [62] in the context of time-reversible integration, one could, in fact, take (4.20) as the starting point, i.e., define an appropriate scaling function $\rho(\boldsymbol{x})$ and discretize the scaled differential equation by, for example, a time-reversible method. That this can lead to highly efficient methods has been demonstrated in [62] for time-reversible Hamiltonian systems of the form

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

$\boldsymbol{q}, \boldsymbol{p} \in \mathrm{R}^n$, with Hamiltonian

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

As suggested in [62], the scaling function is defined by

$$\rho\,(\boldsymbol{q},\boldsymbol{p}) \;=\; \frac{1}{\sqrt{\boldsymbol{p}^T\boldsymbol{M}^{-2}\boldsymbol{p} + (\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}))^T\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q})}}\,. \tag{4.21}$$

Note that this choice makes a lot of sense in the context of Theorem 2.1: The constant M there is proportional to the supremum of $||\boldsymbol{Z}(\boldsymbol{x})||$, $\boldsymbol{x} \in \mathcal{B}_R\mathcal{K}$, $||\,.\,||$ the l^∞-norm in C^n. Replacing the l^∞-norm by the Euclidian norm and using (4.20) with the scaling function (4.21), which corresponds to $\rho(\boldsymbol{x}) = 1/||\boldsymbol{Z}(\boldsymbol{x})||$, yields that the constant M for the scaled differential equation (4.20) is equal to one on the compact set $\mathcal{K}$. (Of course, we need an estimate for $||\boldsymbol{Z}(\boldsymbol{x})||$ on a complex neighborhood $\mathcal{B}_R\mathcal{K}$ of $\mathcal{K}$. Thus scaling of the vector field does not imply that the constant M in Theorem 2.1 is equal to one.)

It has not been shown yet that reversible (non-symplectic) methods show the same excellent long-term behavior as symplectic methods do; namely: conservation of energy over exponentially long periods of time. They also do not preserve volume. Thus it seems reasonable to look for a symplectic discretization of the scaled Hamiltonian equations of motion: Following ZARE & SZEBEHELY [125], we introduce the modified Hamiltonian function

$$E(\boldsymbol{q},\boldsymbol{p},t,e) \quad := \quad \rho(\boldsymbol{q},\boldsymbol{p})\,(H(\boldsymbol{q},\boldsymbol{p})-e)$$

with corresponding equations of motion

$$\begin{aligned}
\frac{d}{d\tau}\boldsymbol{q} &= \rho(\boldsymbol{q},\boldsymbol{p})\,\boldsymbol{M}^{-1}\boldsymbol{p} + (H(\boldsymbol{q},\boldsymbol{p})-e)\,\boldsymbol{\nabla}_{\boldsymbol{p}}\,\rho(\boldsymbol{q},\boldsymbol{p})\,,\\
\frac{d}{d\tau}\boldsymbol{p} &= -\rho(\boldsymbol{q},\boldsymbol{p})\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,V(\boldsymbol{q}) - (H(\boldsymbol{q},\boldsymbol{p})-e)\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,\rho(\boldsymbol{q},\boldsymbol{p})\,,\\
\frac{d}{d\tau}t &= \rho(\boldsymbol{q},\boldsymbol{p})\,,\\
\frac{d}{d\tau}e &= 0
\end{aligned}$$

in extended phase space $\mathrm{R}^{2n}\times\mathrm{R}^2$. In particular, let us consider the case $e = H(\boldsymbol{q}(0),\boldsymbol{p}(0))$ and ρ only a function of $\boldsymbol{q}$. Then

$$\begin{aligned}
\frac{d}{d\tau}\boldsymbol{q} &= \rho(\boldsymbol{q})\,\boldsymbol{M}^{-1}\boldsymbol{p}\,,\\
\frac{d}{d\tau}\boldsymbol{p} &= -\rho(\boldsymbol{q})\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,V(\boldsymbol{q}) - (H(\boldsymbol{q},\boldsymbol{p})-e)\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,\rho(\boldsymbol{q}) \;=\; -\rho(\boldsymbol{q})\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,V(\boldsymbol{q})\,,\\
\frac{d}{d\tau}t &= \rho\,(\boldsymbol{q})\,,\\
\frac{d}{d\tau}e &= 0
\end{aligned}$$

which is just our scaled Hamiltonian vector field and can be discretized by the symplectic Euler method, i.e.

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

Note that, for symplecticity, one has to keep the term $(H(\boldsymbol{q}_n,\boldsymbol{p}_{n+1})-e)\,\boldsymbol{\nabla}_{\boldsymbol{q}}\,\rho(\boldsymbol{q}_n)$. Let us now define our scaling function ρ. For simplicity, we set $\boldsymbol{M}=\boldsymbol{I}$ which can always be achieved by an appropriate scaling of the positions $\boldsymbol{q}$ and the momenta $\boldsymbol{p}$. According to (4.21), we obtain

$$\begin{aligned}
\rho\,(\boldsymbol{q}) &= \frac{1}{\sqrt{\boldsymbol{p}^T\boldsymbol{p} + (\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}))^T\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q})}}\\
&= \frac{1}{\sqrt{2(e-V(\boldsymbol{q})) + (\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}))^T\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q})}}\,.
\end{aligned} \tag{4.22}$$

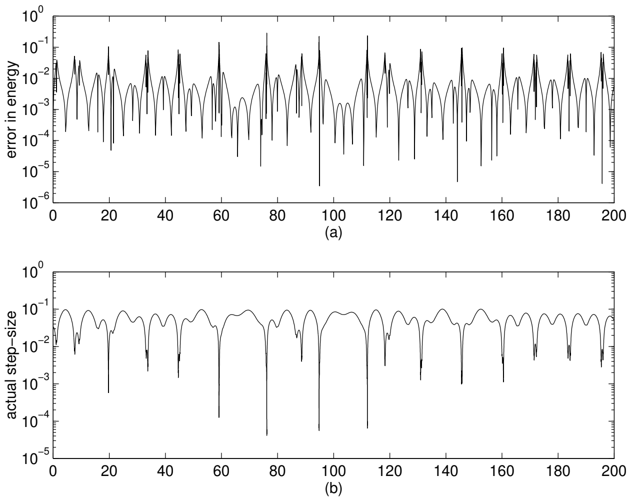


Figure 2.3: Error in energy (a) and actual step-size Δt (b) as a function of time.

The method is explicit in the variable $\boldsymbol{q}$. Unfortunately this implies that the method is only first order in Δt. However, the method is symplectic and, therefore, the Hamiltonian $E = (H(\boldsymbol{q}, \boldsymbol{p}) - e)\rho(\boldsymbol{q})$ is conserved to $\mathcal{O}(\Delta t)$ over exponentially long periods of time. This implies $H(\boldsymbol{q}_n, \boldsymbol{p}_n) - e = \mathcal{O}(\Delta\tau)$ over exponentially long periods of time. Thus the proposed method seem suitable for long term, relatively low precision, variable step-size simulations as they occur, for example, in systems with an ergodic behavior. A second-order symplectic discretization could be obtained by using the second-order Lobatto IIIa-b partitioned Runge-Kutta formula [115], i.e.

$$\begin{aligned}
\boldsymbol{p}_{n+1/2} &= \boldsymbol{p}_n - \Delta\tau \left[\rho(\boldsymbol{q}_n)\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}_n) - [H(\boldsymbol{q}_n, \boldsymbol{p}_{n+1/2}) - e]\boldsymbol{\nabla}_{\boldsymbol{q}}\rho(\boldsymbol{q}_n)\right], \\
\boldsymbol{q}_{n+1} &= \boldsymbol{q}_n + \frac{\Delta t}{2}[\rho(\boldsymbol{q}_{n+1}) + \rho(\boldsymbol{q}_n)]\,\boldsymbol{p}_{n+1/2}, \\
\boldsymbol{p}_{n+1} &= \boldsymbol{p}_{n+1/2} - \Delta\tau \left[\rho(\boldsymbol{q}_{n+1})\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}_{n+1}) - [H(\boldsymbol{q}_{n+1}, \boldsymbol{p}_{n+1/2}) - e]\boldsymbol{\nabla}_{\boldsymbol{q}}\rho(\boldsymbol{q}_{n+1})\right], \\
t_{n+1} &= t_n + \frac{\Delta\tau}{2}[\rho(\boldsymbol{q}_n) + \rho(\boldsymbol{q}_{n+1})].
\end{aligned}$$

The resulting scheme is implicit in $\rho(\boldsymbol{q})$. However, in many cases the scaling function $\rho(\boldsymbol{q})$ can be greatly simplified and its evaluation is cheap compared to the evaluation of the force field $\boldsymbol{F}(\boldsymbol{q}) = -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q})$. Independently of us, the same approach to symplectic variable step-size integration has been formulated by HAIRER [54].

Example 2.6. As a numerical example, we look at the following modified Kepler problem:

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{q} &= \boldsymbol{p}, \\
\frac{d}{dt}\boldsymbol{p} &= -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}),
\end{aligned}$$

$\boldsymbol{q}, \boldsymbol{p} \in \mathrm{R}^2$, and

$$V(q_x, q_y) = -\frac{1}{\sqrt{(q_x/10)^2 + (q_y)^2}}\,.$$

The problem is non-integrable and, in fact, the dynamics is chaotic, i.e., can be reduced to the Bernoulli shift [52, 69]. We chose initial values $\boldsymbol{q} = (0, 1)$ and $\boldsymbol{p} = (1, 0)$. The equations of motion are integrated using the "variable" step-size symplectic Euler method with scaling function (4.22) and $\Delta\tau = 0.05$. The error in energy $\Delta H = |\,H(\boldsymbol{q}, \boldsymbol{p}) - e\,|$ and the variation in the actual step-size $\Delta t = \rho(\boldsymbol{q})\,\Delta\tau$ can be found in Fig. 2.3.

2.5 Another Application: Ergodic Hamiltonian Systems

Let us consider a (real analytic) Hamiltonian system

$$\frac{d}{dt}\boldsymbol{q} = \boldsymbol{M}^{-1}\boldsymbol{p}\,, \tag{5.23}$$

$$\frac{d}{dt}\boldsymbol{p} = -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q})\,, \tag{5.24}$$

$\boldsymbol{q}, \boldsymbol{p} \in \mathrm{R}^n$, 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 (5.23)-(5.24), 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

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

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 (5.23)-(5.24) is ergodic[4] (or even mixing [81],[122]). Thus

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

[4]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.

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 (5.23)-(5.24) 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 $H = 0$,

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

is a compact subset of R^{2n}. We also assume that there is a constant $\gamma_1 > 0$ such that $||\boldsymbol{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x})||_2 > \gamma_1$ 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 diffeomorph 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 (5.23)-(5.24) is ergodic/mixing.

In a first step we construct a Poincaré return map [51]. 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

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

We assume that the Poincaré return times $t_p(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{V}$, are bounded by some moderate 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 $\mathbf{\Pi}_0$ which is defined as the restriction of $\mathbf{\Pi}$ to

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

Thus we have reduced the study of the dynamical properties of the Hamiltonian system (5.23)-(5.24) on the energy shell $\mathcal{M}_0$ to the study of the properties of the Poincaré map $\mathbf{\Pi}_0$. If $\mathbf{\Pi}_0$ is an ergodic (mixing) map, then the Hamiltonian system is ergodic (mixing) on $\mathcal{M}_0$. Note that $\mathbf{\Pi}_0$ is volume preserving, i.e. $\det \boldsymbol{\partial_x} \mathbf{\Pi}_0(\boldsymbol{x}) = 1$.

From now on we assume that $\mathbf{\Pi}_0$ is a uniformly hyperbolic map, i.e., for each $\boldsymbol{x} \in \mathcal{D}$, the linearization $\boldsymbol{\partial_x} \mathbf{\Pi}_0(\boldsymbol{x})$ at $\boldsymbol{x}$ possesses strictly expanding and contracting directions only [51],[121]. The "stochastic" behavior of such a (deterministic) map has been investigated in [79],[121]. Here we only point out the four main results:

- There is a unique invariant density μ_0 on $\mathcal{D}$ that is invariant under $\mathbf{\Pi}_0$. Furthermore, μ_0 is given by the Lebesgue measure on $\mathcal{D}$.
- The autocorrelation function $\langle A \circ [\mathbf{\Pi}_0]^n, A\rangle$ decays exponentially fast, i.e.

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

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

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

  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: Given any $c > 0$ there is a $h(c) > 0$ such that

$$\mu_0(\{\boldsymbol{x}_0 \in \mathcal{D} : |\langle A\rangle_N - \langle A, \mu_0\rangle| > c\}) \leq e^{-Nh(c)} \tag{5.25}$$

  for all large $N \geq 1$.

These results can be proven (see, for example, [121]) 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

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

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

Definition. We call a Hamiltonian system (5.23)-(5.24) with the above introduced properties *Poincaré hyperbolic*[5]. 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 $||\boldsymbol{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x})||_2 > \gamma_1$ for all $\boldsymbol{x} \in \mathcal{U}$, (iii) that a global Poincaré map $\boldsymbol{\Pi}$ 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 are bounded by some moderate constant $K > 0$, and (v) that there is a constant $\gamma_2 > 0$ such that $|\{\psi, H\}(\boldsymbol{x})| > \gamma_2$ on $\mathcal{V}$. □

Lemma. The property of being Poincaré hyperbolic is stable under sufficiently small perturbations[6] of the Hamiltonian H. □

Proof. The assumption $||\boldsymbol{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x})||_2 > \gamma_1$ 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 [5]. □

Let us discretize (5.23)-(5.24) by a symplectic (real analytic) integrator $\boldsymbol{\Psi}_{\Delta t}$ of order $p \geq 1$.

Assumptions. 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 $\tilde{H}$. Let the step-size Δt be sufficiently small such that the perturbed Hamiltonian system is also Poincaré hyperbolic. □

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 $\tilde{H}$ in the definition). Furthermore,

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

$s \in (-\Delta s, +\Delta s)$. 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.

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 $\boldsymbol{\Phi}_{t,\tilde{H}}$ of the modified problem as an interpolation for $t \in [0, \Delta t)$. The map is then

[5] For example, the modified Kepler problem from Example 2.6 is Poincaré hyperbolic.

[6] Small perturbation means that $|H(\boldsymbol{x}) - \tilde{H}(\boldsymbol{x})| + ||\boldsymbol{\nabla}_{\boldsymbol{x}} H(\boldsymbol{x}) - \boldsymbol{\nabla}_{\boldsymbol{x}} \tilde{H}(\boldsymbol{x})||_2$ is uniformly small on $\mathcal{U}$, $\tilde{H}$ the perturbed Hamiltonian.

extended to $t \geq \Delta t$ in the obvious way as the composition of k steps with $\boldsymbol{\Psi}_{\Delta t}$ and one step with $\boldsymbol{\Phi}_{dt,\tilde{H}}$ where $t = k\Delta t + dt$, $\Delta t > dt \geq 0$. Thus, in correspondence with the definition of the global Poincaré map

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

we define

$$\widehat{\boldsymbol{\Pi}}(\boldsymbol{x}) \quad := \quad \boldsymbol{\Psi}_{\tilde{t}_p(\boldsymbol{x})}(\boldsymbol{x}) \,.$$

Here the Poincaré return times $\tilde{t}_p(\boldsymbol{x})$ are the same as in the definition of $\widetilde{\boldsymbol{\Pi}}$. We assume that

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

$\tilde{K} > 0$ some moderate constant.

It follows from backward error analysis that there are constants $c_1, c_2 > 0$ such that

$$||\, \widetilde{\boldsymbol{\Pi}}(\boldsymbol{x}) - \widehat{\boldsymbol{\Pi}}(\boldsymbol{x})\, || \quad \leq \quad c_1\, e^{-p}\, e^{-c_2/\Delta t}$$

for Δt sufficiently small. More importantly, let $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$ be a 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 $\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0 \in \widetilde{\mathcal{D}}$. The sequence $\{\boldsymbol{x}_i\}$ generates two sequences $\{E_i\}$ and $\{s_i\}$ which are defined by $E_i = \tilde{H}(\boldsymbol{x}_i)$ and $s_i = \psi(\boldsymbol{x}_i)$. For the sequence $\{\tilde{\boldsymbol{x}}_i\}$ we obviously have $E_i = 0$ and $s_i = 0$. The "drift" in the values of E_i, s_i per step is exponentially small and sums up linearly with the number of steps. The energy conserving property of a symplectic map has already been discussed in Example 2.4. The same property follows for the sequence of values $\{s_i\}$ from

$$\begin{aligned} |\psi(\boldsymbol{x}_N) - \psi(\boldsymbol{x}_0)| \quad &\leq \quad \sum_{i=1}^{N} |\psi(\boldsymbol{x}_i) - \psi(\boldsymbol{x}_{i-1})| \\ &\leq \quad \sum_{i=1}^{N} |\psi(\widetilde{\boldsymbol{\Pi}}(\boldsymbol{x}_{i-1})) - \psi(\widehat{\boldsymbol{\Pi}}(\boldsymbol{x}_{i-1}))| \\ &\leq \quad N\,\lambda\, c_1\, e^{-p}\, e^{-c_2/\Delta t} \,, \end{aligned}$$

$\lambda > 0$ 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. 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

$$N \quad \leq \quad c_3\, e^{c_4/\Delta t} \,, \tag{5.26}$$

$c_3, c_4 > 0$ appropriate constants.

Now the *Shadowing Lemma* [110] is applied to the sequence $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$.

Proposition 2.1. 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 if the number of iterates N satisfies (5.26). □

Proof. We first project the sequence $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$ down onto $\mathcal{D}$. The projected sequence and the sequence $\{\boldsymbol{x}_i\}$ are $\mathcal{O}(\Delta t^p)$ close to each other provided N satisfies (5.26). 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 [110] for uniformly hyperbolic maps can be applied to the Poincaré map $\boldsymbol{\Pi}_0 : \mathcal{D} \to \mathcal{D}$ and the projected sequence 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 an observable A up to a certain accuracy $c > 0$. The large deviation theorem (5.25) 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 (5.23)-(5.24), then we know from Proposition 2.1 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

$$\langle A\rangle_N - \langle A\rangle_N^e \quad = \quad \mathcal{O}(\Delta t^p) \tag{5.27}$$

for all N satisfying a bound of type (1.9). Thus we obtain the following:

Proposition 2.2. Let (5.23)-(5.24) 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 $\{\boldsymbol{x}_n\}_{n=1,\dots,N}$,

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

satisfies (5.27) 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 (5.26). Furthermore, assume we want to compute the ensemble average of A within a given accuracy[7] $c > 0$. 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 (5.26) of steps, the

[7] We assume, for simplicity, that the constant c is larger than the difference between the time averages (5.27) which is always true for sufficiently small step-sizes Δt.

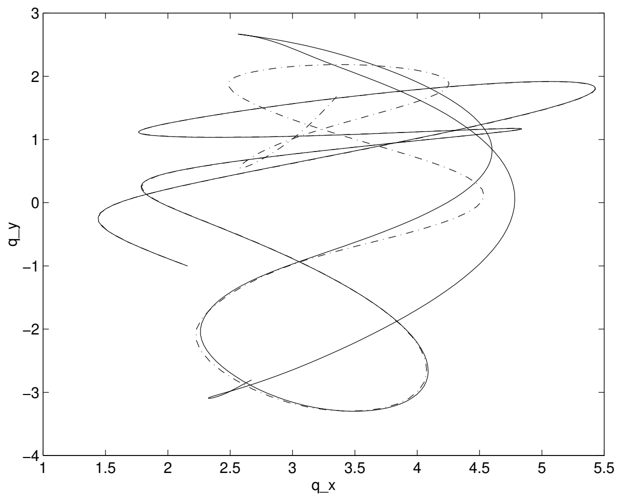


Figure 2.4: Numerical orbits for one particle with identical initial data over a time period $T \leq 20$ and different step-sizes $\Delta t = 0.1$ and $\Delta t = 0.025$.

probability can be made double exponentially close to one in (5.25) as $\Delta t \to 0$. □

Example 2.7 Let us consider a planar "billiard" system of three particles with unit mass and interaction potential $U(r) = 1/r$, $r > 0$ the mutual distance of two particles.[8] To keep the particles in a finite area, we also add a potential energy term

$$U(q_x, q_y) \quad = \quad e^{(q_x/4)^2} + e^{(q_y/4)^2} - 2e$$

for each particle. The equations of motion are discretized by the Verlet method [120]. As observable $\mathcal{A}$, we chose the mean distance between the particles. In Fig. 2.4, we have plotted the computed orbits of one particle for identical initial data over a time period $T \leq 20$ and step-sizes $\Delta t = 0.1$, $\Delta t = 0.025$ respectively. The fast divergence of the numerical orbits is clearly seen [2]. However, as shown in Fig. 2.5, the error in the energy remains bounded and the typical $\mathcal{O}(\Delta t^2)$ is observed. Note the spikes in the energy error. These are due to close interactions of particles and could be resolved with a variable step-size integrator. See the previous section. Finally, in Fig. 2.6, the time evolution of the time-average $\mathcal{A}_T$ is shown for step-sizes $\Delta t = 0.1$ and $\Delta t = 0.025$. The average $\mathcal{A}_T$ still fluctuates for $T \approx 3.0e + 05$ with a standard deviation σ of approximately $\sigma(\mathcal{A}_T) = 0.004$ (see Fig. 2.7) [2],[30]. However, it is obvious that the expectation value of $\mathcal{A}$ is well approximated. □

[8]There is no proof that we know off that this system is Poincaré hyperbolic.

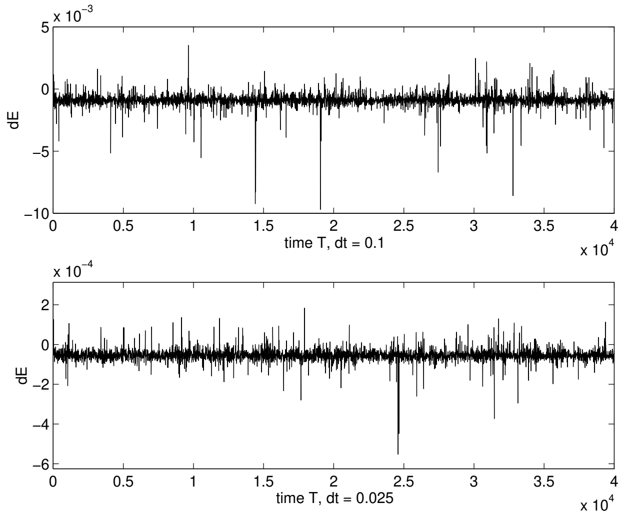


Figure 2.5: Error in energy vs. time T for two different step-sizes.

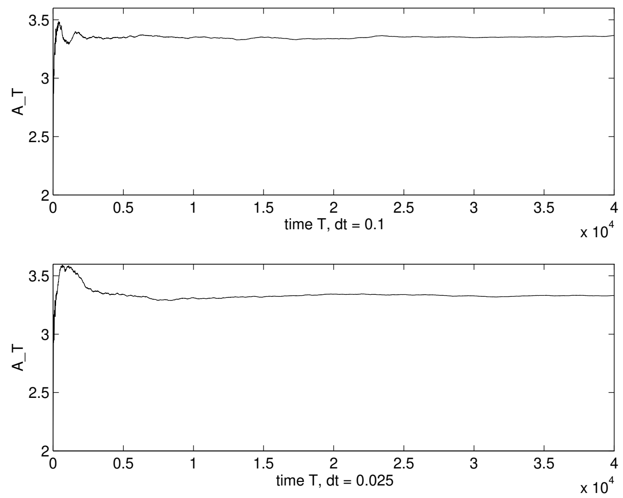
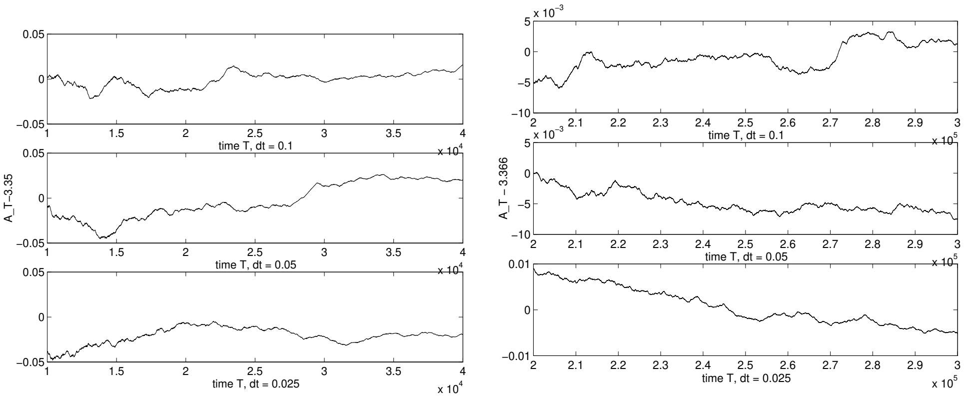


Figure 2.6: Time evolution of the average $\mathcal{A}_T$ vs. time T for two different step-sizes.



(a) $T \leq 4.0e+04$

(b) $T \leq 3.0e+05$

Figure 2.7: Time evolution of $\mathcal{A}_T$ vs. time T for different time intervals and different step-sizes.

3

Normal Form Theory

In this chapter, we discuss normal form theory and exponentially small truncation errors for differential equations

$$\begin{aligned} \frac{d}{dt}\boldsymbol{x} &= \boldsymbol{Y}(\boldsymbol{x};\epsilon)\,, \\ &= \boldsymbol{A}(\boldsymbol{x}) + \epsilon\,\boldsymbol{B}(\boldsymbol{x};\epsilon)\,, \end{aligned} \tag{0.1}$$

where ϵ is a small parameter, $\boldsymbol{x} \in \mathrm{R}^n$. The aim is to find a coordinate transformation (diffeomorphism)

$$\boldsymbol{x} = \boldsymbol{\Psi}(\bar{\boldsymbol{x}};\epsilon) \tag{0.2}$$

to new coordinates $\bar{\boldsymbol{x}} \in \mathrm{R}^n$ such that the transformed vector field

$$\bar{\boldsymbol{Y}}(\bar{\boldsymbol{x}};\epsilon) := \left[\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\bar{\boldsymbol{x}}}\boldsymbol{\Psi}(\bar{\boldsymbol{x}};\epsilon)\right]^{-1} \cdot \boldsymbol{Y}(\epsilon) \circ \boldsymbol{\Psi}(\bar{\boldsymbol{x}};\epsilon)\,, \tag{0.3}$$

is of "simpler" form than the original problem formulation (0.1). With

$$\begin{aligned} \frac{d}{dt}\bar{\boldsymbol{x}} &= \bar{\boldsymbol{Y}}(\bar{\boldsymbol{x}},\epsilon)\,, \\ &= \boldsymbol{A}(\bar{\boldsymbol{x}}) + \epsilon\,\bar{\boldsymbol{B}}(\bar{\boldsymbol{x}};\epsilon)\,, \end{aligned}$$

"simpler" means that the two vector fields $\boldsymbol{A}$ and $\bar{\boldsymbol{B}}$ commute up to terms of order $p \geq 1$, i.e.

$$[\boldsymbol{A},\bar{\boldsymbol{B}}] = \mathcal{O}(\epsilon^p)\,,$$

while, for the original problem (0.1), we only have

$$[\boldsymbol{A},\boldsymbol{B}] = \mathcal{O}(1)\,,$$

in general. Here $[\boldsymbol{X},\boldsymbol{Y}]$ denotes the Lie bracket (commutator) of two vector fields $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e.

$$[\boldsymbol{X},\boldsymbol{Y}](\boldsymbol{x}) := \frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{X}(\boldsymbol{x}) \cdot \boldsymbol{Y}(\boldsymbol{x}) - \frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{Y}(\boldsymbol{x}) \cdot \boldsymbol{X}(\boldsymbol{x})\,.$$

(In many references, the Lie bracket is defined as the negative of the above formula on the right hand side [90],[6].) The ultimate goal is to make $[\boldsymbol{A},\bar{\boldsymbol{B}}]$ as small as possible.

Normal theory (or the principle of averaging) has a long history. In the context of planetary motion it goes back to LAGRANGE and LAPLACE. It was subsequently rediscovered by VAN DER POL and used by him to solve problems in the theory of nonlinear oscillations. The wide-scale application of the averaging principle was stimulated by the investigations of MANDEL'SHTAM, PAPALESKI, N.M. KRYLOV, BOGOLYUBOV, & MITROPOL'SKII [72],[22]. There are various approaches to derive the asymptotic normal form expansion and its corresponding coordinate transformation [5],[7],[80]. For example: (i) The coordinate change is sought as a composition of successive coordinate changes. This seems the most popular method to date. (ii) The coordinate transformation (0.2) is defined as the time-one-flow map of an appropriate non-autonomous differential equation

$$\frac{d}{d\epsilon}\boldsymbol{x} \;=\; \epsilon \boldsymbol{W}(\boldsymbol{x};\epsilon)\,, \qquad \boldsymbol{x}(\epsilon = 0) \;=\; \bar{\boldsymbol{x}}\,.$$

This approach is due to KAMEL [67]. (iii) Same as (ii) except that the time-one-flow map of an autonomous vector field

$$\frac{d}{d\tau}\boldsymbol{x} \;=\; \epsilon \boldsymbol{W}(\boldsymbol{x};\epsilon)\,, \qquad \boldsymbol{x}(\tau = 0) \;=\; \bar{\boldsymbol{x}}$$

is used. This approach was suggested by HORI [61]. (iv) Formal expansion of the coordinate transformation (0.2) in terms of ϵ (Linstedt series). All three methods lead to an asymptotic expansion in the parameter ϵ that, in general, diverges. Thus one is interested in the truncation error introduced by only considering finitely many terms in the asymptotic expansion. Recently, exponentially small estimates for the truncation error have become important. Such estimates were first popularized through the work of NEKHOROSHEV [89] on perturbed integrable Hamiltonian systems. A similar estimate was subsequently derived by NEISHTADT [88] for systems with rapidly rotating phase. Since then exponentially small estimates have been derived for various other systems. In particular, we like to mention the work of BENETTIN, GIORGILLI & GALVANI [13],[14],[15] on conservative mechanical systems with highly oscillatory internal degrees of freedom and the work by FASSÒ [36] on Lie series methods of type (i) for general vector fields (0.1).

Our approach to the normal form recursion is of type (iii). Using Theorem 1.1 from Chapter 1, we know that any diffeomorphism ϵ-close to the identity can be approximated by the time-one-flow map of an autonomous vector field up to terms exponentially small in ϵ. Thus we look for an appropriate vector field $\boldsymbol{W}(\epsilon) : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$ and define the sought change of coordinates by means of

$$\boldsymbol{x} \;=\; \boldsymbol{\Phi}_{1,\boldsymbol{W}(\epsilon)}(\bar{\boldsymbol{x}})\,, \tag{0.4}$$

$\mathcal{U}$ an appropriate subset of R^n. In our opinion, the main advantage of (0.4) lies in the simplicity of the recursive definition of the vector field $\boldsymbol{W}(\epsilon)$ which also allows one to derive simple estimates for the normal form truncation error. More specifically: In Section 3.1, we will give the recursive definition of the transforming vector field

$\boldsymbol{W}(\epsilon)$ in terms of an asymptotic series expansion. We show in Section 3.2 that, under fairly general assumptions, the truncation error in this series expansion can be made exponentially small. In this form, the proof seems to be novel and more elementary than existing proofs for special systems of type (0.1). Note that our approach is close to the method discussed by FASSÒ in [36]. However, while FASSÒ considers normal forms defined by a sequence of time-one-flow maps (which is, therefore, an approach of type (i)), we define the normal form by a single flow map – see Remark 3.1 in Section 3.1 for more details.

Applications are discussed in Section 3.3. In particular, we consider systems of time-varying oscillators and discuss the existence of adiabatic invariants. The results are of interest for discussing the solution behavior of the (truncated finite-dimensional) time-varying Schrödinger equation [74]. We also show that NEISHTADT's result on systems with rapidly rotating phase follows from our proof of Section 3.2 as a special case. We then apply our results to conservative mechanical systems with a single fast degree of motion as discussed, for example, by RUBIN & UNGAR in [105] (see also TAKENS [116] and BORNEMANN & SCHÜTTE [24]). Our approach to this problem is novel in the sense that we give a canonical coordinate transformation that makes the problem accessible to normal form theory. This allows us to get sharper results than those previously derived by means of different techniques such as homogenization (see BORNEMANN & SCHÜTTE [24]). In particular, we investigate the existence of an adiabatic invariant and show that the adiabatic invariant is preserved over exponentially long periods of time. Finally, in Section 3.7, we discuss implications for the numerical treatment of systems of type (0.1) and show that symplectic methods preserve adiabatic invariants associated with Hamiltonian systems of type (0.1).

The following example will serve as a guide through the rather abstract formulations used in the subsequent sections.

Example 3.1. Let us consider the system of differential equations

$$\frac{d}{dt}\boldsymbol{\phi} = \boldsymbol{\omega} + \epsilon\, \boldsymbol{f}(\boldsymbol{\phi}, \boldsymbol{I}) \tag{0.5}$$

$$\frac{d}{dt}\boldsymbol{I} = \epsilon\, \boldsymbol{g}(\boldsymbol{\phi}, \boldsymbol{I}), \tag{0.6}$$

$\boldsymbol{\phi} \in \mathrm{T}^d$, $\boldsymbol{I} \in \mathrm{R}^d$, and the functions $\boldsymbol{f}$, $\boldsymbol{g}$ are 2π-periodic in the argument $\boldsymbol{\phi}$. Here T^d denotes the d-dimensional standard torus, i.e. $\mathrm{T}^d = \mathrm{R}^d/2\pi\mathrm{Z}^d$. The system (0.5)-(0.6) describes the motion of d harmonic oscillators coupled to a small non-linear perturbation. With $\boldsymbol{x} = (\boldsymbol{\phi}^T, \boldsymbol{I}^T)^T$, this corresponds to

$$\boldsymbol{A}(\boldsymbol{x}) := \begin{bmatrix} \boldsymbol{\omega} \\ \mathbf{0} \end{bmatrix}$$

and

$$\boldsymbol{B}(\boldsymbol{x}) := \begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}) \\ \boldsymbol{g}(\boldsymbol{x}) \end{bmatrix}.$$

We would like to find a coordinate transformation such that, in the new coordinates $(\bar{\boldsymbol{\phi}}, \bar{\boldsymbol{I}})$, the equations of motion are

$$\begin{aligned} \frac{d}{dt}\bar{\boldsymbol{\phi}} &= \boldsymbol{\omega} + \epsilon\, \bar{\boldsymbol{f}}(\bar{\boldsymbol{I}};\epsilon) + \mathcal{O}(\epsilon^{p+1})\,, \\ \frac{d}{dt}\bar{\boldsymbol{I}} &= \epsilon\, \bar{\boldsymbol{g}}(\bar{\boldsymbol{I}};\epsilon) + \mathcal{O}(\epsilon^{p+1})\,. \end{aligned}$$

Thus

$$\bar{\boldsymbol{B}}(\bar{\boldsymbol{x}}) := \begin{bmatrix} \bar{\boldsymbol{f}}(\bar{\boldsymbol{I}};\epsilon) \\ \bar{\boldsymbol{g}}(\bar{\boldsymbol{I}};\epsilon) \end{bmatrix} + \mathcal{O}(\epsilon^{p})$$

and $[\boldsymbol{A}, \bar{\boldsymbol{B}}] = \mathcal{O}(\epsilon^p)$. In other words, the motion decouples into fast oscillations in $\bar{\boldsymbol{\phi}}$ and a slow motion in $\bar{\boldsymbol{I}}$ governed by a differential equation solely in the variable $\bar{\boldsymbol{I}}$. □

3.1 The Normal Form Recursion

We start with a definition.

Definition. Let $\boldsymbol{Y}$ be a vector field on R^n and let $\boldsymbol{\Psi}$ be a diffeomorphism on an appropriate subset $\mathcal{U}$ of R^n. Then we define the *pull-back* $\boldsymbol{\Psi}^*\boldsymbol{Y}$ of $\boldsymbol{Y}$ under $\boldsymbol{\Psi}$ by

$$\boldsymbol{\Psi}^*\boldsymbol{Y}\,(\bar{\boldsymbol{x}}) := \left[\frac{\partial}{\partial \bar{\boldsymbol{x}}}\boldsymbol{\Psi}(\bar{\boldsymbol{x}})\right]^{-1} \cdot \boldsymbol{Y} \circ \boldsymbol{\Psi}\,(\bar{\boldsymbol{x}})\,.$$

In other words,

$$\bar{\boldsymbol{Y}}(\bar{\boldsymbol{x}}) := \boldsymbol{\Psi}^*\boldsymbol{Y}(\bar{\boldsymbol{x}})$$

is the vector field obtained by applying the coordinate transformation

$$\boldsymbol{x} = \boldsymbol{\Psi}(\bar{\boldsymbol{x}})$$

to the vector field $\boldsymbol{Y}(\boldsymbol{x})$. Note that the pull-back $\boldsymbol{\Psi}^*\boldsymbol{Y}$ can also be defined by [36]

$$\boldsymbol{\Psi}^*\boldsymbol{Y} = \left[\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{\Psi}^{-1} \cdot \boldsymbol{Y}\right] \circ \boldsymbol{\Psi}\,. \tag{1.7}$$

To be more precise:

$$\boldsymbol{\Psi}^*\boldsymbol{Y}\,(\bar{\boldsymbol{x}}) = \frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{\Psi}^{-1}(\boldsymbol{x}) \cdot \boldsymbol{Y}(\boldsymbol{x})\,, \qquad \text{with } \boldsymbol{x} = \boldsymbol{\Psi}(\bar{\boldsymbol{x}})\,.$$

Throughout this chapter, we will use this definition of the pull-back. □

Using the above notation, our normal form recursion can now be stated as follows: We define a sequence of *transforming vector fields* $\boldsymbol{W}_i(\epsilon)$, $i \geq 1$, by means of the following recursion:

$$\boldsymbol{W}_{i+1}(\epsilon) := \boldsymbol{W}_i(\epsilon) + \epsilon^{i+1}\, \boldsymbol{\Delta W}_{i+1}\,, \qquad \boldsymbol{W}_0(\epsilon) = \boldsymbol{0}\,.$$

Let us assume that we have already obtained $\boldsymbol{W}_i(\epsilon)$. The vector field $\boldsymbol{\Delta W}_{i+1}$ will be defined below. The *transformed vector field* $\boldsymbol{Y}_i(\epsilon)$ is then given by the pull-back of $\boldsymbol{Y}(\epsilon)$ under the time-one-flow map of $\boldsymbol{W}_i(\epsilon)$, i.e.

$$\begin{aligned}\boldsymbol{Y}_i(\epsilon) &:= \boldsymbol{\Phi}^*_{1,\boldsymbol{W}_i(\epsilon)}\boldsymbol{Y}(\epsilon)\,, \\ &= \boldsymbol{A} + \sum_{j=1}^{i} \epsilon^j\, \boldsymbol{\Delta X}^r_j + \mathcal{O}(\epsilon^{i+1})\,.\end{aligned}$$

The vector fields $\boldsymbol{\Delta X}^r_j$, $j = 1, \ldots, i$, are defined in the following way: First we introduce the vector fields $\boldsymbol{\Delta X}_j$ by

$$\boldsymbol{\Delta X}_j \;:=\; \frac{1}{j!}\left[\frac{\partial^j}{\partial\epsilon^j}\boldsymbol{Y}_{j-1}\right]_{\epsilon=0}\,,$$

Then we split the vector field $\boldsymbol{\Delta X}_j$ into two parts $\boldsymbol{\Delta X}^r_j$ and $\boldsymbol{\Delta X}^n_j$, i.e.

$$\boldsymbol{\Delta X}_j \;=:\; \boldsymbol{\Delta X}^n_j + \boldsymbol{\Delta X}^r_j\,,$$

such that the linear partial differential equation

$$[\boldsymbol{A}, \boldsymbol{\Delta W}_j] + \boldsymbol{\Delta X}^n_j \;=\; 0 \tag{1.8}$$

is solvable for $\boldsymbol{\Delta W}_j$. The vector field $\boldsymbol{\Delta X}^r_j$ is called the *resonant part* of $\boldsymbol{\Delta X}_j$ while $\boldsymbol{\Delta X}^n_j$ is called the *non-resonant part* of $\boldsymbol{\Delta X}_j$ with respect to the vector field $\boldsymbol{A}$. The partial differential equation (1.8) is called the *homological equation*. Upon taking $j = i + 1$, we obtain a defining equation for $\boldsymbol{\Delta W}_{i+1}$ which closes our recursion.

<u>Normal Form Recursion</u>

Initial data:

$$\boldsymbol{W}_0(\epsilon) \;:=\; \boldsymbol{0} \tag{1.9}$$

For $i \geq 0$:

$$\boldsymbol{Y}_i(\epsilon) \;:=\; \boldsymbol{\Phi}^*_{1,\boldsymbol{W}_i(\epsilon)}\boldsymbol{Y}(\epsilon)\,, \tag{1.10}$$

$$\boldsymbol{\Delta X}_{i+1} \;:=\; \frac{1}{(i+1)!}\left[\frac{\partial^{i+1}}{\partial\epsilon^{i+1}}\boldsymbol{Y}_i\right]_{\epsilon=0}\,, \tag{1.11}$$

$$\boldsymbol{\Delta X}_{i+1} \;=:\; \boldsymbol{\Delta X}^r_{i+1} + \boldsymbol{\Delta X}^n_{i+1}\,, \tag{1.12}$$

$$0 \;=\; [\boldsymbol{A}, \boldsymbol{\Delta W}_{i+1}] + \boldsymbol{\Delta X}^n_{i+1}\,, \tag{1.13}$$

$$\boldsymbol{W}_{i+1}(\epsilon) \;:=\; \boldsymbol{W}_i(\epsilon) + \epsilon^{i+1}\boldsymbol{\Delta W}_{i+1}\,. \tag{1.14}$$

The splitting (1.12) of the vector field $\boldsymbol{\Delta X}_{i+1}$ is determined by two requirements:

(i) The homological equation (1.13) has to be solvable for $\boldsymbol{\Delta W}_{i+1}$.

(ii) The vector field $\boldsymbol{\Delta X}^r_{i+1}$ should commute with the vector field $\boldsymbol{A}$, i.e.

$$[\boldsymbol{A}, \boldsymbol{\Delta X}^r_{i+1}] \quad = \quad 0\,.$$

This is not always possible. So the minimal requirement would be that the sup-norm of the commutator is bounded by some small constant $\nu \geq 0$.

Example 3.1 (cont.) Any analytic function $u(\boldsymbol{\phi}, \boldsymbol{I})$ that is 2π-periodic in $\boldsymbol{\phi} \in \mathrm{T}^d$ can be written as a multi-dimensional Fourier series

$$u(\boldsymbol{\phi}, \boldsymbol{I}) \quad = \quad \sum_{\boldsymbol{k}\in\mathbb{Z}^d} u_{\boldsymbol{k}}(\boldsymbol{I}) e^{\mathrm{i}\boldsymbol{k}^T\boldsymbol{\phi}}\,.$$

Let us assume that the vector $\boldsymbol{\omega} \in \mathrm{R}^d$ of frequencies $\omega_i > 0$ satisfies a non-resonance condition

$$|\boldsymbol{\omega}^T\boldsymbol{k}| \quad \geq \quad \gamma \; > \; 0\,, \qquad \text{for all } \boldsymbol{k} \in \mathrm{Z}^d_K \setminus \{\boldsymbol{0}\}\,,$$

where $\mathrm{Z}^d_K = \{\boldsymbol{k} \in \mathrm{Z}^d : |\boldsymbol{k}| \leq K\}$. Here $K \gg 1$ is a positive integer sufficiently large and $|\boldsymbol{k}| = |k_1| + \ldots + |k_d|$. Then we define the non-resonant part of u by

$$u^n(\boldsymbol{\phi}, \boldsymbol{I}) \quad := \quad \sum_{\boldsymbol{k}\in\mathbb{Z}^d_K\setminus\{\boldsymbol{0}\}} u_{\boldsymbol{k}}(\boldsymbol{I}) e^{\mathrm{i}\boldsymbol{k}^T\boldsymbol{\phi}}$$

and the resonant part by

$$u^r(\boldsymbol{\phi}, \boldsymbol{I}) \quad := \quad u(\boldsymbol{\phi}, \boldsymbol{I}) - u^n(\boldsymbol{\phi}, \boldsymbol{I})\,.$$

Note that u^r consists of $u_{\boldsymbol{0}}$ and terms that are exponentially small in K. This follows from the exponentially fast decay of the Fourier coefficients $u_{\boldsymbol{k}}$ as $K = |\boldsymbol{k}| \gg 1$. Vector-valued functions $\boldsymbol{u}$ can be treated the same way by considering each component separately. Solving the homological equation (1.13) is equivalent to solving equations of the form

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\phi}} s(\boldsymbol{\phi}, \boldsymbol{I})\boldsymbol{\omega} + u^n(\boldsymbol{\phi}, \boldsymbol{I}) \quad = \quad 0$$

In terms of Fourier expansion, the solution

$$s(\boldsymbol{\phi}, \boldsymbol{I}) \quad = \quad \sum_{\boldsymbol{k}\in\mathbb{Z}^d_K\setminus\{\boldsymbol{0}\}} s_{\boldsymbol{k}}(\boldsymbol{I}) e^{\mathrm{i}\boldsymbol{k}^T\boldsymbol{\phi}}$$

is given by

$$s_{\boldsymbol{k}}(\boldsymbol{I}) \quad := \quad \frac{u_{\boldsymbol{k}}(\boldsymbol{I})}{-\mathrm{i}\,\boldsymbol{\omega}^T\boldsymbol{k}}\,.$$

Note that, by definition of u^n, $|\boldsymbol{\omega}^T\boldsymbol{k}| \geq \gamma$. The commutator of $\boldsymbol{A}$ with the resonant part u^r is exponentially small in K, i.e. $\nu \sim e^{-cK}$, $c > 0$ some constant. □

Remark 3.1. In contrast to the normal form recursion (1.10)-(1.14), the method considered by FASSÒ [36] defines transformed vector fields $\boldsymbol{Y}_i(\epsilon)$ recursively by

$$\boldsymbol{Y}_i(\epsilon) \quad := \quad \boldsymbol{\Psi}_i^*(\epsilon)\boldsymbol{Y}_{i-1}(\epsilon)$$

where $\boldsymbol{\Psi}_i(\epsilon)$ is the time-one-flow map of an appropriate vector field. Thus the overall coordinate transformation is given by the composition of the transformations $\boldsymbol{\Psi}_i(\epsilon)$, i.e.

$$\boldsymbol{\Psi}(\epsilon) \quad = \quad \boldsymbol{\Psi}_1(\epsilon) \circ \boldsymbol{\Psi}_2(\epsilon) \circ \ldots .$$

The normal form truncation error for this method has been discussed by FASSÒ in [36]. □

For simplicity of notation, we will often write $\boldsymbol{W}_i$, $\boldsymbol{Y}_i$, etc. instead of $\boldsymbol{W}_i(\epsilon)$, $\boldsymbol{Y}_i(\epsilon)$, etc. Next we show that the above recursion (1.10)-(1.14) indeed defines a normal form in terms of an asymptotic expansion.

Lemma 3.1. The vector fields $\boldsymbol{Y}_i$, $i \geq 1$, satisfy

$$\boldsymbol{Y}_i \quad = \quad \boldsymbol{A} + \sum_{j=1}^{i} \epsilon^j \boldsymbol{\Delta X}_j^r + \mathcal{O}(\epsilon^{i+1}) .$$

□

Proof. The statement is certainly true for $i = 1$. For $i > 1$, we have (for simplicity, we suppress the arguments)

$$\begin{aligned}
\boldsymbol{Y}_i \quad &:= \quad \boldsymbol{\Phi}_{1,\boldsymbol{W}_i}^* \boldsymbol{Y} , \\
&= \quad \boldsymbol{\Phi}_{1,\epsilon^i \boldsymbol{\Delta W}_i}^* \left[\boldsymbol{\Phi}_{1,\boldsymbol{W}_{i-1}}^* \boldsymbol{Y} \right] + \mathcal{O}(\epsilon^{i+1}) \\
&= \quad \boldsymbol{\Phi}_{1,\epsilon^i \boldsymbol{\Delta W}_i}^* \boldsymbol{Y}_{i-1} + \mathcal{O}(\epsilon^{i+1}) \\
&= \quad \boldsymbol{Y}_{i-1} + [\boldsymbol{Y}_{i-1}, \epsilon^i \boldsymbol{\Delta W}_i] + \mathcal{O}(\epsilon^{i+1}) , \\
&= \quad \boldsymbol{Y}_{i-1} + [\boldsymbol{A}, \epsilon^i \boldsymbol{\Delta W}_i] + \mathcal{O}(\epsilon^{i+1}) , \\
&= \quad \boldsymbol{A} + \sum_{j=1}^{i-1} \epsilon^j \boldsymbol{\Delta X}_j^r + \epsilon^i \left(\boldsymbol{\Delta X}_i + [\boldsymbol{A}, \boldsymbol{\Delta W}_i] \right) + \mathcal{O}(\epsilon^{i+1}) , \\
&= \quad \boldsymbol{A} + \sum_{j=1}^{i-1} \epsilon^j \boldsymbol{\Delta X}_j^r + \epsilon^i \left(\boldsymbol{\Delta X}_i - \boldsymbol{\Delta X}_i^n \right) + \mathcal{O}(\epsilon^{i+1}) , \\
&= \quad \boldsymbol{A} + \sum_{j=1}^{i-1} \epsilon^j \boldsymbol{\Delta X}_j^r + \epsilon^i \boldsymbol{\Delta X}_i^r + \mathcal{O}(\epsilon^{i+1}) , \\
&= \quad \boldsymbol{A} + \sum_{j=1}^{i} \epsilon^j \boldsymbol{\Delta X}_j^r + \mathcal{O}(\epsilon^{i+1}) .
\end{aligned}$$

□

Thus, given the transformed vector field $\boldsymbol{Y}_i$, we also consider its *truncation to normal form of order i*

$$\bar{\boldsymbol{Y}}_i \quad := \quad \boldsymbol{A} + \sum_{j=1}^{i} \epsilon^j \boldsymbol{\Delta X}_j^r$$

This truncation introduces an error of size $\mathcal{O}(\epsilon^{i+1})$. We will denote the *truncation error* by

$$\boldsymbol{T}_i \quad := \quad \boldsymbol{Y}_i - \bar{\boldsymbol{Y}}_i \,. \tag{1.15}$$

The following considerations will be useful for estimating the truncation error in the normal form recursion: Let us define the family of vector fields

$$\boldsymbol{Y}_i(t) \quad := \quad \boldsymbol{\Phi}^*_{t,\boldsymbol{W}_i} \boldsymbol{Y} \,, \qquad t \in [0,1] \,.$$

Then [36]

$$\frac{\partial}{\partial t} \boldsymbol{Y}_i(t = \tau) \quad = \quad \boldsymbol{\Phi}^*_{\tau,\boldsymbol{W}_i} [\boldsymbol{Y}, \boldsymbol{W}_i] \,,$$

$[\boldsymbol{Y}, \boldsymbol{W}_i]$ the Lie bracket (commutator) [90] of the two vector fields $\boldsymbol{Y}$, $\boldsymbol{W}_i$, and the defining equation (1.10) is equivalent to

$$\boldsymbol{Y}_i \quad = \quad \boldsymbol{Y} + \int_0^1 \boldsymbol{\Phi}^*_{t,\boldsymbol{W}_i} [\boldsymbol{Y}, \boldsymbol{W}_i] \, dt \,. \tag{1.16}$$

Upon introducing

$$\boldsymbol{X}_i \quad := \quad \sum_{j=1}^{i} \epsilon^j \boldsymbol{\Delta X}_j \tag{1.17}$$

and the corresponding splitting into the resonant and non-resonant part, i.e.

$$\boldsymbol{X}_i \quad =: \quad \boldsymbol{X}_i^r + \boldsymbol{X}_i^n \,,$$

the equations (1.13)-(1.14) in the normal form recursion can now be replaced by

$$0 \quad = \quad [\boldsymbol{A}, \boldsymbol{W}_i] + \boldsymbol{X}_i^n \tag{1.18}$$

and (1.16) becomes equivalent to

$$\boldsymbol{Y}_i \quad = \quad \boldsymbol{Y} + \int_0^1 \boldsymbol{\Phi}^*_{t,\boldsymbol{W}_i} \left(-\boldsymbol{X}_i^n + \epsilon \, [\boldsymbol{B}, \boldsymbol{W}_i] \right) \, dt \,. \tag{1.19}$$

Furthermore, let us introduce the function $\boldsymbol{f}(\epsilon)$ by

$$\boldsymbol{f}(\epsilon) \quad := \quad \int_0^1 \boldsymbol{\Phi}^*_{t,\boldsymbol{W}_i}[\boldsymbol{Y}, \boldsymbol{W}_i]\, dt\,.$$

Then, for $i \geq 1$,

$$\begin{aligned} \boldsymbol{\Delta X}_{i+1} \quad &:= \quad \frac{1}{(i+1)!}\left[\frac{\partial^{i+1}}{\partial\epsilon^{i+1}}\boldsymbol{Y}_i\right]_{\epsilon=0} \\ &= \quad \frac{1}{(i+1)!}\left[\frac{\partial^{i+1}}{\partial\epsilon^{i+1}}\boldsymbol{Y}\right]_{\epsilon=0} + \frac{1}{(i+1)!}\left[\frac{\partial^{i+1}}{\partial\epsilon^{i+1}}\boldsymbol{f}\right]_{\epsilon=0}\,. \end{aligned} \tag{1.20}$$

3.2 Comments on the Homological Equation

The solvability of the homological equation depends on the solution properties of the vector field $\boldsymbol{A}$ in (0.1). In many applications, the vector field $\boldsymbol{A}$ possesses $d \geq n/2$ first integrals I_j, $j = 1, \ldots, d$. Furthermore, the manifolds

$$\mathcal{M}_{\boldsymbol{c}} \quad := \quad \{\boldsymbol{x} \in \mathrm{R}^n : \boldsymbol{I}(\boldsymbol{x}) = \boldsymbol{c}\}\,, \qquad (\boldsymbol{c} \in \mathcal{V} \subset \mathrm{R}^d)\,,$$

$\mathcal{V}$ an appropriate open subset of R^d, are diffeomorph to the $(n-d)$-torus T^{n-d}. Then the vector field $\boldsymbol{A}$ can be transformed (locally) to

$$\begin{aligned} \frac{d}{dt}\boldsymbol{\phi} \quad &= \quad \boldsymbol{\omega}(\boldsymbol{I})\,, \\ \frac{d}{dt}\boldsymbol{I} \quad &= \quad \boldsymbol{0}\,. \end{aligned}$$

Next we assume that the motion in $\boldsymbol{\phi}$ is ergodic. Then the ensemble average of a vector field $\boldsymbol{X}$ and the time-average of $\boldsymbol{X}$ along trajectories are equal. This also implies the solvability of the homological equation

$$[\boldsymbol{A}, \boldsymbol{W}] + \boldsymbol{X}^n \quad = \quad \boldsymbol{0}\,,$$

where $\boldsymbol{X}^r$ denotes the ensemble average of $\boldsymbol{X}$ and $\boldsymbol{X}^n := \boldsymbol{X} - \boldsymbol{X}^r$. Note that, in this case, $[\boldsymbol{A}, \boldsymbol{X}^r] = \boldsymbol{0}$. This is basically Anosov's averaging principle [4],[7]. To see this, let us consider the time-dependent coordinate transformation

$$\boldsymbol{x} \quad = \quad \boldsymbol{\Phi}_{t,\boldsymbol{A}}(\boldsymbol{u})\,.$$

In the new coordinate $\boldsymbol{u}$, the system (0.1) is equivalent to

$$\frac{d}{dt}\boldsymbol{u} \quad = \quad \epsilon\,\tilde{\boldsymbol{B}}(\boldsymbol{u}, t, \epsilon)$$

with

$$\tilde{\boldsymbol{B}}(\boldsymbol{u}, t, \epsilon) \quad = \quad \boldsymbol{\Phi}^*_{t,\boldsymbol{A}}\boldsymbol{B}\,.$$

This system is highly oscillatory in the variable t. These oscillations can be eliminated by averaging in time, i.e.

$$\bar{\boldsymbol{B}}(\boldsymbol{u},\epsilon) \quad := \quad \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{+T} \tilde{\boldsymbol{B}}(\boldsymbol{u},t,\epsilon)\,dt\,.$$

Upon assuming ergodicity, this time-average is equivalent to the ensemble average and we obtain

$$\boldsymbol{\Delta X}_1^r \quad = \quad \bar{\boldsymbol{B}}(\epsilon = 0)\,.$$

Often the motion in $\boldsymbol{\phi}$ is not completely ergodic (resonances) or only over very long periods of time (problem of small denominators), then we first have to find a truncation $\boldsymbol{X}^t$ of $\boldsymbol{X}$ for which the time average and the ensemble average are identical, converge faster respectively. Denote the ensemble average of the truncation $\boldsymbol{X}^t$ by $\boldsymbol{X}^a$. Then the resonant part of $\boldsymbol{X}$ is defined by $\boldsymbol{X}^r := (\boldsymbol{X} - \boldsymbol{X}^t) + \boldsymbol{X}^a$ and the non-resonant part by $\boldsymbol{X}^n := \boldsymbol{X}^t - \boldsymbol{X}^a$. This again insures the solvability of the homological equation. Unfortunately we have $[\boldsymbol{A},\boldsymbol{X}^r] = [\boldsymbol{A},\boldsymbol{X} - \boldsymbol{X}^t] \neq \boldsymbol{0}$, in general. However, we can often make the truncation "error" $\boldsymbol{X} - \boldsymbol{X}^t$ exponentially small (see Example 3.1 below).

Example 3.1 (cont.) If the vector $\boldsymbol{\omega}$ of frequencies is resonant, i.e., there exists a $\boldsymbol{k} \in \mathrm{Z}^d \setminus \{\boldsymbol{0}\}$ such that $\boldsymbol{\omega}^T\boldsymbol{k} = 0$, then the system

$$\frac{d}{dt}\boldsymbol{\phi} \quad = \quad \boldsymbol{\omega}$$

is not ergodic. Even if $\boldsymbol{\omega}^T\boldsymbol{k} \neq 0$ for all $\boldsymbol{k} \in \mathrm{Z}^d \setminus \{\boldsymbol{0}\}$, we typically have an estimate

$$|\boldsymbol{\omega}^T\boldsymbol{k}| \quad \geq \quad \frac{\sigma}{K^{d-1}}\,, \qquad K = |\boldsymbol{k}|\,,\, 0 < \sigma \leq 1\,,$$

and $\boldsymbol{\omega}^T\boldsymbol{k}$ can become arbitrarily small for $|\boldsymbol{k}|$ large enough.

Thus ergodicity will hold only when averaging over extremely long periods of time. This motivates the use of a truncated Fourier expansion in the definition of $\boldsymbol{X}^n$. Since the Fourier coefficients decay exponentially fast with $K = |\boldsymbol{k}| \gg 1$, $[\boldsymbol{A},\boldsymbol{X}^r]$ can be made exponentially small in K. □

3.3 The Theorem on the Exponential Estimate

3.3.1 Definitions and Assumptions

Let $\mathcal{K} \subset \mathrm{R}^n$ be a compact subset of R^n. Then $\mathcal{B}_R\mathcal{K} \subset \mathrm{C}^n$ denotes the complex neighborhood of radius $R > 0$ around $\mathcal{K}$ with respect to the norm

$$\begin{aligned} ||\boldsymbol{x}|| \quad &:= \quad \max_{i=1,\dots,n} ||x_i|| \qquad \boldsymbol{x} \in \mathrm{C}^n\,, \\ ||x_i|| \quad &:= \quad \left([\mathrm{Re}(x_i)]^2 + [\mathrm{Im}(x_i)]^2\right)^{1/2}\,, \end{aligned}$$

$\boldsymbol{x} = (x_1, \dots, x_n)^T$. In other words,

$$\mathcal{B}_R\mathcal{K} \quad := \quad \bigcup_{\boldsymbol{x}_0 \in \mathcal{K}} \mathcal{B}_R(\boldsymbol{x}_0)$$

and

$$\mathcal{B}_R(\boldsymbol{x}_0) \quad := \quad \{\boldsymbol{x} \in \mathrm{C}^n : ||\boldsymbol{x} - \boldsymbol{x}_0|| \leq R\}\,.$$

Let $\boldsymbol{X} : \mathcal{U} \subset \mathrm{R}^n \to \mathrm{R}^n$, $\mathcal{K} \subset \mathcal{U}$, be a real analytic function. Then the usual sup-norm on $\mathcal{B}_R\mathcal{K}$ is defined by

$$||\boldsymbol{X}||_R \quad := \quad \sup_{\boldsymbol{x} \in \mathcal{B}_R\mathcal{K}} ||\boldsymbol{X}(\boldsymbol{x})||\,.$$

Assumptions. We assume that there is a second norm[1] $|\boldsymbol{X}|_R$ of $\boldsymbol{X}$ on $\mathcal{B}_R\mathcal{K}$ such that

(i) $||\boldsymbol{X}||_R \leq |\boldsymbol{X}|_R$,

(ii) the splitting of $\boldsymbol{X}$ into its resonant part $\boldsymbol{X}^r$ and it non-resonant part $\boldsymbol{X}^n$ satisfies $|\boldsymbol{X}^n|_R \leq |\boldsymbol{X}|_R$ and $|\boldsymbol{X}^r|_R \leq |\boldsymbol{X}|_R$,

(iii) The solution of the homological equation

$$[\boldsymbol{A}, \boldsymbol{W}] + \boldsymbol{X}^n \quad = \quad \boldsymbol{0}$$

satisfies

$$|\boldsymbol{W}|_R \quad \leq \quad \frac{1}{\gamma}\,|\boldsymbol{X}^n|_R\,. \tag{3.21}$$

Here $\gamma > 0$ is some positive constant.

We assume that the vector field $\boldsymbol{B}$ in (0.1) satisfies

$$|\,\boldsymbol{B}(\epsilon)\,|_R \quad \leq \quad 1 \tag{3.22}$$

for all $\epsilon \leq \gamma R/3$. □

We will also need the following two estimates:

Lemma 3.2. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two real analytic vector fields with $|\boldsymbol{X}|_{r+\sigma} \leq M_1$ and $|\boldsymbol{Y}|_{r+\sigma} \leq M_2$, then the vector field $\boldsymbol{Z} = [\boldsymbol{X}, \boldsymbol{Y}]$ satisfies

$$|\boldsymbol{Z}|_r \quad \leq \quad \frac{2\,M_1\,M_2}{\sigma}\,.$$

[1] This norm is first defined for scalar-valued functions on $\mathcal{B}_R\mathcal{K}$. The corresponding norm for vector-valued functions is obtained by taking the maximum over the norm of all components. See Example 3.1 below.

□

Proof. This estimate follows from Cauchy's estimate [64],[93].

Lemma 3.3. Let $\boldsymbol{W}$ be a real analytic vector field with $|\boldsymbol{W}|_{r+k\sigma} \leq \sigma$ where $r, \sigma > 0$ and $k > 1$. Consider a second real analytic vector field $\boldsymbol{Z}$ with $|\boldsymbol{Z}|_r \leq m$. Then

$$|\,\boldsymbol{\partial_x \Phi}_{t,\boldsymbol{W}} \cdot \boldsymbol{Z}\,|_r \quad \leq \quad \frac{k}{k-1}\,|\,\boldsymbol{Z}\,|_r \tag{3.23}$$

for all $t \in [0,1]$. □

Proof. Let us define the function

$$\boldsymbol{f}(\boldsymbol{x},t,\lambda) \quad := \quad \boldsymbol{\Phi}_{t,\boldsymbol{W}}(\boldsymbol{x} + \lambda \boldsymbol{Z}(\boldsymbol{x})) - \boldsymbol{x}\,.$$

Note that $\boldsymbol{f}$ can also be written as

$$\boldsymbol{f}(\boldsymbol{x},t,\lambda) \quad = \quad \lambda \boldsymbol{Z}(\boldsymbol{x}) + \int_0^t \boldsymbol{W}(\boldsymbol{\Phi}_{\tau,\boldsymbol{W}}(\boldsymbol{x} + \lambda \boldsymbol{Z}(\boldsymbol{x})))\,d\tau\,.$$

Next we derive an estimate for $|\boldsymbol{f}(t,\lambda)|_r$ for $t \in [0,1]$ and $|\lambda| \leq \lambda_0$ with $\lambda_0 > 0$ still to be specified. First we obtain

$$|\,\boldsymbol{f}(t,\lambda)\,|_r \quad \leq \quad \lambda\,|\,\boldsymbol{Z}\,|_r + \int_0^t |\,\boldsymbol{W}(\boldsymbol{\Phi}_{\tau,\boldsymbol{W}}(\mathbf{id} + \lambda \boldsymbol{Z}))\,|_r\,d\tau\,.$$

From this we get

$$|\,\boldsymbol{f}(t,\lambda)\,|_r \quad \leq \quad \lambda_0\,m + \sigma \;=\; k\,\sigma$$

for

$$|\lambda| \quad \leq \quad \lambda_0 \;:=\; \frac{(k-1)\,\sigma}{m}\,.$$

Here λ_0 was determined from the requirement that

$$\boldsymbol{x} + \lambda \boldsymbol{Z}(\boldsymbol{x}) \quad \in \quad \mathcal{B}_{r+(k-1)\sigma}\mathcal{K}$$

for all $\boldsymbol{x} \in \mathcal{B}_r\mathcal{K}$. This implies

$$\boldsymbol{\Phi}_{\tau,\boldsymbol{W}}(\boldsymbol{x} + \lambda \boldsymbol{Z}(\boldsymbol{x})) \quad \in \quad \mathcal{B}_{r+k\sigma}\mathcal{K}$$

for all $\tau \in [0,1]$ and all $\boldsymbol{x} \in \mathcal{B}_r\mathcal{K}$. Finally, we apply the usual Cauchy inequality [93] to obtain

$$\begin{aligned} |\,\boldsymbol{\partial_x \Phi}_{t,\boldsymbol{W}} \cdot \boldsymbol{Z}\,|_r \quad &= \quad |\,\partial_\lambda \boldsymbol{f}(t,\lambda = 0)\,|_r\,, \\ &\leq \quad \sigma\,k\,\lambda_0^{-1} \;=\; \frac{k}{k-1}\,|\,\boldsymbol{Z}\,|_r\,. \end{aligned}$$

□

Example 3.1 (cont.) For a problem of type (0.5)-(0.6), the following exponentially weighted norm is suitable: If u is a real analytic function on $\mathcal{B}_r\mathcal{K} = \mathcal{B}_r\mathrm{T}^d \times \mathcal{B}_r\tilde{\mathcal{K}}$, $\tilde{\mathcal{K}} \subset \mathrm{R}^d$ a compact subset, $0 < r \leq 1$, with Fourier expansion

$$u(\boldsymbol{\phi}, \boldsymbol{I}) \quad = \quad \sum_{\boldsymbol{k}\in\mathbb{Z}^d} u_{\boldsymbol{k}}(\boldsymbol{I}) e^{\mathrm{i}\boldsymbol{k}^T\boldsymbol{\phi}}\,,$$

then

$$|u|_r \quad := \quad \sup_{\boldsymbol{I}\in\mathcal{B}_r\tilde{\mathcal{K}}} \sum_{\boldsymbol{k}\in\mathbb{Z}^d} ||u_{\boldsymbol{k}}(\boldsymbol{I})||\, e^{|\boldsymbol{k}|r}$$

and [93]

$$||u||_r \quad \leq \quad |u|_r \;\leq\; \coth^d\sigma\, ||u||_{r+\sigma}\,.$$

For vector-valued functions $\boldsymbol{u} = (u_1, \ldots, u_d)^T$ we define

$$|\boldsymbol{u}|_r \quad := \quad \max_{j=1,\ldots,d} |u_j|_r\,.$$

If the vector $\boldsymbol{\omega}$ of frequencies satisfies the non-resonance condition

$$|\boldsymbol{\omega}^T\boldsymbol{k}| \quad \geq \quad \gamma\,, \quad \text{for all } \boldsymbol{k} \in \mathrm{Z}^d_K \setminus \{\boldsymbol{0}\}\,, \tag{3.24}$$

then the solution s of the homological equation

$$\frac{\boldsymbol{\partial}}{\boldsymbol{\partial\phi}} s(\boldsymbol{\phi}, \boldsymbol{I})\, \boldsymbol{\omega} + u^n(\boldsymbol{\phi}, \boldsymbol{I}) \quad = \quad 0 \tag{3.25}$$

satisfies

$$|s|_r \quad \leq \quad \gamma^{-1}|u^n|_r \;\leq\; \gamma^{-1}|u|_r$$

with

$$u^n(\boldsymbol{\phi}, \boldsymbol{I}) \quad := \quad \sum_{\boldsymbol{k}\in\mathbb{Z}^d_K\setminus\{\boldsymbol{0}\}} u_{\boldsymbol{k}}(\boldsymbol{I})\, e^{\mathrm{i}\boldsymbol{k}^T\boldsymbol{\phi}}\,.$$

The truncated expansion term

$$u^t(\boldsymbol{\phi}, \boldsymbol{I}) \quad := \quad \sum_{|\boldsymbol{k}|>K} u(\boldsymbol{I}) e^{\mathrm{i}\boldsymbol{k}^T\boldsymbol{\phi}}$$

can be estimated by

$$|u^t|_{r/2} \quad \leq \quad |u|_r\, e^{-Kr/2}\,.$$

One could also work in the $||u||_r$ norm and use an estimate due to RÜSSMAN [91]. Namely, if the vector of frequencies $\boldsymbol{\omega}$ satisfies the diophantine condition

$$|\,\boldsymbol{\omega}^T \boldsymbol{k}\,| \quad > \quad \frac{\sigma}{|\boldsymbol{k}|^{d-1}}\,, \qquad \text{for all}\, \boldsymbol{k} \in \mathrm{Z}^d \setminus \{\boldsymbol{0}\}\,,$$

then the corresponding homological equation (3.25) has solution s with

$$||\, s\,||_{r-\delta} \quad \leq \quad \frac{c}{\sigma\,\delta^{d-1}}\,||\,u^n\,||_r$$

where $c = \sqrt{(2d-2)!}$ and

$$u^n(\boldsymbol{\phi}, \boldsymbol{I}) \quad := \sum_{\boldsymbol{k}\in\mathbb{Z}^d\setminus\{\boldsymbol{0}\}} u_{\boldsymbol{k}}(\boldsymbol{I})\, e^{\mathrm{i}\boldsymbol{k}^T\boldsymbol{\phi}}\,.$$

The application of this estimate would require some minor modifications in the result and the proof of Theorem 3.1 (Proposition 3.1, respectively). □

3.3.2 The Theorem

We would like to have an estimate for the difference between the transformed vector field $\boldsymbol{Y}_i$ and its normal form truncation of order i

$$\bar{\boldsymbol{Y}}_i \quad := \quad \boldsymbol{A} + \sum_{j=1}^{i} \epsilon^j \boldsymbol{\Delta X}^r_j\,.$$

As for the backward error analysis, the key to the solution is to find an optimal truncation index $i = i_*(\epsilon)$.

Theorem 3.1. Let us assume that the Assumptions made in Section 3.3.1 hold for a given system (0.1). Define $i_*(\epsilon)$ as the integer part of

$$i_o(\epsilon) \quad := \quad \frac{\gamma\, R}{4\, c\, e\, \epsilon}\,.$$

Then

$$\begin{aligned} ||\,\boldsymbol{T}_{i_*}\,||_{R/2} \quad &= \quad ||\,\boldsymbol{Y}_{i_*} - \bar{\boldsymbol{Y}}_{i_*}\,||_{R/2}\,, \\ &\leq \quad 6\,\epsilon\, b\, e^{-\mu/\epsilon} \end{aligned}$$

with $\mu = \gamma R/(4ce)$, $b = 60$, and $c = 36000$.

The transforming vector field $\boldsymbol{W}_{i_*}$ satisfies the estimate

$$||\,\boldsymbol{W}_{i_*}\,||_{R/2} \quad \leq \quad b\,\epsilon\,\gamma^{-1}$$

while for the corresponding time-one-flow map $\boldsymbol{\Phi}_{1,\boldsymbol{W}_{i_*}}$ the estimate

$$||\,\boldsymbol{\Phi}_{1,\boldsymbol{W}_{i_*}} - \mathbf{id}\,||_{R/4} \quad \leq \quad b\,\epsilon\,\gamma^{-1}$$

holds. □

Example 3.1 (cont.) Let the vector $\boldsymbol{\omega}$ of frequencies satisfy a non-resonance condition (3.24). Then it follows from Theorem 3.1 that the system (0.5)-(0.6) can be transformed to normal form

$$\begin{aligned}\frac{d}{dt}\bar{\boldsymbol{\phi}} &= \boldsymbol{\omega} + \epsilon\,\bar{\boldsymbol{f}}(\bar{\boldsymbol{I}};\epsilon)\,,\\ \frac{d}{dt}\bar{\boldsymbol{I}} &= \quad + \epsilon\,\bar{\boldsymbol{g}}(\bar{\boldsymbol{I}};\epsilon)\end{aligned}$$

up to terms exponentially small in ϵ and terms exponentially small in K, $K \gg 1$ the truncation index in the Fourier series expansion. □

3.4 Proof of the Theorem

3.4.1 The First Two Estimates

The estimate (3.22) and the assumption (3.21) imply that

$$|\boldsymbol{\Delta X}_1|_R \leq 1 \quad \text{and} \quad |\boldsymbol{\Delta W}_1|_R \leq \gamma^{-1}\,. \tag{4.26}$$

It follows that $||\boldsymbol{W}_1||_R \leq |\boldsymbol{W}_1|_R \leq \epsilon\,\gamma^{-1}$ and

$$\boldsymbol{\Phi}_{t,\boldsymbol{W}_1}(\bar{\boldsymbol{x}}) \quad \in \quad \mathcal{B}_{\alpha R+(1-\alpha)R/3}\mathcal{K}$$

for all $t \in [0,1]$, all $\bar{\boldsymbol{x}} \in \mathcal{B}_{\alpha R}\mathcal{K}$, and all

$$\epsilon \leq \epsilon_0 \quad := \quad \frac{\gamma\,(1-\alpha)\,R}{3}\,.$$

Let us define the family of functions $\boldsymbol{f}(\epsilon)$ by

$$\boldsymbol{f}(\bar{\boldsymbol{x}};\epsilon) \quad := \quad \int_0^1 \left[\frac{\boldsymbol{\partial}}{\boldsymbol{\partial x}}\boldsymbol{\Phi}_{t,-\boldsymbol{W}_1}\cdot[\boldsymbol{Y},\boldsymbol{W}_1]\right]\circ\boldsymbol{\Phi}_{t,\boldsymbol{W}_1}(\bar{\boldsymbol{x}})\,dt$$

with

$$\boldsymbol{Y} \quad = \quad \boldsymbol{A} + \epsilon\,\boldsymbol{B}(\epsilon)\,.$$

Then, because of (1.20),

$$\boldsymbol{\Delta X}_2 \quad = \quad \frac{1}{2!}\left[\frac{\partial^2}{\partial\epsilon^2}\epsilon\,\boldsymbol{B}\right]_{\epsilon=0} + \frac{1}{2!}\left[\frac{\partial^2}{\partial\epsilon^2}\boldsymbol{f}\right]_{\epsilon=0}$$

which we find an estimate for by applying Cauchy's estimate. In particular, since

$$|\epsilon\,\boldsymbol{B}|_{\alpha R} \quad \leq \quad \epsilon_0$$

for all $\epsilon \leq \epsilon_0 \leq \gamma R/3$, and (3.23) with $r = \alpha R + \sigma$,

$$\sigma \;=\; \epsilon_0/\gamma \;=\; \frac{1}{3}\,(1-\alpha)\,R\,,$$

and $k = 2$, we obtain

$$\begin{aligned}
|\boldsymbol{f}|_{\alpha R} &\leq \max_{t\in[0,1]} |\,\boldsymbol{\partial_x \Phi}_{t,-\boldsymbol{W}_1} \cdot [\boldsymbol{Y}, \boldsymbol{W}_1]\,|_{\alpha R+\sigma}\,,\\
&\leq 2\,|\,[\boldsymbol{Y}, \boldsymbol{W}_1]\,|_{\alpha R+\sigma}\,,\\
&\leq 2\,|\,[\boldsymbol{A}, \boldsymbol{W}_1]\,|_{\alpha R+\sigma} + 2\,\epsilon_0\,|\,[\boldsymbol{B}, \boldsymbol{W}_1]\,|_{\alpha R+\sigma}\,,\\
&\leq 2\,\epsilon_0\,|\boldsymbol{\Delta X}_1|_R + 2\,\epsilon_0\,\frac{2\,|\boldsymbol{B}|_R\,|\boldsymbol{W}_1|_R}{2\,\sigma}\,,\\
&\leq 2\,\epsilon_0 + 2\,\epsilon_0\,\frac{2\,\sigma}{2\,\sigma}\,,\\
&\leq 4\,\epsilon_0
\end{aligned}$$

for all $|\epsilon| \leq \epsilon_0$. Thus Cauchy's estimate yields

$$\begin{aligned}
|\epsilon^2\,\boldsymbol{\Delta X}_2|_{\alpha R} &\leq \epsilon_0\left(\frac{\epsilon}{\epsilon_0}\right)^2 + 4\,\epsilon_0\left(\frac{\epsilon}{\epsilon_0}\right)^2,\\
&\leq \frac{15\,\epsilon^2}{\gamma\,(1-\alpha)\,R}\,.
\end{aligned} \tag{4.27}$$

3.4.2 The General Estimate

Lemma 3.4. The vector fields $\boldsymbol{\Delta X}_i$ (1.11) satisfy

$$\epsilon^i\,|\,\boldsymbol{\Delta X}_i\,|_{\alpha R} \;\leq\; b\,\epsilon\left(\frac{c\,(i-1)\,\epsilon}{\gamma\,(1-\alpha)R}\right)^{i-1} \tag{4.28}$$

for $i \geq 3$ and $\alpha \in [0,1)$. The constants b and c can be chosen as

$$b \;=\; 60 \qquad \text{and} \qquad c = 36000\,.$$

□

Proof. Let us assume that (4.28) holds for $i = 3, \ldots, j$. Then

$$\begin{aligned}
|\,\boldsymbol{X}_j\,|_{\alpha R} &\leq \sum_{i=1}^{j} \epsilon^i\,|\,\boldsymbol{\Delta X}_i\,|_{\alpha R}\\
&\leq \epsilon\left[1 + \frac{15\,\epsilon}{\gamma\,(1-\alpha)\,R} + \sum_{i=3}^{j} b\left(\frac{c\,(i-1)\,\epsilon}{\gamma\,(1-\alpha)\,R}\right)^{i-1}\right]
\end{aligned} \tag{4.29}$$

which implies

$$|\,\boldsymbol{X}_j\,|_{\alpha R} \;\leq\; 0.95\,b\,\epsilon_0 \;\leq\; \delta\,\gamma\,(1-\alpha)\,R \tag{4.30}$$

for

$$\epsilon \quad \leq \quad \epsilon_0 \ := \ \frac{\gamma\,(1-\alpha)R}{c\,j}\,,$$

$b \geq 60$, $c \geq 36000$, and

$$\delta \quad := \quad \frac{0.95\,b}{c\,j}\,.$$

Here we have used that

$$\sum_{i=3}^{j}\left(\frac{i-1}{j}\right)^{i-1} \quad \leq \quad 0.85$$

for $j \geq 3$ which implies that

$$\begin{aligned} 1+\frac{15\,\epsilon_0}{\gamma\,(1-\alpha)\,R}+\sum_{i=3}^{j} b\left(\frac{c\,(i-1)\,\epsilon_0}{\gamma\,(1-\alpha)\,R}\right)^{i-1} \quad &= \quad 1+\frac{15}{c\,j}+b\left[\sum_{i=3}^{j}\left(\frac{i-1}{j}\right)^{i-1}\right] \\ &\leq \quad 0.95\,b \end{aligned}$$

for $j \geq 3$, $b \geq 60$, and $c \geq 36000$. From (4.30) and (3.21), we obtain immediately

$$|\,\boldsymbol{W}_j\,|_{\alpha R} \quad \leq \quad \gamma^{-1}|\,\boldsymbol{X}_j\,|_{\alpha R} \ \leq \ \delta\,(1-\alpha)\,R\,.$$

Next we verify that for $b \geq 60$ and $c \geq 36000$

$$|\,\boldsymbol{W}_j\,|_{(\alpha+40\delta(1-\alpha))R} \quad \leq \quad \delta(1-\alpha)R \tag{4.31}$$

for $\epsilon \leq \epsilon_0$ as well. In other words, we chose b and c such that

$$1+\frac{15}{(1-40\delta)\,c\,j}+b\sum_{i=3}^{j}\left(\frac{i-1}{(1-40\delta)\,j}\right)^{i-1} \quad \leq \quad 0.95\,b$$

where we have used (4.29) with α replaced by $\alpha+40\delta(1-\alpha)$ and

$$1-(\alpha+(1-\alpha)\,40\delta) \quad = \quad (1-\alpha)\,(1-40\delta)\,.$$

In particular, for $b = 60$ and $c = 36000$, we obtain

$$\sum_{i=3}^{j}\left(\frac{i-1}{(1-40\delta)\,j}\right)^{i-1} \quad \leq \quad 0.89$$

for $j \geq 3$. Let us now consider the vector-valued function

$$\begin{aligned} \boldsymbol{f}\,(\bar{\boldsymbol{x}};\epsilon) \quad &:= \quad \int_0^1 \left[\boldsymbol{\partial_x}\boldsymbol{\Phi}_{t,-\boldsymbol{W}_j}\cdot[\boldsymbol{Y}\,,\boldsymbol{W}_j]\right]\circ\boldsymbol{\Phi}_{t,\boldsymbol{W}_j}\,(\bar{\boldsymbol{x}})\,dt\,, \\ &= \quad \boldsymbol{Y}_j(\bar{\boldsymbol{x}};\epsilon)-\boldsymbol{Y}(\bar{\boldsymbol{x}};\epsilon) \end{aligned}$$

for $\bar{\boldsymbol{x}} \in \mathcal{B}_{\alpha R}\mathcal{K}$. Since (4.31), we have

$$||\,\boldsymbol{W}_j||_{\alpha R+40\sigma} \quad \leq \quad \sigma\,, \qquad \sigma \;:=\; \delta\,(1-\alpha)\,R\,.$$

Thus

$$\boldsymbol{\Phi}_{t,\boldsymbol{W}_j}(\bar{\boldsymbol{x}}) \quad \in \quad \mathcal{B}_{\alpha R+\sigma}\mathcal{K}$$

for $t \in [0,1]$, $\bar{\boldsymbol{x}} \in \mathcal{B}_{\alpha R}\mathcal{K}$. Next we use (3.23) with $r = \alpha R + \sigma$, $\sigma = \delta(1-\alpha)R$, and $k = 39$ to obtain

$$\begin{aligned}
|\boldsymbol{f}|_{\alpha R} \quad &\leq \quad \max_{t\in[0,1]} |\,\boldsymbol{\partial}_{\boldsymbol{x}}\boldsymbol{\Phi}_{t,-\boldsymbol{W}_j}\cdot[\boldsymbol{Y},\boldsymbol{W}_j]\,|_{\alpha R+\sigma}\,,\\
&\leq \quad (39/38)\,|\,[\boldsymbol{Y},\boldsymbol{W}_j]\,|_{\alpha R+\sigma}\\
&\leq \quad 1.027\,|\,[\boldsymbol{A},\boldsymbol{W}_j]\,|_{\alpha R+\sigma} + 1.027\,\epsilon_0\,|\,[\boldsymbol{B},\boldsymbol{W}_j]\,|_{\alpha R+\sigma}\,.\\
&\leq \quad 1.027\,|\boldsymbol{X}_j|_{\alpha R+\sigma} + 1.027\,\epsilon_0\,\frac{2\,|\boldsymbol{B}|_R\,|\boldsymbol{W}_j|_{\alpha R+40\sigma}}{39\,\sigma}\,,\\
&\leq \quad 1.027\,\epsilon_0\ (0.95\,b + 2/39)\ ,\\
&\leq \quad \epsilon_0\,(b-1)
\end{aligned}$$

for all $|\epsilon| \leq \epsilon_0$, and, by Cauchy's estimate,

$$\begin{aligned}
|\epsilon^{j+1}\,\boldsymbol{\Delta X}_{j+1}|_{\alpha R} \quad &= \quad |\,\frac{\epsilon^{j+1}}{(j+1)!}\left[\frac{\partial^{j+1}}{\partial\epsilon^{j+1}}\epsilon\boldsymbol{B}\right]_{\epsilon=0}|_{\alpha R} + |\,\frac{\epsilon^{j+1}}{(j+1)!}\left[\frac{\partial^{j+1}}{\partial\epsilon^{j+1}}\boldsymbol{f}\right]_{\epsilon=0}|_{\alpha R}\\
&\leq \quad \epsilon_0\left(\frac{\epsilon}{\epsilon_0}\right)^{j+1} + (b-1)\,\epsilon_0\left(\frac{\epsilon}{\epsilon_0}\right)^{j+1}\,,\\
&\leq \quad b\,\epsilon\left(\frac{c\,j\,\epsilon}{\gamma\,(1-\alpha)\,R}\right)^j\,.
\end{aligned}$$

as required. □

3.4.3 Optimal Truncation Index

We need an estimate for the difference between $\boldsymbol{Y}_i$ and its normal form truncation

$$\bar{\boldsymbol{Y}}_i \quad = \quad \boldsymbol{A} + \sum_{j=1}^{i}\epsilon^j\boldsymbol{\Delta X}^r_j\,.$$

By standard Taylor series expansion, we know that

$$\begin{aligned}
||\,\boldsymbol{Y}_i(\bar{\boldsymbol{x}};\epsilon) - \bar{\boldsymbol{Y}}_i(\bar{\boldsymbol{x}};\epsilon)\,|| \quad &= \quad ||\,\epsilon^{i+1}\,\boldsymbol{\Delta X}_{i+1}(\bar{\boldsymbol{x}})\,|| + \mathcal{O}(\epsilon^{i+2})\,,\\
&\leq \quad \frac{\epsilon^{i+1}}{(i+1)!}\sup_{0\leq\hat{\epsilon}\leq\epsilon}||\,\partial_\epsilon^{i+1}\boldsymbol{Y}_i(\bar{\boldsymbol{x}};\epsilon=\hat{\epsilon})\,||\,.
\end{aligned}$$

This requires an estimate for $||\partial_\epsilon^{i+1}\boldsymbol{Y}_i(\bar{\boldsymbol{x}};\epsilon=\hat{\epsilon})||$, $0 \leq \hat{\epsilon} \leq \epsilon$. Following the proof of Lemma 3.4 and taking $\alpha = 1/2$, we obtain

$$||\,\boldsymbol{Y}_i(\epsilon) - \boldsymbol{A}\,||_{R/2} \quad \leq \quad b\,\epsilon_0$$

for $|\epsilon - \hat{\epsilon}| \leq \epsilon_0/2$, $\hat{\epsilon} \leq \epsilon_0/2$, and $\epsilon_0 = \gamma R/(2ci)$. Thus Cauchy's estimate implies

$$\frac{\epsilon^{i+1}}{(i+1)!} \, || \, \partial_\epsilon^{i+1} \boldsymbol{Y}_i(\epsilon = \hat{\epsilon}) \, ||_{R/2} \quad \leq \quad b \, \epsilon_0 \left(\frac{2\,\epsilon}{\epsilon_0} \right)^{i+1}$$

for $\hat{\epsilon} \leq \epsilon_0/2$ and the estimate

$$|| \, \boldsymbol{Y}_i(\bar{\boldsymbol{x}}; \epsilon) - \bar{\boldsymbol{Y}}_i(\bar{\boldsymbol{x}}; \epsilon) \, || \quad \leq \quad 2\, b\, \epsilon \left(\frac{4\, c\, i\, \epsilon}{\gamma\, R} \right)^i , \qquad (\bar{\boldsymbol{x}} \in \mathcal{B}_{R/2}\mathcal{K}) \, ,$$

follows for $\epsilon \leq \epsilon_0/2$. We now determine an "optimal" number of iterations $i_*(\epsilon)$: Let $i_o(\epsilon)$ be defined by

$$\frac{4\, c\, \epsilon\, i_o}{\gamma\, R} \quad = \quad e^{-1}$$

and take $i_*(\epsilon)$ as the integer part of $i_o(\epsilon)$, i.e.,

$$i_*(\epsilon) \quad := \quad \left[\frac{\gamma\, R}{4\, e\, c\, \epsilon} \right]$$

were $[x]$ denotes the integer part of a positive real number x. Finally, we use $e^{-i_*} \leq e^{-i_o+1} < 3e^{-\mu/\epsilon}$. This concludes the first part of the proof.

The transforming vector field $\boldsymbol{W}_{i_*}$ satisfies

$$|| \, \boldsymbol{W}_{i_*} \, ||_{R/2} \quad \leq \quad \epsilon\, \gamma^{-1} \left[1 + \frac{30\, \epsilon}{\gamma\, R} + \sum_{i=3}^{i_*} b \left(\frac{2\, c\, (i-1)\, \epsilon}{\gamma\, R} \right)^{i-1} \right] .$$

With

$$\epsilon \quad \leq \quad \frac{\gamma\, R}{4\, e\, c\, i_*}$$

the estimate

$$|| \, \boldsymbol{W}_{i_*} \, ||_{R/2} \quad \leq \quad b\, \epsilon\, \gamma^{-1}$$

follows. Since $b\, \epsilon\, \gamma^{-1} < R/4$, the desired estimate for the time-one-flow map $\boldsymbol{\Phi}_{1, \boldsymbol{W}_{i_*}}$ is obvious. $\square$

One often encounters problems where the parameter γ depends in the iteration index i. For example, the vector $\boldsymbol{\omega}$ of frequencies in Example 3.1 typically satisfies

$$|\boldsymbol{\omega}^T \boldsymbol{k}| \quad \geq \quad \frac{\sigma}{K^{d-1}} \, , \qquad K = |\boldsymbol{k}| \, , \, \sigma > 0 \, .$$

Then, working with truncated Fourier expansions, we chose K as $K = i_o L$. $L > 0$ an integer sufficiently large such that the exponentially small truncation error in the Fourier expansion is negligible. However this implies that $i_o(\epsilon)$ is now determined by

$$\frac{4\, c\, \epsilon\, i_o}{\gamma(i_o)\, R} \quad = \quad \frac{4\, c\, \epsilon\, i_o^d\, L^{d-1}}{\sigma\, R} \quad = \quad e^{-1}$$

and, therefore,

$$i_*(\epsilon) \quad := \quad \left[\left(\frac{\sigma\, R}{4\, e\, c\, \epsilon\, L^{d-1}} \right)^{1/d} \right] .$$

Thus Theorem 3.1 has to be replaced by:

Proposition 3.1. Let us assume that the assumptions made in Section 3.3.1 hold for a given system (0.1). Let us also assume that the parameter γ in (3.21) satisfies

$$\gamma(i) \quad = \quad \frac{\delta}{i^{d-1}} \,,$$

$\delta > 0$, $d \geq 1$ some appropriate constants and i the iteration index of the normal form iteration (1.10)-(1.14). Define $i_*(\epsilon)$ as the integer part of

$$i_o(\epsilon) \quad := \quad \left(\frac{\delta\, R}{4\, c\, e\, \epsilon} \right)^{1/d} .$$

Then

$$\begin{aligned} || \boldsymbol{T}_{i_*} ||_{R/2} \quad &= \quad || \boldsymbol{Y}_{i_*} - \bar{\boldsymbol{Y}}_{i_*} ||_{R/2} \,, \\ &\leq \quad 6\, \epsilon\, b\, e^{-(\mu/\epsilon)^{1/d}} \end{aligned} \tag{4.32}$$

with $\mu = \delta R/(4ce)$, $b = 60$, and $c = 36000$.

The transforming vector field $\boldsymbol{W}_{i_*}$ satisfies the estimate

$$|| \boldsymbol{W}_{i_*} ||_{R/2} \quad \leq \quad \frac{b\, R}{4\, c\, e} \left(\frac{4\, c\, e\, \epsilon}{\delta\, R} \right)^{1/d}$$

while for the corresponding time-one-flow map $\boldsymbol{\Phi}_{1,\boldsymbol{W}_{i_*}}$ the estimate

$$|| \boldsymbol{\Phi}_{1,\boldsymbol{W}_{i_*}} - \mathbf{id} ||_{R/4} \quad \leq \quad \frac{b\, R}{4\, c\, e} \left(\frac{4\, c\, e\, \epsilon}{\delta\, R} \right)^{1/d}$$

holds. □

Remark 3.2. Note that d is typically proportional to the degrees of freedom in the system. Thus, for many degrees of freedom systems, exponential estimates of type (4.32) are probably not as useful.

As for backward error analysis, better estimates can be obtained by making the constants c, b, and γ dependent on the normal form iteration index i. Using the exponentially weighted norm from Example 3.1, one might even be able to obtain results qualitatively similar to the Jeans-Teller-Landau approximation [12]. □

3.5 Geometric Properties of the Normal Form Truncation

If the vector fields $\boldsymbol{A}$ and $\boldsymbol{B}$ belong to a certain sub-algebra $\mathfrak{g}$ of the algebra of all vector fields on R^n, then the transforming vector fields $\boldsymbol{W}_i$ can be chosen to be in the same sub-algebra. This only requires that the vector fields $\boldsymbol{\Delta X}_i \in \mathfrak{g}$ are split such that $\boldsymbol{\Delta X}_i^n \in \mathfrak{g}$. Then the transformed vector fields satisfy $\boldsymbol{Y}_i \in \mathfrak{g}$ as well. This is particularly important for Hamiltonian vector fields. Since it often implies severe and important restrictions on the possible solution behavior of the normal form equations.

Example 3.1 (cont.) If the system (0.5)-(0.6) is Hamiltonian and satisfies a non-resonance condition (3.24), then all the transforming vector fields can be chosen to be Hamiltonian as well. This implies that the truncated normal form system

$$\begin{aligned} \frac{d}{dt}\bar{\boldsymbol{\phi}} &= \boldsymbol{\omega} + \epsilon\, \bar{\boldsymbol{f}}(\bar{\boldsymbol{I}};\epsilon)\,, \\ \frac{d}{dt}\bar{\boldsymbol{I}} &= \quad + \epsilon\, \bar{\boldsymbol{g}}(\bar{\boldsymbol{I}};\epsilon) \end{aligned}$$

is Hamiltonian. In particular, the functions $\bar{\boldsymbol{f}}$ and $\bar{\boldsymbol{g}}$ are the gradients of a Hamiltonian $\bar{h}$, i.e. $\bar{\boldsymbol{f}} = \boldsymbol{\nabla}_{\bar{\boldsymbol{I}}}\bar{h}$ and $\bar{\boldsymbol{g}} = \boldsymbol{\nabla}_{\bar{\boldsymbol{\phi}}}\bar{h}$. This and the fact that $\bar{h}$ would have to be 2π-periodic in the argument $\bar{\boldsymbol{phi}}$ immediately yield that $\bar{h}$ depends only on $\bar{\boldsymbol{I}}$ and

$$\bar{\boldsymbol{g}}(\bar{\boldsymbol{I}};\epsilon) \quad = \quad \boldsymbol{0}\,!$$

Thus each entry of the vector-valued variable $\boldsymbol{I}$ is an *adiabatic invariant* that is preserved over exponentially long periods of time up to small fluctuations. This problem has, for example, been discussed in detail by Perry & Wiggins in [91]. □

For Hamiltonian systems one could also transform the corresponding Hamiltonian instead of transforming the vector field [61],[7]. The crucial observation is that the Lie bracket of two Hamiltonian vector fields $\boldsymbol{X}_F$ and $\boldsymbol{X}_G$ with Hamiltonian F, G respectively, satisfies

$$[\boldsymbol{X}_F, \boldsymbol{X}_G] \quad = \quad \boldsymbol{X}_{\{F,G\}}\,,$$

where $\{G, F\}$ is the Lie-Poisson bracket of the two functions F and G. Thus, for a Hamiltonian

$$H(\boldsymbol{x};\epsilon) \quad = \quad \Omega(\boldsymbol{x}) + \epsilon\, h(\boldsymbol{x};\epsilon)$$

the normal form recursion is:

<u>HAMILTONIAN NORMAL FORM RECURSION</u>

Initial data:

$$S_0(\epsilon) \quad := \quad 0$$

For $i \geq 0$:

$$\begin{aligned}
\boldsymbol{W}_i(\epsilon) &:= \{\mathbf{id}, S_i(\epsilon)\}\,, & (5.33)\\
H_i(\epsilon) &:= H(\epsilon) \circ \boldsymbol{\Phi}_{1,\boldsymbol{W}_i(\epsilon)}\,, & (5.34)\\
\Delta h_{i+1} &:= \frac{1}{(i+1)!}\left[\frac{\partial^{i+1}}{\partial \epsilon^{i+1}} H_i\right]_{\epsilon=0}\,, & (5.35)\\
\Delta h_{i+1} &=: \Delta h^r_{i+1} + \Delta h^n_{i+1}\,, & (5.36)\\
0 &= \{\Omega, \Delta S_{i+1}\} + \Delta h^n_{i+1}\,, & (5.37)\\
S_{i+1}(\epsilon) &:= S_i(\epsilon) + \epsilon^{i+1}\,\Delta S_{i+1}\,. & (5.38)
\end{aligned}$$

The truncated normal form is then

$$\bar{H}_i(\bar{\boldsymbol{x}};\boldsymbol{\epsilon}) \quad = \quad \Omega(\bar{\boldsymbol{x}}) + \sum_{j=1}^{i} \epsilon^i\, \Delta h^r_i(\bar{\boldsymbol{x}})\,.$$

This or similar recursions are often used. However, the corresponding estimates for the truncation error $T_i = H_i - \bar{H}_i$ are slightly more difficult to prove. (Knowing an estimate for the Hamiltonian S_i, we need an additional estimate for the vector field $\boldsymbol{W}_i$. This can be obtained from Cauchy's estimate. See, for example, PÖSCHEL [93].) Otherwise the proof of Theorem 3.1 can also be applied to the Hamiltonian normal form recursion.

Another interesting class of problems with strong geometric properties is provided by reversible differential equations [86].

3.6 Applications

3.6.1 Linear Time-Varying Systems

We consider the system of linear time-varying differential equations

$$\frac{d}{d\tau}\tilde{\boldsymbol{u}} \quad = \quad \frac{\mathrm{i}}{\epsilon}\boldsymbol{S}(\tau)\tilde{\boldsymbol{u}}\,, \tag{6.39}$$

$\tilde{\boldsymbol{u}} \in \mathrm{C}^n$ and $\boldsymbol{S}(\tau)$ a real symmetric positive definite matrix. Such equations arise in the finite-dimensional truncation of the time-varying Schrödinger equation [74]. Let $\boldsymbol{\Psi}(\tau)$ denote the orthogonal matrix of eigenvectors of $\boldsymbol{S}(\tau)$ and assume that $\boldsymbol{S}$ and $\boldsymbol{\Psi}$ are analytic in τ. Then upon rescaling time by $1/\epsilon$, i.e. $t = \tau/\epsilon$, and using the linear time-varying coordinate transformation

$$\tilde{\boldsymbol{u}} \quad = \quad \boldsymbol{\Psi}(\tau)\boldsymbol{u}\,,$$

the system (6.39) is equivalent to the analytic system

$$\begin{aligned}\frac{d}{dt}\boldsymbol{u} &= \mathrm{i}\boldsymbol{D}(\tau)\boldsymbol{u} + \epsilon\, \boldsymbol{E}(\tau)\boldsymbol{u}\,,\\ \frac{d}{dt}\tau &= \epsilon\,,\end{aligned}$$

with $\boldsymbol{D} = \boldsymbol{\Psi}^{-1}\boldsymbol{S}\boldsymbol{\Psi}$ a diagonal matrix and $\boldsymbol{E} = -\boldsymbol{\Psi}^{-1}\partial_t\boldsymbol{\Psi}$ skew-symmetric. Thus we have obtained a system of type (0.1). In particular,

$$\boldsymbol{A}(\boldsymbol{x}) \;:=\; \begin{bmatrix} \mathrm{i}\,\boldsymbol{D}(\tau)\boldsymbol{u} \\ 0 \end{bmatrix}$$

and

$$\boldsymbol{B}(\boldsymbol{x}) \;:=\; \begin{bmatrix} \boldsymbol{E}(\tau)\boldsymbol{u} \\ 1 \end{bmatrix},$$

$\boldsymbol{x} = (\boldsymbol{u}^T, \tau)^T$. The homological equation

$$[\boldsymbol{A}, \boldsymbol{\Delta W}_1] + \boldsymbol{\Delta X}_1^n \;=\; 0\,, \quad \boldsymbol{\Delta X}_1^n(\tau) \;=\; \begin{bmatrix} \boldsymbol{E}(\tau)\boldsymbol{u} \\ 0 \end{bmatrix}$$

is solvable whenever the diagonal entries $d_{ii}(\tau)$ of $\boldsymbol{D}(\tau)$ satisfy $d_{ii}(\tau) \neq d_{jj}(\tau)$ for all $i \neq j$. Note that

$$\boldsymbol{\Delta X}_1^r \;:=\; \begin{bmatrix} \boldsymbol{0} \\ 1 \end{bmatrix}.$$

The corresponding transformed system is again linear and time-varying. Thus we can continue and solve in each step the homological (matrix) equation

$$[\boldsymbol{A}(\tau), \boldsymbol{\Delta W}_i(\tau)] + \boldsymbol{\Delta X}_i^n(\tau) \;=\; 0$$

with $\boldsymbol{\Delta X}_i^r$ the diagonal part and $\boldsymbol{\Delta X}_i^n$ the off-diagonal part of $\boldsymbol{\Delta X}_i$ corresponding to the matrix representation of $\boldsymbol{\Delta X}_i$. Let us look at this equation in more detail: First we write $\boldsymbol{A}$, $\boldsymbol{\Delta W}_i$ and $\boldsymbol{\Delta X}_i$ in matrix form, i.e.

$$\boldsymbol{A}(\boldsymbol{x}) \;=\; \mathrm{i}\boldsymbol{D}(\tau)\boldsymbol{u}\,, \quad \boldsymbol{\Delta W}_i \;=\; \boldsymbol{\Delta F}_i(\tau)\boldsymbol{u}\,, \quad \boldsymbol{\Delta X}_i \;=\; \boldsymbol{\Delta E}_i(\tau)\boldsymbol{u}\,.$$

Then the solution of the homological equation is given by

$$\Delta f^i_{kl}(\tau) \;=\; \frac{-\Delta e^i_{kl}(\tau)}{\mathrm{i}\,[d_{kk}(\tau) - d_{ll}(\tau)]}\,, \qquad k \neq l\,.$$

Here Δf^i_{kl} denotes the matrix entries of $\boldsymbol{\Delta F}_i$ etc. Let us now introduce an appropriate norm on $\mathcal{B}_r\mathcal{K}$, $\mathcal{K} = \boldsymbol{0} \times \mathcal{I}$, $\mathcal{I} \subset \mathrm{R}$: For

$$\boldsymbol{Y}(\boldsymbol{x}) \;=\; \boldsymbol{T}(\tau)\boldsymbol{u}\,,$$

we define

$$|\boldsymbol{Y}|_r \;=\; \sup_{\tau\in\mathcal{B}_r\mathcal{I}} \sum_{i,j} |t_{ij}(\tau)|\, r\,.$$

Let us assume that

$$|d_{ii}(\tau) - d_{jj}(\tau)| \;\geq\; \gamma\,, \quad i \neq j\,, \tau \in \mathcal{I}\,. \tag{6.40}$$

Then the solution $\boldsymbol{\Delta W}_i$ of the homological equation certainly satisfies the estimate

$$|\boldsymbol{\Delta W}_i|_r \;\leq\; \gamma^{-1}|\boldsymbol{\Delta X}_i^r|_r \;\leq\; \gamma^{-1}|\boldsymbol{\Delta X}_i|_r$$

and $||\boldsymbol{\Delta W}_i||_r \leq |\boldsymbol{\Delta W}_i|_r$. Thus, by virtue of Theorem 3.1, we know then that there exists a time-dependent linear coordinate transformation such that the problem (6.39) is equivalent to

$$\frac{d}{d\tau}\bar{\boldsymbol{u}} \;=\; \frac{\mathrm{i}}{\epsilon}\bar{\boldsymbol{D}}(\tau;\epsilon)\bar{\boldsymbol{u}}$$

up to terms exponentially small. Here $\bar{\boldsymbol{D}}(\tau;\epsilon)$ is a real diagonal matrix with $\bar{\boldsymbol{D}} = \boldsymbol{D} + \mathcal{O}(\epsilon^2)$. Thus, for ϵ small enough, the system (6.39) effectively decouples into n time-varying harmonic oscillators each of which gives rise to a first integral

$$\bar{J}_i \;=\; [\mathrm{Re}(\bar{u}_i)]^2 + [\mathrm{Im}(\bar{u}_i)]^2\,, \quad i = 1,\ldots,n\,.$$

The adiabatic invariance of the J_i's has already been discussed by Born & Fock [23]. For previous results on exponentially small transition rates see [66].

If the system (6.39) is discretized by a unitary integrator

$$\tilde{\boldsymbol{u}}_{n+1} \;=\; \boldsymbol{G}_{\Delta t}(\tau)\,\tilde{\boldsymbol{u}}_n\,,$$

$\boldsymbol{G}_{\Delta t}(\tau) \in \mathrm{C}^{n\times n}$ a unitary matrix, then backward error analysis implies that the numerical solutions are equivalent to the solutions of a perturbed problem of type (6.39) up to terms exponentially small. Applying normal form theory to this system, we obtain the adiabatic invariance of the corresponding J_i's over exponentially long periods of time. The specific arguments behind this result will be discussed in Section 3.7 in more detail.

In case of resonances, i.e., $d_{ii}(\tau) = d_{jj}(\tau)$ for some $i \neq j$, the normal form $\bar{\boldsymbol{D}}(\tau)$ contains off-diagonal elements that lead to an energy exchange between J_i and J_j. Still the normal form truncation is valid up to terms exponentially small in ϵ and the sum of J_i and J_j is an adiabatic invariant. The specific evolution of J_i and J_j under the constraint $J_i + J_j = const.$ depends now on the corresponding terms in the normal form.

Let us mention another interesting point: Assume that $\boldsymbol{D}(\tau)$ is 2π-periodic in τ. Then the normal form $\bar{\boldsymbol{D}}(\tau)$ is also 2π-periodic in τ. However the time evolution corresponding to $\boldsymbol{D}(\tau)$, to $\bar{\boldsymbol{D}}(\tau)$ respectively, will show a phase difference after one period $\tau = 2\pi$ (assuming identical phase at $\tau = 0$). This phase is called the Berry phase and has many interesting consequences in quantum mechanics [17],[33].

3.6.2 Non-Linear Systems With a Single Fast Degree

Systems

$$\begin{aligned} \frac{d}{dt}\phi &= \omega(\boldsymbol{I}) + \epsilon\, f(\phi, \boldsymbol{I})\,, \\ \frac{d}{dt}\boldsymbol{I} &= \epsilon\, \boldsymbol{g}(\phi, \boldsymbol{I}) \end{aligned}$$

with one fast degree $\phi \in \mathrm{T}$ and slow degrees $\boldsymbol{I} \in \mathrm{R}^d$ are among the best understood [5],[7]. This is because the corresponding homological equation

$$[\boldsymbol{A}, \boldsymbol{W}] + \boldsymbol{X}^n = 0$$

is alway solvable if $\omega(\boldsymbol{I}) \geq d_1 > 0$. Here the resonant part $\boldsymbol{X}^r$ of a function $\boldsymbol{X}$ is defined by

$$\boldsymbol{X}^r(\boldsymbol{I}) := \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{X}(\phi, \boldsymbol{I})\, d\phi\,.$$

Let us write the functions $\boldsymbol{X}^n$ and $\boldsymbol{W}$ in terms of a Fourier series, i.e.

$$\boldsymbol{X}^n(\phi, \boldsymbol{I}) = \begin{bmatrix} \sum_{k\neq 0} X^n_{1,k}(\boldsymbol{I})\, e^{\mathrm{i}k\phi} \\ \sum_{k\neq 0} \boldsymbol{X}^n_{2,k}(\boldsymbol{I})\, e^{\mathrm{i}k\phi} \end{bmatrix}$$

and

$$\boldsymbol{W}(\phi, \boldsymbol{I}) = \begin{bmatrix} \sum_{k\neq 0} W_{1,k}(\boldsymbol{I})\, e^{\mathrm{i}k\phi} \\ \sum_{k\neq 0} \boldsymbol{W}_{2,k}(\boldsymbol{I})\, e^{\mathrm{i}k\phi} \end{bmatrix}.$$

Thus the homological equation is equivalent to

$$\begin{aligned} \mathrm{i}\,\omega(\boldsymbol{I})\, k\, W_{1,k}(\boldsymbol{I}) - \left[\frac{\boldsymbol{\partial}}{\boldsymbol{\partial I}}\omega(\boldsymbol{I})\right] \cdot \boldsymbol{W}_{2,k}(\boldsymbol{I}) &= X^n_{1,k}(\boldsymbol{I})\,, \\ \mathrm{i}\,\omega(\boldsymbol{I})\, k\, \boldsymbol{W}_{2,k}(\boldsymbol{I}) &= \boldsymbol{X}^n_{2,k}(\boldsymbol{I})\,, \end{aligned}$$

$k \neq 0$. For each $k \neq 0$, this system is solvable for $W_{1,k}$ and $\boldsymbol{W}_{2,k}$ in terms of $X^n_{1,k}$ and $\boldsymbol{X}^n_{2,k}$. Furthermore, if $\boldsymbol{\partial_I}\omega$ is bounded from above by some constant d_2 and all the involved functions are real analytic, then we certainly obtain a bound

$$|\boldsymbol{W}|_r \leq \gamma^{-1}|\boldsymbol{X}^n|_r \leq \gamma^{-1}|\boldsymbol{X}|_r\,, \qquad \gamma^{-1} = (d_1)^{-1} + d_2(d_1)^{-2}\,,$$

with $|.|_r$ the exponentially weighted norm introduced in Example 3.1. Thus we can apply Theorem 3.1 which implies the existence of a coordinate transformation such that the transformed system is

$$\begin{aligned} \frac{d}{dt}\bar{\phi} &= \omega(\bar{\boldsymbol{I}}) + \epsilon\, \bar{f}(\bar{\boldsymbol{I}}; \epsilon)\,, \\ \frac{d}{dt}\bar{\boldsymbol{I}} &= \epsilon\, \bar{\boldsymbol{g}}(\bar{\boldsymbol{I}}; \epsilon) \end{aligned}$$

up to terms exponentially small, i.e., the truncation error satisfies

$$||\, \boldsymbol{T}(\bar{\phi}, \bar{\boldsymbol{I}})\, ||_\infty \leq \epsilon\, c_1\, e^{c_2/\epsilon}\,,$$

$c_1, c_2 > 0$ appropriate constants. An exponential estimate for this class of problems was first stated by Neishtadt in [88] by means of a different recursion.

Time-Varying Harmonic Oscillator

Let us consider the time-varying harmonic oscillator

$$\begin{aligned} \frac{d}{dt}q &= p\,, \\ \frac{d}{dt}p &= -\omega(\tau)^2 q\,, \\ \frac{d}{dt}\tau &= \epsilon \end{aligned}$$

with $\epsilon > 0$ a small parameter. We introduce action-angle variables (J, ϕ) [78],[6] via the generating function [78],[33]

$$S(q,\phi,\tau) \;:=\; \frac{1}{2}\,\omega(\tau)\,q^2\cot\phi\,.$$

Thus the new coordinates are implicitly defined by

$$\begin{aligned} p &= \frac{\partial S}{\partial q} = \omega(\tau)\,q\cot\phi\,, \\ J &= \frac{\partial S}{\partial \phi} = \frac{1}{2}\,\omega(\tau)\,q^2\,\frac{1}{\sin^2\phi}\,, \end{aligned}$$

or, solved for (q, p):

$$\begin{aligned} q &= \sqrt{\frac{2J}{\omega}}\,\sin\phi\,, \\ p &= \sqrt{2\omega J}\,\cos\phi\,. \end{aligned}$$

The corresponding transformed Hamiltonian is [33]

$$\begin{aligned} H &= \frac{1}{2}\left(p^2+\omega(\tau)^2q^2\right)+\frac{\partial S}{\partial t}(q,\phi,\tau)\,, \\ &= \omega(\tau)\,J+\frac{\epsilon}{2}\,\frac{\omega'(\tau)}{\omega(\tau)}\,J\,\sin(2\phi) \end{aligned}$$

and the equations of motion become

$$\begin{aligned} \frac{d}{dt}\phi &= \omega(\tau)+\frac{\epsilon}{2}\,\frac{\omega'(\tau)}{\omega(\tau)}\,\sin(2\phi)\,, \\ \frac{d}{dt}J &= -\,\epsilon\,\frac{\omega'(\tau)}{\omega(\tau)}J\cos(2\phi)\,, \\ \frac{d}{dt}\tau &= \epsilon\,. \end{aligned}$$

If ω is real analytic in $\tau \in \mathrm{R}$, $\omega(\tau) > c_1$, $|\omega'(\tau)| < c_2$, and both constants c_1, c_2 are of moderate size, then the system can be brought into normal form[2]

$$\begin{aligned}
\frac{d}{dt}\bar{\phi} &= \omega(\tau) + \epsilon\, \Delta\omega(\bar{J}, \tau, \epsilon)\,, \\
\frac{d}{dt}\bar{J} &= 0\,, \\
\frac{d}{dt}\tau &= \epsilon
\end{aligned}$$

up to terms exponentially small with respect to ϵ. Note that this result was used in Section 2.3.1 to show that a symplectic integrator will preserve the adiabatic invariant J over an exponentially long period of time. The trick is that the result holds for all Hamiltonian functions of type

$$H = \omega(\tau)\, J + \epsilon\, h(J, \phi, \tau, \epsilon)\,,$$

h an arbitrary (bounded, real analytic) function 2π-periodic in ϕ.

Let us mention another interesting point: Assume that we have a system of two perturbed time-varying oscillators described by the Hamiltonian

$$H(\boldsymbol{J}, \boldsymbol{\phi}, \tau, \epsilon) := \boldsymbol{\omega}(\tau)^T \boldsymbol{J} + \epsilon\, h(\boldsymbol{J}, \boldsymbol{\phi}, \tau, \epsilon)$$

with the vector $\boldsymbol{\omega}(\tau) \in \mathrm{R}^2$ of frequencies satisfying

$$\omega_1(\tau) + \omega_2(\tau) > 0 \qquad \text{for all}\, \tau\,. \tag{6.41}$$

Due to resonances, the individual action variables J_i will not be adiabatic invariants. However, an appropriate linear combination of the action variables J_i will be preserved while passing through a resonance. For example, let as assume that $\omega_1(\tau_0) = \omega_2(\tau_0)$ at some time τ_0. Then, upon introducing new angles

$$\begin{aligned}
\psi_1 &= \frac{1}{2}(\phi_1 + \phi_2)\,, \\
\psi_2 &= \phi_1 - \phi_2
\end{aligned}$$

with corresponding action variables

$$\begin{aligned}
I_1 &= \frac{1}{2}(J_1 + J_2)\,, \\
I_2 &= J_1 - J_2\,,
\end{aligned}$$

we can treat the resulting system with Hamiltonian

$$H = \frac{1}{2}(\omega_1(\tau) + \omega_2(\tau))I_1 + (\omega_1(\tau) - \omega_2(\tau))I_2 + \epsilon\, \tilde{h}(I_1, I_2, \psi_1, \psi_2, \tau, \epsilon)$$

[2] In each step, the normal form transformation is constructed by formally fixing τ and taking J as the slow variable $\boldsymbol{I}$.

as a single frequency system in the variable ψ_1 as long as $\omega_1(\tau) - \omega_2(\tau)$ stays small enough (for example, $|\omega_1(\tau) - \omega_2(\tau)| \leq \sqrt{\epsilon}$). Thus we perform a normal form recursion where we only averages over the variable ψ_1. For that reason, we use the Fourier series expansion

$$h(\boldsymbol{J}, \boldsymbol{\phi}, \tau, \epsilon) \;=\; \sum_{\boldsymbol{k} \in \mathbb{Z}^d} h_{\boldsymbol{k}}(\boldsymbol{J}, \tau, \epsilon)\, e^{\mathrm{i}\boldsymbol{k}^T \boldsymbol{\phi}}$$

of h. This immediately yields the corresponding expansion in ψ_1. The corresponding homological equation is solvable due to (6.41) and the variable conjugate to ψ_1, i.e. I_1, is an adiabatic invariant up to terms exponentially small in $\sqrt{\epsilon}$. □

Hamiltonian Systems with a Stiff Spring

Let us consider a diatomic molecule where two atoms with unit mass are "bonded" by a stiff spring and move under an external conservative (real analytic) force field. In terms of external coordinates $(\boldsymbol{Q}, \boldsymbol{P}) \in \mathrm{R}^4$ and internal coordinates $(r, p_r) \in \mathrm{R}^2$ ($r > 0$ the distance between the two atoms) the corresponding Hamiltonian function is

$$H \;=\; \frac{p_r^2 + \epsilon^{-2}(r - r_0)^2}{2} + h(\boldsymbol{Q}, \boldsymbol{P}) + (r - r_0)\, f(r - r_0, p_r, \boldsymbol{P}, \boldsymbol{Q})$$

with h and f real analytic functions and ϵ^{-2} the spring constant of the "chemical bond". The equations of motion are then derived from H and the Lie-Poisson bracket

$$\{F, G\} \;=\; \{F, G\}_{r,p_r} + \{F, G\}_{\boldsymbol{Q},\boldsymbol{P}}\,, \tag{6.42}$$

where $\{F, G\}_{r,p_r}$ and $\{F, G\}_{\boldsymbol{Q},\boldsymbol{P}}$ are the standard canonical brackets. Upon introducing new canonical variables

$$\begin{aligned} R &:= \epsilon^{-1/2}(r - r_0)\,, \\ P_R &= \epsilon^{1/2} p_r\,, \end{aligned}$$

the Hamiltonian becomes

$$H \;=\; \frac{1}{2\epsilon}(P_R^2 + R^2) + h(\boldsymbol{Q}, \boldsymbol{P}) + \epsilon^{1/2}\, R\, f(\epsilon^{1/2} R, \epsilon^{-1/2} P_R, \boldsymbol{Q}, \boldsymbol{P})\,.$$

Next we introduce action-angle variables $(J, \phi) \in \mathrm{R}^+ \times \mathrm{T}$ by means of

$$\begin{aligned} R &= \sqrt{2\,\epsilon\, J}\, \cos\phi\,, \\ P_R &= \sqrt{2\,\epsilon\, J}\, \sin\phi\,. \end{aligned}$$

This defines a symplectic coordinate transformation from the variables $(R, P_R, \boldsymbol{Q}, \boldsymbol{P})$ and the canonical Poisson bracket (6.42) to the new variables $(J, \phi, \boldsymbol{Q}, \boldsymbol{P})$ and the scaled Lie-Poisson bracket

$$\{F, G\}_s \;:=\; \epsilon^{-1}\{F, G\}_{\phi,J} + \{F, G\}_{\boldsymbol{Q},\boldsymbol{P}}\,. \tag{6.43}$$

The transformed Hamiltonian is

$$H = J + h(\boldsymbol{Q},\boldsymbol{P}) + \epsilon\,\tilde{f}(\phi,J,\boldsymbol{Q},\boldsymbol{P},\epsilon)\,.$$

Upon rescaling time by ϵ, i.e. $t = \epsilon\tau$, the resulting equations of motion are

$$\begin{aligned}
\frac{d}{d\tau}\phi &= 1 + \ \epsilon\,\nabla_J \tilde{f}(\phi,J,\boldsymbol{Q},\boldsymbol{P},\epsilon)\,,\\
\frac{d}{d\tau}J &= -\epsilon\,\nabla_\phi \tilde{f}(\phi,J,\boldsymbol{Q},\boldsymbol{P},\epsilon)\,,\\
\frac{d}{d\tau}\boldsymbol{Q} &= \epsilon\,\boldsymbol{\nabla}_{\boldsymbol{P}} h(\boldsymbol{Q},\boldsymbol{P}) + \epsilon^2\,\boldsymbol{\nabla}_{\boldsymbol{P}} \tilde{f}(\phi,J,\boldsymbol{Q},\boldsymbol{P},\epsilon)\,,\\
\frac{d}{d\tau}\boldsymbol{P} &= -\epsilon\,\boldsymbol{\nabla}_{\boldsymbol{Q}} h(\boldsymbol{Q},\boldsymbol{P}) - \epsilon^2\,\boldsymbol{\nabla}_{\boldsymbol{Q}} \tilde{f}(\phi,J,\boldsymbol{Q},\boldsymbol{P},\epsilon)\,.
\end{aligned}$$

These equations are certainly amenable to Theorem 3.1 and the previous remarks in this subsection (with $\boldsymbol{I} = (J,\boldsymbol{Q}^T,\boldsymbol{P}^T)^T$). We only have to be careful to not include $J = 0$ into the domain $\mathcal{B}_R\mathcal{K}$. Since the function $\tilde{f}$ will, in general, be not analytic at $J = 0$. Thus, for $J > 0$, there exists a symplectic coordinate transformation such that the transformed Hamiltonian system is

$$\begin{aligned}
\frac{d}{d\tau}\bar{\phi} &= 1 + \ \epsilon\,\nabla_{\bar{J}} \bar{f}(\bar{J},\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}},\epsilon)\,,\\
\frac{d}{d\tau}\bar{J} &= -\epsilon\,\nabla_{\bar{\phi}} \bar{f}(\bar{J},\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}},\epsilon) \ = \ 0\,,\\
\frac{d}{d\tau}\bar{\boldsymbol{Q}} &= \epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{P}}} h(\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}}) + \epsilon^2\,\boldsymbol{\nabla}_{\bar{\boldsymbol{P}}} \bar{f}(\bar{J},\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}},\epsilon)\,,\\
\frac{d}{d\tau}\bar{\boldsymbol{P}} &= -\epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} h(\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}}) - \epsilon^2\,\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} \bar{f}(\bar{J},\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}},\epsilon)
\end{aligned}$$

up to terms exponentially small in ϵ.

Remark 3.3. A comment is necessary at this point. We can associate a sequence of Hamiltonian functions S_i with the sequence of transforming vector fields $\boldsymbol{W}_i$. However, because of the non-trivial Poisson structure (6.43), the standard relation $\boldsymbol{W}_i = \{\mathbf{id}, S_i\}_s$ has to be modified. This modification is discussed in the Appendix. □

The action J is an adiabatic invariant and, therefore, the system can effectively be reduced to its motion in the external degrees of freedom $(\boldsymbol{Q},\boldsymbol{P})$. Note that J corresponds to the energy in the fast bond stretching motion. Furthermore, upon neglecting terms of order ϵ^2, the stiff spring can be replaced by a rigid rod, i.e., with $t = \epsilon\tau$, the reduced equations of motion are

$$\begin{aligned}
r &= r_0\,,\\
p_r &= 0\\
\frac{d}{dt}\bar{\boldsymbol{Q}} &= +\boldsymbol{\nabla}_{\bar{\boldsymbol{P}}} h(\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}})\,,\\
\frac{d}{dt}\bar{\boldsymbol{P}} &= -\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} h(\bar{\boldsymbol{Q}},\bar{\boldsymbol{P}})\,.
\end{aligned}$$

In Cartesian coordinates $(\boldsymbol{q},\boldsymbol{p}) \in \mathrm{R}^6 \times \mathrm{R}^6$ the motion of the diatomic molecule is given by

$$
\begin{aligned}
\frac{d}{dt}\boldsymbol{q} &= \boldsymbol{p}\,, \\
\frac{d}{dt}\boldsymbol{p} &= -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}) - \epsilon^{-2}(r(\boldsymbol{q}) - r_0)\boldsymbol{\nabla}_{\boldsymbol{q}}\, r(\boldsymbol{q})\,.
\end{aligned}
$$

Replacing the stiff spring by a rigid rod results in the constrained system

$$
\begin{aligned}
\frac{d}{dt}\boldsymbol{q} &= \boldsymbol{p}\,, \\
\frac{d}{dt}\boldsymbol{p} &= -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}) - \boldsymbol{\nabla}_{\boldsymbol{q}}\, r(\boldsymbol{q})\,\lambda\,, \\
0 &= r(\boldsymbol{q}) - r_0\,.
\end{aligned}
$$

Exponential estimates where first derived by Benettin, Galgani & Giorgilli [14]. They derive a normal form for the Hamiltonian in a different set of coordinates without using a non-canonical Poisson bracket. This implies that they could not apply Neishtadt's proof for systems with rapidly rotating phase [88]. More recently, Benettin, Carati & Gallavotti [12] have given a rigorous justification for the so called Jeans-Landau-Teller approximation for adiabatic invariants [75]. This approach also yields exponentially small estimates but the derivation is different from the Nekhoroshev-type approaches used in this chapter. The Jeans-Landau-Teller approximation is more intuitive and yields sharper estimates [12]. However, so far, only special cases have been treated.

Again it can be shown by means of backward error analysis that a symplectic discretization of the equations of motion will result in an adiabatic invariance of the energy in the fast degree of motion over exponentially long periods of time. See Section 3.7 for details. □

The Rubin & Ungar Problem

In the previous subsection, we considered a mechanical system with one fast degree of motion which, in appropriate coordinates, reduces to a harmonic oscillator with constant frequency. Here we look at a Hamiltonian system with a fast degree of motion whose frequency depends on the slowly changing solution components, i.e.

$$
H = \frac{\omega^2(\tilde{\boldsymbol{Q}})\, p_r^2 + \epsilon^{-2}(r - r_0)^2}{2} + h(\tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{P}}) + (r - r_0)\, f(r - r_0, p_r, \tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}})
$$

with ω, h, and f real analytic functions and ϵ^{-2} the force constant of the fast oscillator. The equations of motion are then derived from H and the Lie-Poisson bracket

$$
\{F,G\} = \{F,G\}_{r,p_r} + \{F,G\}_{\tilde{\boldsymbol{Q}},\tilde{\boldsymbol{P}}}\,,
$$

where $\{F,G\}_{r,p_r}$ and $\{F,G\}_{\tilde{\boldsymbol{Q}},\tilde{\boldsymbol{P}}}$ are the standard canonical brackets. This problem and its reduced dynamics in the limit $\epsilon \to 0$ was first discussed by Rubin & Ungar

[105]. For a more recent account see Bornemann & Schütte [24]. Upon introducing the canonical variables

$$\begin{aligned} R &:= \epsilon^{-1/2}(r - r_0)\,, \\ P_R &= \epsilon^{1/2} p_r\,, \end{aligned}$$

the Hamiltonian becomes

$$H = \frac{1}{2\epsilon}(\omega^2(\tilde{\boldsymbol{Q}})\, P_R^2 + R^2) + h(\tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{P}}) + \epsilon^{1/2}\, R\, f(\epsilon^{1/2} R, \epsilon^{-1/2} P_R, \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{P}})\,.$$

Next we introduce another set of variables $(\hat{J}, \phi, \boldsymbol{Q}, \boldsymbol{P}) \in \mathrm{R}^+ \times \mathrm{T} \times \mathrm{R}^m \times \mathrm{R}^m$ by means of the generating function

$$S = \tilde{\boldsymbol{Q}}^T \boldsymbol{P} + \frac{\omega(\tilde{\boldsymbol{Q}})}{2}\, P_R^2 \cot\phi\,.$$

From this generating function, we obtain the equations [6]

$$\begin{aligned} R &= \frac{\partial S}{\partial P_R} = \omega(\tilde{\boldsymbol{Q}})\, P_R \cot\phi\,, \\ \hat{J} &= \frac{\partial S}{\partial \phi} = \frac{\omega(\tilde{\boldsymbol{Q}})}{2}\, P_R^2 \sin^{-2}\phi\,, \\ \boldsymbol{Q} &= \frac{\boldsymbol{\partial} S}{\boldsymbol{\partial P}} = \tilde{\boldsymbol{Q}}\,, \\ \tilde{\boldsymbol{P}} &= \frac{\boldsymbol{\partial} S}{\boldsymbol{\partial} \tilde{\boldsymbol{Q}}} = \boldsymbol{P} + \boldsymbol{\nabla}_{\tilde{\boldsymbol{Q}}} \frac{\omega(\tilde{\boldsymbol{Q}})}{2}\, P_R^2 \cot\phi\,. \end{aligned}$$

Using the manipulations described in [33] (page 54ff.) and $\hat{J} = \epsilon J$, this is equivalent to

$$\begin{aligned} R &= \sqrt{2\,\epsilon\,\omega(\boldsymbol{Q})\,J}\,\cos\phi\,, \\ P_R &= \sqrt{\frac{2\,\epsilon\,J}{\omega(\boldsymbol{Q})}}\,\sin\phi\,, \\ \tilde{\boldsymbol{Q}} &= \boldsymbol{Q}\,, \\ \tilde{\boldsymbol{P}} &= \boldsymbol{P} + \frac{\epsilon}{2}\,\boldsymbol{\nabla}_{\boldsymbol{Q}} \ln[\omega(\boldsymbol{Q})]\, J\, \sin(2\phi)\,. \end{aligned}$$

The transformation is symplectic from the variables $(R, P_R, \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{P}})$ and the canonical Lie-Poisson bracket to the new variables $(\phi, J, \boldsymbol{Q}, \boldsymbol{P})$ and the scaled Lie-Poisson bracket

$$\{F, G\}_s := \epsilon^{-1}\{F, G\}_{\phi,J} + \{F, G\}_{\boldsymbol{Q},\boldsymbol{P}}\,.$$

The new transformed Hamiltonian is

$$H = \omega(\boldsymbol{Q})\, J + h(\boldsymbol{Q}, \boldsymbol{P}) + \epsilon\, \tilde{f}(\phi, J, \boldsymbol{Q}, \boldsymbol{P}, \epsilon)$$

with $\tilde{f}$ appropriately defined. Upon rescaling time by ϵ, i.e. $t = \epsilon\tau$, the resulting equations of motion become

$$\begin{aligned}
\frac{d}{d\tau}\phi &= \omega(\boldsymbol{Q}) + \epsilon\nabla_J \tilde{f}(\phi, J, \boldsymbol{Q}, \boldsymbol{P}, \epsilon)\,, \\
\frac{d}{d\tau}J &= \qquad - \epsilon\nabla_\phi \tilde{f}(\phi, J, \boldsymbol{Q}, \boldsymbol{P}, \epsilon)\,, \\
\frac{d}{d\tau}\boldsymbol{Q} &= \qquad + \epsilon\,\boldsymbol{\nabla}_{\boldsymbol{P}} h(\boldsymbol{Q}, \boldsymbol{P}) + \epsilon^2\,\boldsymbol{\nabla}_{\boldsymbol{P}} \tilde{f}(\phi, J, \boldsymbol{Q}, \boldsymbol{P}, \epsilon)\,, \\
\frac{d}{d\tau}\boldsymbol{P} &= \qquad - \epsilon\,\boldsymbol{\nabla}_{\boldsymbol{Q}} h(\boldsymbol{Q}, \boldsymbol{P}) - \epsilon^2\,\boldsymbol{\nabla}_{\boldsymbol{Q}} \tilde{f}(\phi, J, \boldsymbol{Q}, \boldsymbol{P}, \epsilon) - \epsilon\,\boldsymbol{\nabla}_{\boldsymbol{Q}}\omega(\boldsymbol{Q})\,J\,.
\end{aligned}$$

These equations are certainly amenable to Theorem 3.1 (with $\boldsymbol{I} = (J, \boldsymbol{Q}^T, \boldsymbol{P}^T)^T$) and the comments made in the Appendix. Thus, for $J > 0$, there exists a symplectic coordinate transformation such that the transformed Hamiltonian system is

$$\begin{aligned}
\frac{d}{d\tau}\bar{\phi} &= \omega(\bar{\boldsymbol{Q}}) + \epsilon\nabla_{\bar{J}} \bar{f}(\bar{J}, \bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}}, \epsilon)\,, \\
\frac{d}{d\tau}\bar{J} &= \qquad - \epsilon\nabla_{\bar{\phi}} \bar{f}(\bar{J}, \bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}}, \epsilon) \;=\; 0\,, \\
\frac{d}{d\tau}\bar{\boldsymbol{Q}} &= \qquad + \epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{P}}} h(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}}) + \epsilon^2\,\boldsymbol{\nabla}_{\bar{\boldsymbol{P}}} \bar{f}(\bar{J}, \bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}}, \epsilon)\,, \\
\frac{d}{d\tau}\bar{\boldsymbol{P}} &= \qquad - \epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} h(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}}) - \epsilon^2\,\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} \bar{f}(\bar{J}, \bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}}, \epsilon) - \epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}}\,\omega(\bar{\boldsymbol{Q}})\,\bar{J}
\end{aligned}$$

up to terms exponentially small in ϵ. Note that the action $\bar{J}$ is a first integral and that, therefore, the system can effectively be reduced to its motion in the external degrees of freedom $(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}})$. Furthermore, upon neglecting terms of order ϵ^2, the motion in the slow variable $(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}})$ is given by

$$\begin{aligned}
\frac{d}{d\tau}\bar{\boldsymbol{Q}} &= +\epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{P}}} h(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}})\,, \\
\frac{d}{d\tau}\bar{\boldsymbol{P}} &= -\epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} h(\bar{\boldsymbol{Q}}, \bar{\boldsymbol{P}}) - \epsilon\,\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}}\,\omega(\bar{\boldsymbol{Q}})\,\bar{J}\,.
\end{aligned}$$

Note the additional force term $\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}}\omega(\bar{\boldsymbol{Q}})\,\bar{J}$. This term was first derived by Rubin & Ungar [105]. More recently, the same term was derived by Bornemann & Schütte [24] using homogenization techniques. However, neither Rubin & Ungar nor Bornemann & Schütte gave exponentially small estimates for the remainder in the normal form truncation which implies the preservation of the adiabatic invariant $\bar{J}$ over exponentially long periods of time. In [100], we have discussed the influence of a heat-bath (Langevin dynamics) on the slow dynamics. It turns out that, in the limit $\epsilon \to 0$ and for sufficiently strong coupling to the heat bath, a different correction term appears (the so called Fixman potential [39],[100]). See also the paper by Helfand [59].

3.7 Numerical Conservation of Adiabatic Invariants

In this section, we discuss the effect of a symplectic discretization on a system of type

$$\frac{d}{dt}\boldsymbol{x} \quad = \quad \boldsymbol{A}(\boldsymbol{x}) + \epsilon \boldsymbol{B}(\boldsymbol{x};\epsilon) \tag{7.44}$$

where both $\boldsymbol{A}$ and $\boldsymbol{B}$ are real analytic Hamiltonian vector fields on R^{2n}, $n \geq 1$. We give a detailed treatment for the case satisfying the assumptions below. But the described approach can certainly be applied to more general problems which can be treated by Theorem 3.1 or Proposition 3.1, respectively, and which lead to adiabatic invariants that are conserved over exponentially long periods of time.

Assumptions. We assume (i) that $\boldsymbol{A}$ possesses periodic solutions with action variable J (the area[3] enclosed by the periodic motion) and (ii) that (7.44) satisfies the conditions of Theorem 3.1 for some $R > 0$, $\gamma > 0$, and some compact set $\mathcal{K}$. □

Let us discretize the system (7.44) by a symplectic integrator of order $p \geq 1$, i.e.

$$\begin{aligned} \boldsymbol{x}_{n+1} &= \boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}_n)\,, \\ t_{n+1} &= t_n + \Delta t\,. \end{aligned} \tag{7.45}$$

Backward error analysis implies that there exists a modified Hamiltonian differential equation

$$\frac{d}{dt}\boldsymbol{x} \quad = \quad \tilde{\boldsymbol{X}}(\boldsymbol{x},\Delta t)$$

such that its time-Δt-flow map is exponentially close to the discrete time map $\boldsymbol{\Psi}_{\Delta t}$. More specifically, let us introduce the vector field

$$\boldsymbol{Y}(\epsilon) \quad = \quad \boldsymbol{A} + \epsilon \boldsymbol{B}(\epsilon)$$

By our assumptions, $\boldsymbol{Y}$ is real analytic on $\mathcal{B}_R\mathcal{K}$ (using the notations of Section 3.3) and an estimate

$$||\,\boldsymbol{Y}(\epsilon)\,||_R \quad \leq \quad M\,, \qquad M\,>\,0\,,$$

holds for all $\epsilon > 0$ small enough. Then, according to Theorem 2.1, there exist constants $c_i > 0$, $i = 1,2,3,4$, such that

$$||\,\boldsymbol{\Psi}_{\Delta t} - \boldsymbol{\Phi}_{\Delta t,\tilde{\boldsymbol{X}}}\,||_{R/2} \quad \leq \quad c_1 M \Delta t\, e^{-p}\, e^{-c_2/(M\Delta t)} \tag{7.46}$$

and

$$||\,\tilde{\boldsymbol{X}}(\Delta t) - \boldsymbol{Y}\,||_{R/2} \quad \leq \quad c_3\, M\, (c_4 M \Delta t)^p\,.$$

Next we define $\epsilon\tilde{\boldsymbol{B}} := \tilde{\boldsymbol{X}} - \boldsymbol{A}$ or, in other words,

$$\tilde{\boldsymbol{X}}(\epsilon,\Delta t) \quad := \quad \boldsymbol{A} + \epsilon\tilde{\boldsymbol{B}}(\epsilon,\Delta t)\,. \tag{7.47}$$

[3]The periodic solutions are assumed, for simplicity, to lie on a plane.

We conclude that

$$||\tilde{\boldsymbol{B}}||_{R/2} \leq 1 + \epsilon^{-1} c_3 M \,(c_4 M \Delta t)^p$$

and apply Theorem 3.1 to the modified vector field (7.47). We only have to replace ϵ by

$$\tilde{\epsilon} \quad := \quad \epsilon + c_3 M \,(c_4 M \Delta t)^p \tag{7.48}$$

and R by $R/2$. Our assumptions and Theorem 3.1 imply that there exists a function $\hat{J}(\boldsymbol{x})$ which is $\tilde{\epsilon}$-close to the action variable J and is a first integral of the vector field $\tilde{\boldsymbol{X}}(\Delta t)$ up to terms exponentially small in $\tilde{\epsilon}$. Thus there exist constants $c_5, c_6 > 0$ such that

$$|\,\hat{J}(\boldsymbol{x}) - \hat{J}(\boldsymbol{\Phi}_{\Delta t, \tilde{\boldsymbol{X}}}(\boldsymbol{x}))\,| \quad \leq \quad c_5\, \tilde{\epsilon}\, \Delta t\, e^{-c_6/\tilde{\epsilon}} \tag{7.49}$$

for all $\boldsymbol{x} \in \mathcal{K}$.

Theorem 3.2. Let us assume (i) that the Assumptions hold, (ii) that the equations are discretized by a symplectic method of order $p \geq 1$, and (iii) that the numerically computed solutions stay in the compact subset $\mathcal{K}$ of phase space. Then there exist the above introduced constants c_i, $i = 1, \ldots, 6$, and a transformed action $\hat{J}$ which is $\tilde{\epsilon}$-close to the action J of the vector field $\boldsymbol{A}$ such that the numerical solution $\boldsymbol{x}_n$ after n integration steps satisfies

$$|\,\hat{J}(\boldsymbol{x}_n) - \hat{J}(\boldsymbol{x}_0)\,| \leq n\, \Delta t\, \left[\lambda\, c_1\, M\, e^{-p}\, e^{-c_2/(M\Delta t)} + c_5\, \tilde{\epsilon}\, e^{-c_6/\tilde{\epsilon}}\right]$$

with $\tilde{\epsilon}$ defined by (7.48) and $\lambda > 0$ the Lipschitz constant of $\hat{J}$ on $\mathcal{K}$. This implies

$$|\,J(\boldsymbol{x}_n) - J(\boldsymbol{x}_0)\,| \quad \leq \quad n\, \Delta t\, \left[\lambda\, c_1\, M\, e^{-p}\, e^{-c_2/(M\Delta t)} + c_5\, \tilde{\epsilon}\, e^{-c_6/\tilde{\epsilon}}\right] + c_7\, \tilde{\epsilon}, \tag{7.50}$$

$c_7 > 0$ an appropriate constant such that

$$|\,\hat{J}(\boldsymbol{x}) - J(\boldsymbol{x})\,| \leq c_7\, \tilde{\epsilon}$$

on $\mathcal{K}$. □

Proof. We have already shown (7.46) and (7.49). Thus

$$\begin{aligned}
|\hat{J}(\boldsymbol{x}_n) - \hat{J}(\boldsymbol{x}_0)| \quad &\leq \quad \sum_{j=0}^{n-1} |\hat{J}(\boldsymbol{x}_{j+1}) - \hat{J}(\boldsymbol{x}_j)|\,, \\
&\leq \quad \sum_{j=0}^{n-1} |\hat{J}(\boldsymbol{\Psi}_{\Delta t}(\boldsymbol{x}_j)) - \hat{J}(\boldsymbol{\Phi}_{\Delta t, \tilde{\boldsymbol{X}}}(\boldsymbol{x}_j)) + \hat{J}(\boldsymbol{\Phi}_{\Delta t, \tilde{\boldsymbol{X}}}(\boldsymbol{x}_j)) - \hat{J}(\boldsymbol{x}_j)| \\
&\leq \quad n\, \left[\lambda\, c_1\, M\, \Delta t\, e^{-p}\, e^{-c_2/(M\Delta t)} + c_5\, \Delta t\, \tilde{\epsilon}\, e^{-c_6/\tilde{\epsilon}}\right]\,.
\end{aligned}$$

□

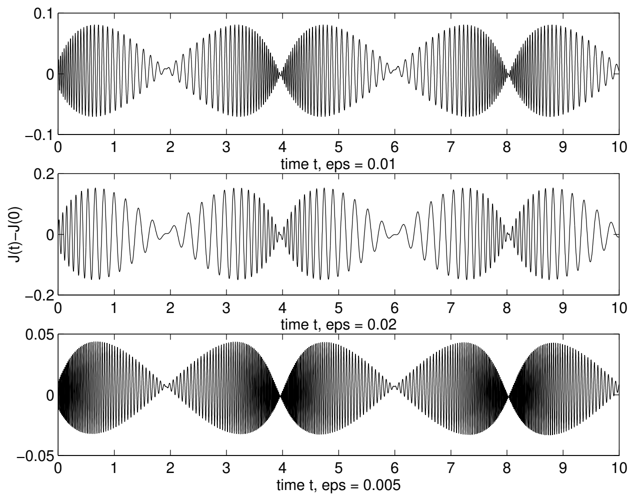


Figure 3.1: Time evolution of the adiabatic invariant J for three different values of ϵ and step-size $\Delta t = \epsilon/10$.

We conclude from Theorem 3.2 that, for sufficiently small step-sizes Δt, $\epsilon \approx \tilde{\epsilon}$ and the adiabatic invariant J is conserved under symplectic discretization in the same manner as it is for the analytic solutions. See also [102].

Example 3.2. Let us consider the stiff "reversed" pendulum [25]

$$\frac{d}{dt}\boldsymbol{q} \quad = \quad \boldsymbol{p}\,, \tag{7.51}$$

$$\frac{d}{dt}\boldsymbol{p} \quad = \quad -\epsilon^{-2}(\phi(\boldsymbol{q}) - \phi_0)\boldsymbol{\nabla}_{\boldsymbol{q}}\,\phi(\boldsymbol{q}) - (r(\boldsymbol{q}) - r_0)\boldsymbol{\nabla}_{\boldsymbol{q}}\,r(\boldsymbol{q}) \tag{7.52}$$

where $\boldsymbol{q}, \boldsymbol{p} \in \mathrm{R}^2$, $r(\boldsymbol{q}) = |\boldsymbol{q}| = \sqrt{q_1^2 + q_2^2}$ and $\phi(\boldsymbol{q}) = \mathrm{acos}\,(q_1/|\boldsymbol{q}|)$. The Hamiltonian is

$$H(\boldsymbol{q}, \boldsymbol{p}) \quad = \quad \frac{1}{2}[\boldsymbol{p}^T\boldsymbol{p} + \epsilon^{-2}(\phi(\boldsymbol{q}) - \phi_0)^2 + (r(\boldsymbol{q}) - r_0)^2]$$

The equations of motion (7.51)-(7.52) can be transformed to (r, ϕ)-coordinates by introducing corresponding conjugate momenta $p_r, p_\phi \in \mathrm{R}$. It can be directly verified that

$$\boldsymbol{\nabla}_{\boldsymbol{q}}\,r(\boldsymbol{q}) = \frac{1}{r}\boldsymbol{q} =: \boldsymbol{G}^T, \qquad \boldsymbol{\nabla}_{\boldsymbol{q}}\,\phi(\boldsymbol{q}) = \frac{1}{r^2}(-q_2, q_1)^T =: \boldsymbol{B}^T$$

Hence

$$\boldsymbol{p} = \boldsymbol{G}^T p_r + \boldsymbol{B}^T p_\phi = p_r \boldsymbol{\nabla}_{\boldsymbol{q}}\,r + p_\phi \boldsymbol{\nabla}_{\boldsymbol{q}}\,\phi\,.$$

(Note that $\boldsymbol{B}\boldsymbol{G}^T = 0$.) The transformed Hamiltonian is

$$H(\boldsymbol{p}, \boldsymbol{q}) = E(r, \phi, p_r, p_\phi) = \frac{1}{2}[p_r^2 + r^{-2}p_\phi^2 + \epsilon^{-2}(\phi - \phi_0)^2 + (r - r_0)^2]$$

and the transformed equations of motion are

$$\begin{aligned}
\frac{d}{dt}r &= p_r\,, \\
\frac{d}{dt}p_r &= -(r-r_0) + p_\phi^2 r^{-3}\,, \\
\frac{d}{dt}\phi &= r^{-2}p_\phi\,, \\
\frac{dt}{dt}p_\phi &= -\epsilon^{-2}(\phi-\phi_0)\,.
\end{aligned}$$

Note that this system is of the type considered in Section 3.6.2. Here (ϕ, p_ϕ) takes the role of (r, p_r) and $(\tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{P}})$ is given by (r, p_r)! (That's why we call it the reversed pendulum.) Thus, in (r, ϕ)-coordinates the fast and slow motion can be separated. Here we are interested in the Cartesian formulation of the problem. First note that, in the limit $\epsilon \to 0$, the dynamics of (7.51)-(7.52) does not reduce to the constrained system

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{q} &= \boldsymbol{p}\,, \\
\frac{d}{dt}\boldsymbol{p} &= -(r(\boldsymbol{q})-r_0)\boldsymbol{\nabla}_{\boldsymbol{q}}\, r(\boldsymbol{q}) - \boldsymbol{\nabla}_{\boldsymbol{q}}\, \phi(\boldsymbol{q})\,\lambda\,, \\
0 &= \phi(\boldsymbol{q}) - \phi_0
\end{aligned}$$

unless the energy in the (fast) (ϕ, p_ϕ) degree of motion is zero. Instead an additional force term has to be added and the dynamics reduces to the modified constrained system [105],[25]

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{q} &= \boldsymbol{p}\,, \\
\frac{d}{dt}\boldsymbol{p} &= -(r(\boldsymbol{q})-r_0)\boldsymbol{\nabla}_{\boldsymbol{q}}\, r(\boldsymbol{q}) - c\,\boldsymbol{\nabla}_{\boldsymbol{q}}\, r(\boldsymbol{q})^{-1} - \boldsymbol{\nabla}_{\boldsymbol{q}}\, \phi(\boldsymbol{q})\,\lambda\,, \\
0 &= \phi(\boldsymbol{q}) - \phi_0\,.
\end{aligned}$$

The constant $c \geq 0$ depends on the initial energy and frequency in the (ϕ, p_ϕ) degree of motion and is related to the existence of an adiabatic invariant for the fast (ϕ, p_ϕ) degree of motion. This adiabatic invariant is given by

$$J \;=\; r\left[\frac{\boldsymbol{p}^T\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{p}}{2\,r^2} + \frac{1}{2\epsilon^2}(\phi-\phi_0)^2\right], \qquad \mathbf{V} := (-q_2, q_1)\,,$$

and $c = J(0)$. The adiabatic invariant is correctly reproduced by the symplectic Verlet method [120] as shown in Fig. 3.1 provided the step-size Δt is chosen sufficiently small. For a more detailed numerical study on a related model problem, see [102]. □

3.8 Appendix

Let us consider a Hamiltonian of type

$$H \quad := \quad \omega(\boldsymbol{Q})J + h(\boldsymbol{Q},\boldsymbol{P}) + \epsilon f(\phi, J, \boldsymbol{Q}, \boldsymbol{P}, \epsilon)\,,$$

$\phi \in \mathrm{T}$, $J \in \mathrm{R}^+$, $\boldsymbol{Q},\boldsymbol{P} \in \mathrm{R}^m$, and the scaled Lie-Poisson bracket

$$\{F,G\}_s \quad := \quad \{F,G\}_{\phi,J} + \epsilon\,\{F,G\}_{\boldsymbol{Q},\boldsymbol{P}}\,.$$

Problems of this type were considered in Section 3.6.2 (for the Hamiltonian system with a stiff spring we have $\omega(\boldsymbol{Q}) = 1$). Because of the non-standard Lie-Poisson bracket, the Hamiltonian normal form recursion (5.33)-(5.38) has to be modified. We use the same notations and point out only the differences. First, we define the transformed Hamiltonian

$$H_i \quad := \quad H \circ \boldsymbol{\Phi}_{1,\hat{\boldsymbol{W}}_i}$$

with $\hat{\boldsymbol{W}}_i := \{\mathbf{id}, S_i\}_s$ and $S_i = S_{i-1} + \epsilon^i \Delta S_i$. The Hamiltonian Δh_{i+1} is defined as before, i.e.

$$\Delta h_{i+1} \quad := \quad \frac{1}{(i+1)!}\left[\frac{\partial^{i+1}}{\partial \epsilon^{i+1}} H_i(\epsilon)\right]_{\epsilon=0}\,.$$

The homological equation is now replaced by

$$\{\Omega, \Delta S_{i+1}\}_{\phi,J} + \Delta h^n_{i+1} \quad = \quad 0$$

with $\Omega = \omega(\boldsymbol{Q})J + h(\boldsymbol{Q},\boldsymbol{P})$. Using

$$\begin{aligned}
H_{i+1} \quad &= \quad H \circ \boldsymbol{\Phi}_{1,\hat{\boldsymbol{W}}_i} \circ \boldsymbol{\Phi}_{1,\epsilon^{i+1}\boldsymbol{\Delta}\hat{\boldsymbol{W}}_{i+1}} + \mathcal{O}(\epsilon^{i+2})\,,\\
&= \quad H_i + \{H, \epsilon^{i+1}\Delta S_{i+1}\}_s + \mathcal{O}(\epsilon^{i+2})\,,\\
&= \quad H_i + \epsilon^{i+1}\{\Omega, \Delta S_{i+1}\}_{\phi,J} + \mathcal{O}(\epsilon^{i+2})\,,
\end{aligned}$$

it follows that this modified recursion indeed leads to the desired normal form, i.e.

$$H_i(\phi, J, \boldsymbol{Q}, \boldsymbol{P}) \quad = \quad \omega(\boldsymbol{Q})J + h(\boldsymbol{Q},\boldsymbol{P}) + \sum_{j=1}^{i} \epsilon^j \Delta h^r_j(J,\boldsymbol{Q},\boldsymbol{P}) + \mathcal{O}(\epsilon^{i+1})\,.$$

Let us now relate this modified Hamiltonian normal form recursion to the vector field normal form recursion (1.10)-(1.14). We make the *ansatz*

$$\boldsymbol{W}_i \quad = \quad \{\mathbf{id}, S_{i-1}\}_s + \epsilon^i\{\mathbf{id}, \Delta S_i\}_{\phi,J}\,, \tag{8.53}$$

i.e. $\hat{\boldsymbol{W}}_i - \boldsymbol{W}_i = \epsilon^{i+1}\{\mathbf{id}, \Delta S_i\}_{\boldsymbol{Q},\boldsymbol{P}}$. This yields the identities

$$\begin{aligned}
\boldsymbol{Y}_i \quad &= \quad \boldsymbol{\Phi}^*_{1,\boldsymbol{W}_i}\boldsymbol{Y}\,,\\
&= \quad \boldsymbol{\Phi}^*_{1,\boldsymbol{W}_i - \hat{\boldsymbol{W}}_i}\boldsymbol{\Phi}^*_{1,\hat{\boldsymbol{W}}_i}\{\mathbf{id}, H\}_s + \mathcal{O}(\epsilon^{i+2})\,,\\
&= \quad \{\mathbf{id}, H_i\}_s + \epsilon^{i+1}\left[\{\mathbf{id}, \Delta S_i\}_{\boldsymbol{Q},\boldsymbol{P}}, \{\mathbf{id}, \Omega\}_{\phi,J}\right] + \mathcal{O}(\epsilon^{i+2})\,.
\end{aligned}$$

where we have used $\boldsymbol{Y} = \{\mathbf{id}, H\}_s$. Thus

$$\boldsymbol{\Delta X}_{i+1} \quad = \quad \{\mathbf{id}, \Delta h_i^r\}_{\boldsymbol{Q},\boldsymbol{P}} + \{\mathbf{id}, \Delta h_{i+1}\}_{\phi,\boldsymbol{J}} + \left[\{\mathbf{id}, \Delta S_i\}_{\boldsymbol{Q},\boldsymbol{P}}, \{\mathbf{id}, \Omega\}_{\phi,\boldsymbol{J}}\right].$$

The corresponding homological equation (1.14) with $\boldsymbol{A} = \{\mathbf{id}, \Omega\}_{\phi,\boldsymbol{J}}$ and

$$\boldsymbol{\Delta X}_{i+1}^n \quad = \quad \{\mathbf{id}, \Delta h_{i+1}^n\}_{\phi,\boldsymbol{J}} + \left[\{\mathbf{id}, \Delta S_i\}_{\boldsymbol{Q},\boldsymbol{P}}, \{\mathbf{id}, \Omega\}_{\phi,\boldsymbol{J}}\right]$$

is solved by

$$\boldsymbol{\Delta W}_{i+1} \quad = \quad \{\mathbf{id}, \Delta S_{i+1}\}_{\phi,\boldsymbol{J}} + \{\mathbf{id}, \Delta S_i\}_{\boldsymbol{Q},\boldsymbol{P}}$$

and $\boldsymbol{W}_{i+1} = \boldsymbol{W}_i + \epsilon^{i+1}\boldsymbol{\Delta W}_{i+1}$ satisfies the *ansatz* (8.53) with i replaced by $i+1$ as desired.

4

Highly-Oscillatory Systems

This chapter is about highly-oscillatory mechanical systems that, in the limit, reduce to rigid bodies. In the first part, we review theoretical results on the elimination of fast internal vibrations. In particular, we give a new proof for Jean's conjecture on the exponentially decoupling of slow rigid body motions and fast internal vibrations. A rigorous proof of this result was first given by Benettin, Galgani & Giorgilli [15]. In the second part, we discuss effective integrators for rigid bodies, the concept of soft-constraints, and a modified multiple-time-stepping method that avoids the resonance problems associated with standard multiple-time-stepping [19]. The explicit symplectic integrator for rigid body motion was first derived by the author in [101]. The concept of soft constraints was introduced in [126],[97],[99]. Here a new implementation is suggested that is easier and cheaper to implement. The suggested projected multiple-time-stepping method was inspired by reading the paper [42] by García-Arcgilla, Sanz-Serna & Skeel.

4.1 Theoretical Results

4.1.1 Systems Near an Equilibrium Point

In this section, we consider systems

$$\frac{d}{dt}\tilde{\boldsymbol{x}} \quad = \quad \boldsymbol{A}\tilde{\boldsymbol{x}} + \boldsymbol{f}(\tilde{\boldsymbol{x}})\,, \tag{1.1}$$

$\tilde{\boldsymbol{x}} \in \mathrm{C}^n$, $\boldsymbol{A} \in \mathrm{C}^{n\times n}$ a diagonal matrix, and $\boldsymbol{f} : \mathcal{B}_r\{\boldsymbol{0}\} \subset \mathrm{C}^n \to \mathrm{C}^n$ an analytic function with $\boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{0}$ and Jacobian $\boldsymbol{\partial}_{\tilde{\boldsymbol{x}}}\boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{0}$. Here $\mathcal{B}_r\{\boldsymbol{0}\}$ denotes the complex ball of radius $r > 0$ around $\boldsymbol{0} \in \mathrm{C}^n$. We denote by $\boldsymbol{e}_s$, $s = 1,\ldots,n$, the basis vectors in R^n. Then the Taylor expansion of $\boldsymbol{f}$ can be written as

$$\boldsymbol{f}(\tilde{\boldsymbol{x}}) \quad = \quad \sum_s \sum_{\boldsymbol{k}\in\mathbb{I}^n} f_{\boldsymbol{k},s}\tilde{\boldsymbol{x}}^{\boldsymbol{k}}\boldsymbol{e}_s\,.$$

Here $\tilde{\boldsymbol{x}}^{\boldsymbol{k}}$ denotes the monomial

$$\tilde{\boldsymbol{x}}^{\boldsymbol{k}} \quad = \quad \tilde{x}_1^{k_1}\tilde{x}_2^{k_2}\cdots\tilde{x}_n^{k_n}\,,$$

$\tilde{\boldsymbol{x}} = (\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_n)^T$, $\boldsymbol{k} = (k_1,k_2,\ldots,k_n)^T$, $\boldsymbol{k} \in \mathrm{I}^n$, I the set of non-negative integers. The function space P_K is the space of all analytic functions $\boldsymbol{f} : \mathcal{B}_r\{\boldsymbol{0}\} \subset \mathrm{C}^n \to \mathrm{C}^n$ whose Taylor series representation contains only monomials $\tilde{\boldsymbol{x}}^{\boldsymbol{k}}$ with $|\boldsymbol{k}| \le K$, i.e.

$$\boldsymbol{k} \in \mathrm{I}_K^n \quad := \quad \{\boldsymbol{k} \in \mathrm{I}^n : |\boldsymbol{k}| \le K\}\,,$$

$|\boldsymbol{k}| = k_1 + \cdots + k_n$. For a bounded analytic function $\boldsymbol{f} : \mathcal{B}_r\{\boldsymbol{0}\} \subset \mathrm{C}^n \to \mathrm{C}^n$, we also introduce the norm

$$|\boldsymbol{f}|_r \quad := \quad \sum_s \sum_{\boldsymbol{k}\in\mathbb{I}^n} |f_{\boldsymbol{k},s}|\, r^{|\boldsymbol{k}|}$$

on a complex ball of radius $r > 0$ around $\tilde{\boldsymbol{x}} = \boldsymbol{0}$. Since we are interested in the normal form of (1.1) near the equilibrium $\tilde{\boldsymbol{x}} = \boldsymbol{0}$, we scale $\tilde{\boldsymbol{x}}$ by the small parameter ϵ, i.e. $\boldsymbol{x} := \tilde{\boldsymbol{x}}/\epsilon$, and obtain the scaled differential equation

$$\begin{aligned} \frac{d}{dt}\boldsymbol{x} \quad &= \quad \boldsymbol{A}\boldsymbol{x} + \epsilon^{-1}\boldsymbol{f}(\epsilon\boldsymbol{x})\,, \\ &= \quad \boldsymbol{A}\boldsymbol{x} + \epsilon\, \boldsymbol{B}(\boldsymbol{x},\epsilon)\,. \end{aligned} \tag{1.2}$$

This scaled system is of the type considered in the previous chapter. The corresponding homological equation

$$[\boldsymbol{A}, \boldsymbol{\Delta W}] + \boldsymbol{\Delta X}^n \quad = \quad \boldsymbol{0}$$

can be solved in terms of the Taylor expansions of $\boldsymbol{\Delta W}$ and $\boldsymbol{\Delta X}^n$: Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T$ denote the vector of eigenvalues of the matrix $\boldsymbol{A}$. Then we define the resonance module $\mathcal{M}_s \subset \mathrm{I}^n$, $s = 1, \ldots, n$, by

$$\mathcal{M}_s \quad := \quad \{\boldsymbol{k} \in \mathrm{I}^n : \lambda_s = \boldsymbol{\lambda}^T\boldsymbol{k}\,\}\,.$$

We assume that the given $\boldsymbol{\Delta X}$ is an element of P_K, $K \geq 2$. The resonant part $\boldsymbol{\Delta X}^r$ of

$$\boldsymbol{\Delta X}(\boldsymbol{x}) \quad = \quad \sum_s \sum_{\boldsymbol{k}\in\mathbb{I}^n_K} \Delta X_{\boldsymbol{k},s}\, \boldsymbol{x}^{\boldsymbol{k}} \boldsymbol{e}_s$$

is defined by

$$\boldsymbol{\Delta X}^r(\boldsymbol{x}) \quad = \quad \sum_s \sum_{\boldsymbol{k}\in\mathcal{M}_s} \Delta X_{\boldsymbol{k},s}\, \boldsymbol{x}^{\boldsymbol{k}} \boldsymbol{e}_s$$

and the non-resonant part by $\boldsymbol{\Delta X}^n = \boldsymbol{\Delta X} - \boldsymbol{\Delta X}^r$. The solution of the homological equations is now given by (in terms of the Taylor series coefficients) [5]:

$$\Delta W_{\boldsymbol{k},s} \quad = \quad \frac{-\Delta X_{\boldsymbol{k},s}}{\lambda_s - \boldsymbol{\lambda}^T\boldsymbol{k}}\,, \qquad \boldsymbol{k} \in \mathrm{I}^n_K \setminus \mathcal{M}_s\,,$$

and $\Delta W_{\boldsymbol{k},s} = 0$ for $\boldsymbol{k} \in \mathcal{M}_s$. Note that $\boldsymbol{\Delta W} \in P_K$ with

$$\boldsymbol{\Delta W}(\boldsymbol{x}) \quad = \quad \sum_s \sum_{\boldsymbol{k}\in\mathbb{I}^n_K} \Delta W_{\boldsymbol{k},s}\, \boldsymbol{x}^{\boldsymbol{k}} \boldsymbol{e}_s\,.$$

We assume that the estimate [5]

$$|\boldsymbol{\lambda}^T\boldsymbol{k} - \lambda_s| \quad \geq \quad \frac{\delta}{(K-1)^{d-1}}\,, \qquad \boldsymbol{k} \in \mathrm{I}^n_K \setminus \mathcal{M}_s$$

holds for all $s = 1, \dots, n$ and all $K \geq 2$ as well as

$$|\boldsymbol{\lambda}^T \boldsymbol{k} - \lambda_s| \quad \geq \quad \delta\,, \qquad \boldsymbol{k} \in \mathrm{I}_K^n \setminus \mathcal{M}_s\,, \qquad K = 0, 1\,.$$

Here $\delta > 0$ and $d \geq 0$ are appropriate constants depending on the eigenvalues of $\boldsymbol{A}$. Then the solution $\boldsymbol{\Delta W}$ satisfies the estimate

$$|\boldsymbol{\Delta W}|_r \quad \leq \quad \gamma^{-1}\,|\boldsymbol{\Delta X}|_r\,, \qquad \gamma = \frac{\delta}{(K-1)^{d-1}}\,. \tag{1.3}$$

With this, we can now start the normal form recursion (1.10)-(1.14). In fact, it is easier to first look at the system

$$\begin{aligned} \frac{d}{dt}\epsilon\boldsymbol{x} \quad &= \quad \boldsymbol{A}\epsilon\boldsymbol{x} + \boldsymbol{f}(\epsilon\boldsymbol{x})\,, \\ &=: \quad \hat{\boldsymbol{Y}}(\epsilon\boldsymbol{x})\,. \end{aligned}$$

For this $\hat{\boldsymbol{Y}}$, the normal form recursion yields vector fields $\boldsymbol{\Delta}\hat{\boldsymbol{X}}_i(\boldsymbol{x})$ that are obviously elements of the space P_i, i.e. $\epsilon^i \boldsymbol{\Delta}\hat{\boldsymbol{X}}_i(\boldsymbol{x}) = \boldsymbol{\Delta}\hat{\boldsymbol{X}}_i(\epsilon\boldsymbol{x}, \epsilon)$. Now $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ are related by $\hat{\boldsymbol{Y}} = \epsilon\boldsymbol{Y}$ and, thus, we have $\boldsymbol{\Delta}\hat{\boldsymbol{X}}_{i+1} = \boldsymbol{\Delta X}_i$. This implies that $\boldsymbol{\Delta X}_i(\boldsymbol{x})$ must be an element of P_{i+1} and

$$\epsilon^i\,\boldsymbol{\Delta X}_i(\boldsymbol{x}) \quad = \quad \epsilon^{-1}\boldsymbol{\Delta X}_i(\epsilon\boldsymbol{x}, \epsilon)\,.$$

Let us now apply Theorem 3.1. With $\mathcal{K} = \{\boldsymbol{0}\}$ the compact set and $\epsilon^{-1}\boldsymbol{f}(\epsilon\boldsymbol{x})$ satisfying an estimate

$$\epsilon^{-1}|\boldsymbol{f}(\epsilon)|_R \quad \leq \quad \epsilon\,, \qquad R > 0\,,$$

we could, in principle, apply Theorem 3.1. We have to be a bit carefully though: Since $\boldsymbol{\Delta X}_i \in P_{i+1}$, the parameter γ in (1.3) depends on the iteration index i with $K = i+1$. In other words, we have to use the estimate (4.32) of Proposition 3.1 for the error in the normal form truncation

$$\bar{\boldsymbol{Y}}_{i_*}(\bar{\boldsymbol{x}}, \epsilon) \quad = \quad \boldsymbol{A}\bar{\boldsymbol{x}} + \sum_{i=1}^{i_*} \epsilon^{-1}\boldsymbol{\Delta f}_i^r(\epsilon\bar{\boldsymbol{x}}, \epsilon)\,, \qquad (\bar{\boldsymbol{x}} \in \mathcal{B}_{R/2}\mathcal{K})\,,$$

i.e., the truncation error $\boldsymbol{T}(\bar{\boldsymbol{x}}, \epsilon)$ satisfies

$$||\,\boldsymbol{T}(\epsilon)\,||_{R/2} \quad \leq \quad c_1\,\epsilon\,e^{-(c_2/\epsilon)^{1/d}}\,,$$

$c_1, c_2 > 0$ appropriate constants and d proportional to the number of non-resonant eigenvalues of $\boldsymbol{A}$. Here

$$\epsilon^{-1}\boldsymbol{\Delta f}_i^r(\epsilon\boldsymbol{x}, \epsilon) \quad := \quad \epsilon^i\boldsymbol{\Delta X}_i^r(\boldsymbol{x})\,.$$

Remark 4.1. For the subsequent sections, it is important to note that (i) the linear vector field $\boldsymbol{A}\bar{\boldsymbol{x}}$ commutes with all the resonant vector fields $\boldsymbol{\Delta f}_i^r$ and (ii) $d = 1$ in

(1.3) if either $n = 1$, i.e., the system has one-degree of freedom, or $\lambda_i = \lambda$, i.e., all the eigenvalues of $\boldsymbol{A}$ are identical. If the vector field $\boldsymbol{A}\bar{\boldsymbol{x}}$ is oscillatory, then d is equal to the number of non-resonant frequencies in the system. □

Remark 4.2. The normal form expansion converges if all the eigenvalues of $\boldsymbol{A}$ are in the Poincaré domain [29],[5]. □

Example 4.1. Let us consider a system of fast linear oscillators subject to a slow non-linear perturbation, i.e.,

$$\begin{aligned} \frac{d}{d\tau}\tilde{\boldsymbol{q}} &= \boldsymbol{M}(\tilde{\boldsymbol{q}})^{-1}\,\boldsymbol{p}\,, \\ \frac{d}{d\tau}\boldsymbol{p} &= -\epsilon^{-2}\tilde{\boldsymbol{q}} - \boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}V(\tilde{\boldsymbol{q}}) - \boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}\frac{\boldsymbol{p}^T\boldsymbol{M}(\tilde{\boldsymbol{q}})^{-1}\boldsymbol{p}}{2}\,, \end{aligned}$$

$\tilde{\boldsymbol{q}}, \boldsymbol{p} \in \mathrm{R}^n$. We write

$$\boldsymbol{M}(\tilde{\boldsymbol{q}})^{-1} \;=\; \boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_1(\tilde{\boldsymbol{q}})$$

and assume that $\boldsymbol{\Omega}_0$ is a positive diagonal matrix. Furthermore, we scale $\tilde{\boldsymbol{q}}$ such that

$$\epsilon^{-1}\tilde{\boldsymbol{q}} \;=:\; \boldsymbol{q}$$

and transform time by a factor of ϵ, i.e. $\tau = \epsilon t$. Thus the scaled equations of motion are

$$\begin{aligned} \frac{d}{dt}\boldsymbol{q} &= \boldsymbol{M}(\epsilon\boldsymbol{q})^{-1}\,\boldsymbol{p}\,, \\ \frac{d}{dt}\boldsymbol{p} &= -\boldsymbol{q} - \boldsymbol{\nabla}_{\boldsymbol{q}}V(\epsilon\boldsymbol{q}) - \boldsymbol{\nabla}_{\boldsymbol{q}}\frac{\boldsymbol{p}^T\boldsymbol{M}(\epsilon\boldsymbol{q})^{-1}\boldsymbol{p}}{2}\,. \end{aligned}$$

These equations of motion are still Hamiltonian. Note that we have used

$$\epsilon\,\boldsymbol{\nabla}_{\epsilon\boldsymbol{q}}V(\epsilon\boldsymbol{q}) \;=\; \boldsymbol{\nabla}_{\boldsymbol{q}}V(\epsilon\boldsymbol{q})\,.$$

For simplicity, we assume that the gradient and the Hessian of V are identical equal to zero at $\boldsymbol{q} = \boldsymbol{0}$. Let us now introduce the new variable $\tilde{\boldsymbol{x}} = (\boldsymbol{q}^T, \boldsymbol{p}^T)^T$. Then the above system can be written as

$$\frac{d}{dt}\tilde{\boldsymbol{x}} \;=\; \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} + \epsilon^{-1}\tilde{\boldsymbol{f}}(\epsilon\tilde{\boldsymbol{x}})\,,$$

where

$$\tilde{\boldsymbol{A}} \;:=\; \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Omega}_0 \\ -\boldsymbol{I} & \boldsymbol{0} \end{pmatrix}$$

and

$$\epsilon^{-1}\tilde{\boldsymbol{f}}(\epsilon\tilde{\boldsymbol{x}}, \epsilon) \;=\; \epsilon^{-1}\begin{pmatrix} \boldsymbol{\Omega}_1(\epsilon\boldsymbol{q})\epsilon\boldsymbol{p} \\ -\epsilon\,\boldsymbol{\nabla}_{\boldsymbol{q}}V(\epsilon\boldsymbol{q}) - \boldsymbol{\nabla}_{\boldsymbol{q}}\frac{(\epsilon\boldsymbol{p})^T\boldsymbol{\Omega}_1(\epsilon\boldsymbol{q})(\epsilon\boldsymbol{p})}{2\,\epsilon} \end{pmatrix}.$$

Note that taking $\boldsymbol{\nabla}_{\boldsymbol{q}}$ yields a factor ϵ and that the equations are Hamiltonian. Next we complexify and transform the matrix $\tilde{\boldsymbol{A}}$ to diagonal form by a unitary transformation $\boldsymbol{T}$. Introducing the new variable $\boldsymbol{x} = \boldsymbol{T}\tilde{\boldsymbol{x}}$, we end up with a system of type (1.2) to which normal form theory can be applied. In particular, let us assume that the function $\boldsymbol{f}$ is analytic in a complex ball of radius $R > 0$ around $\boldsymbol{0} \in \mathrm{C}^{2n}$ such that

$$|\epsilon^{-1}\boldsymbol{f}(\epsilon)|_R \quad \leq \quad \epsilon\,.$$

Furthermore, let us assume that $d\boldsymbol{x}/dt = \boldsymbol{A}\boldsymbol{x}$ decomposes into $d \geq 1$ blocks of oscillators with equal frequency $\pm\omega_l$. Let us denote the corresponding vector of frequencies by $\boldsymbol{\omega} \in \mathrm{R}^d$. Then we assume that

$$|\boldsymbol{k}^T\boldsymbol{\omega}| \quad \geq \quad \frac{\delta}{K^{d-1}}\,, \qquad |\boldsymbol{k}| \in \mathrm{Z}_K^d \setminus \{\boldsymbol{0}\}\,.$$

This implies that the eigenvalues of $\boldsymbol{A}$ satisfy the estimate (1.3) with the resonance modules $\mathcal{M}_s$ defined appropriately. Thus we can apply Proposition 3.1 and (4.32) to estimate the error in the normal form truncation. What are the consequences on the dynamics of the perturbed system of harmonic oscillators? With each block of oscillators of equal frequency ω_l, we can associate the energy

$$E_l \quad = \quad \sum_i \frac{\omega_l^2}{2}p_i^2 + \frac{1}{2\epsilon^2}q_i^2$$

where the index i runs over all the oscillators in the block. The corresponding energy term $\bar{E}_l$ in normal form coordinates is a first integral of the truncated normal form system. Thus, E_l is an adiabatic invariant and is preserved up to fluctuations of size $\mathcal{O}(\epsilon^{1/d})$ over a time period $[0, T]$ with

$$T \quad \leq \quad c_1\,\epsilon\, e^{(c_2/\epsilon)^d}\,,$$

$c_1, c_2 > 0$ appropriate constants, provided that the solutions $\boldsymbol{x}(t)$ stay in $\mathcal{B}_{R/2}\{\boldsymbol{0}\}$. The energy exchange between the oscillators within one block of equal frequency is non-zero, in general, and is determined by the resonant terms in the normal form.

Hamiltonian systems near an equilibrium are typically discussed in terms of the Birkhoff normal form of the corresponding Hamiltonian (see Birkhoff [20], Siegel [113], de Almeida [29]). A proof for the exponentially smallness of the truncation error in the asymptotic expansion has already been given by Giorgilli & Galgani [47], Giorgilli, Delshams, Frontich, Galgani & Simó [46], Delshams & Gutiérrez [31]. □

4.1.2 Elimination of Fast Internal Vibrations

Our interest is now in systems that have fast and slow degrees of freedom. In particular, we are interested in highly oscillatory Hamiltonian systems of type

$$\begin{aligned}
\frac{d}{d\tau}\boldsymbol{q} \quad &= \quad \boldsymbol{M}^{-1}\boldsymbol{p}\,,\\
\frac{d}{d\tau}\boldsymbol{p} \quad &= \quad -\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}) - \epsilon^{-2}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{g}(\boldsymbol{q})\,,
\end{aligned}$$

$\boldsymbol{q}, \boldsymbol{p} \in \mathrm{R}^{3N}$, $\boldsymbol{g} : \mathrm{R}^{3N} \to \mathrm{R}^m$, $m < 3N$. We assume that the $m \times m$ matrix $\boldsymbol{\partial_q g}(\boldsymbol{q})\boldsymbol{M}^{-1}\boldsymbol{\nabla_q g}(\boldsymbol{q})$ is invertible. The Hamiltonian is

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

We like to see under which conditions the highly oscillatory system can be replaced by the constrained system

$$\begin{aligned}
\frac{d}{d\tau}\bar{\boldsymbol{q}} &= \boldsymbol{M}^{-1}\bar{\boldsymbol{p}}\,,\\
\frac{d}{d\tau}\bar{\boldsymbol{p}} &= -\boldsymbol{\nabla}_{\bar{\boldsymbol{q}}}V(\bar{\boldsymbol{q}}) - \boldsymbol{\nabla}_{\bar{\boldsymbol{q}}}\boldsymbol{g}(\bar{\boldsymbol{q}})\,\boldsymbol{\lambda}\,,\\
\boldsymbol{0} &= \boldsymbol{g}(\bar{\boldsymbol{q}})\,,
\end{aligned}$$

$\boldsymbol{\lambda} \in \mathrm{R}^m$ the vector of Lagrange multipliers. Here we will only consider those systems that, as a constrained system, reduce to a system of decoupled rigid bodies. In other words, we assume that the matrix $\boldsymbol{\partial_q g}(\boldsymbol{q})\boldsymbol{M}^{-1}\boldsymbol{\nabla_q g}(\boldsymbol{q})$ is constant along solution curves of the constrained system. Note that the case $m = 1$, i.e., a single fast degree of motion has been discussed in Section 3.6.2. There a single diatomic molecule was considered. Here we are interested in systems of diatomic or other small molecular systems like, for example, water. We assume that the internal bonded interactions are modeled by stiff harmonic forces. The question is under which conditions these harmonic forces can be replaced by rigid constraints.

Let us introduce local coordinates

$$\tilde{\boldsymbol{r}} = \boldsymbol{g}(\boldsymbol{q}) \in \mathrm{R}^m\,, \qquad \boldsymbol{Q} = \boldsymbol{b}(\boldsymbol{q}) \in \mathrm{R}^{3N-m}\,,$$

and corresponding conjugate momenta $\tilde{\boldsymbol{p}}_{\boldsymbol{r}} \in \mathrm{R}^m$ and $\boldsymbol{P} \in \mathrm{R}^{3N-m}$ [97],[99]. Here $\boldsymbol{b} : \mathrm{R}^{3N} \to \mathrm{R}^{3N-m}$ is an appropriate function with

$$\boldsymbol{\partial_q b}(\boldsymbol{q})\,\boldsymbol{M}^{-1}\,\boldsymbol{\nabla_q g}(\boldsymbol{q}) \;=\; \boldsymbol{0}\,.$$

In these coordinates, the highly oscillatory system becomes

$$\begin{aligned}
\frac{d}{d\tau}\tilde{\boldsymbol{r}} &= \boldsymbol{M}_1(\tilde{\boldsymbol{r}})\,\tilde{\boldsymbol{p}}_{\boldsymbol{r}}\,,\\
\frac{d}{d\tau}\tilde{\boldsymbol{p}}_{\boldsymbol{r}} &= -\epsilon^{-2}\tilde{\boldsymbol{r}} - \boldsymbol{\nabla}_{\tilde{\boldsymbol{r}}}V(\tilde{\boldsymbol{r}},\boldsymbol{Q}) - \boldsymbol{\nabla}_{\tilde{\boldsymbol{r}}}\left[\frac{\tilde{\boldsymbol{p}}_{\boldsymbol{r}}{}^T\,\boldsymbol{M}_1(\tilde{\boldsymbol{r}})\,\tilde{\boldsymbol{p}}_{\boldsymbol{r}}}{2} + \frac{\boldsymbol{P}^T\,\boldsymbol{M}_2(\tilde{\boldsymbol{r}},\boldsymbol{Q})\,\boldsymbol{P}}{2}\right],\\
\frac{d}{d\tau}\boldsymbol{Q} &= \boldsymbol{M}_2(\tilde{\boldsymbol{r}},\boldsymbol{Q})\,\boldsymbol{P}\,,\\
\frac{d}{d\tau}\boldsymbol{P} &= -\boldsymbol{\nabla}_{\boldsymbol{Q}}V(\tilde{\boldsymbol{r}},\boldsymbol{Q}) - \boldsymbol{\nabla}_{\boldsymbol{Q}}\frac{\boldsymbol{P}^T\,\boldsymbol{M}_2(\tilde{\boldsymbol{r}},\boldsymbol{Q})\,\boldsymbol{P}}{2}\,.
\end{aligned}$$

Here $\boldsymbol{M}_1 := \boldsymbol{\partial_q g}\boldsymbol{M}^{-1}\boldsymbol{\nabla_q g}$ and $\boldsymbol{M}_2 := \boldsymbol{\partial_q b}\boldsymbol{M}^{-1}\boldsymbol{\nabla_q b}$.

Remark 4.3. In general, the matrix $\boldsymbol{M}_1$ will also depend on the slow degrees of freedom $\boldsymbol{Q}$. (See Section 3.6.2 for an example with a single fast degree of motion.)

But this case is not considered here. Let us just mention a few points: If $\boldsymbol{M}_1$ depends on the slow variable $\boldsymbol{Q}$, then the gradient of the corresponding kinetic energy term with respect to $\boldsymbol{Q}$ leads to a force on the slow degrees of motion that depends on the energy and the frequency of the fast degrees of motion. As long as there are no resonances, the ratio of energy and frequency is an adiabatic invariant for each fast degree of motion. However, as $\boldsymbol{Q}$ varies, the fast system will undergo resonances which lead to a drift in the corresponding adiabatic invariants. Generically, this drift will be slow [7], i.e. of order $\mathcal{O}(\sqrt{\epsilon})$, but cannot be neglected over exponentially long periods of time. Thus, it seems impossible to decouple the slow and fast degrees of motion over exponentially long periods of time. □

Concerning the fast degrees of freedom $(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{p}}_{\boldsymbol{r}}) \in \mathrm{R}^{2m}$, we make the same assumptions as in Example 4.1. Thus we can apply the same transformations: (i) define $\boldsymbol{r} := \epsilon^{-1}\tilde{\boldsymbol{r}}$, $\boldsymbol{p}_{\boldsymbol{r}} = \tilde{\boldsymbol{p}}_{\boldsymbol{r}}$ and consider $(\boldsymbol{Q}, \boldsymbol{P})$ as a parameter, (ii) rescale time: $\tau = \epsilon t$ and arrive at the system

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{r} &= \boldsymbol{M}_1(\epsilon \boldsymbol{r})\, \boldsymbol{p}_{\boldsymbol{r}}\,, \\
\frac{d}{dt}\boldsymbol{p}_{\boldsymbol{r}} &= -\boldsymbol{r} - \boldsymbol{\nabla}_{\boldsymbol{r}} V(\epsilon \boldsymbol{r}, \boldsymbol{Q}) - \boldsymbol{\nabla}_{\boldsymbol{r}} \left[\frac{\boldsymbol{p}_{\boldsymbol{r}}{}^T\, \boldsymbol{M}_1(\epsilon \boldsymbol{r})\, \boldsymbol{p}_{\boldsymbol{r}}}{2} + \frac{\boldsymbol{P}^T\, \boldsymbol{M}_2(\epsilon \boldsymbol{r}, \boldsymbol{Q})\, \boldsymbol{P}}{2} \right], \\
\frac{d}{dt}\boldsymbol{Q} &= \epsilon \boldsymbol{M}_2(\epsilon \boldsymbol{r}, \boldsymbol{Q})\, \boldsymbol{P}\,, \\
\frac{d}{dt}\boldsymbol{P} &= -\epsilon \boldsymbol{\nabla}_{\boldsymbol{Q}} V(\epsilon \boldsymbol{r}, \boldsymbol{Q}) - \epsilon \boldsymbol{\nabla}_{\boldsymbol{Q}} \frac{\boldsymbol{P}^T\, \boldsymbol{M}_2(\epsilon \boldsymbol{r}, \boldsymbol{Q})\, \boldsymbol{P}}{2}\,.
\end{aligned}$$

This system is Hamiltonian with respect to the non-standard Lie-Poisson bracket

$$\{F, G\}_s \;:=\; \{F, G\}_{\boldsymbol{r},\boldsymbol{p}_{\boldsymbol{r}}} + \epsilon\,\{F, G\}_{\boldsymbol{Q},\boldsymbol{P}}$$

where $\{F, G\}_{\boldsymbol{r},\boldsymbol{p}_{\boldsymbol{r}}}$ is the canonical bracket in the $(\boldsymbol{r}, \boldsymbol{p}_{\boldsymbol{r}})$ variable and $\{F, G\}_{\boldsymbol{Q},\boldsymbol{P}}$ is the canonical bracket in the $(\boldsymbol{Q}, \boldsymbol{P})$ variable. We write

$$\boldsymbol{M}_1(\epsilon \boldsymbol{r}) \;=\; \boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_1(\epsilon \boldsymbol{r})\,.$$

Let us introduce the new variables $\tilde{\boldsymbol{x}} = (\boldsymbol{r}^T, \boldsymbol{p}_{\boldsymbol{r}}{}^T)^T \in \mathrm{R}^{2m}$ and $\boldsymbol{y} = (\boldsymbol{Q}^T, \boldsymbol{P}^T)^T \in \mathrm{R}^{6N-2m}$. Then the above system can be written as

$$\begin{aligned}
\frac{d}{dt}\tilde{\boldsymbol{x}} &= \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} + \epsilon^{-1}\, \tilde{\boldsymbol{f}}(\epsilon \tilde{\boldsymbol{x}}, \boldsymbol{y}, \epsilon)\,, \\
\frac{d}{dt}\boldsymbol{y} &= \qquad\quad + \epsilon\, \tilde{\boldsymbol{g}}(\epsilon \tilde{\boldsymbol{x}}, \boldsymbol{y}, \epsilon)
\end{aligned}$$

with

$$\tilde{\boldsymbol{A}} \;:=\; \begin{pmatrix} \boldsymbol{0} & \boldsymbol{\Omega}_0 \\ -\boldsymbol{I} & \boldsymbol{0} \end{pmatrix}.$$

To such a system we like to apply normal form theory. We are mainly interested in the rate of energy exchange between the fast (internal) degrees of freedom $\tilde{\boldsymbol{x}}$ and the slow (external) degrees of freedom $\boldsymbol{y}$. The energy in the fast degrees of freedom is given by

$$E_f \;=\; \frac{\boldsymbol{p_r}^T \boldsymbol{\Omega}_0 \boldsymbol{p_r}}{2} + \frac{\boldsymbol{r}^T \boldsymbol{r}}{2}\,.$$

To be able to apply normal form theory, we complexify the system, i.e. $\tilde{\boldsymbol{x}} \in \mathrm{C}^{2m}$ and $\boldsymbol{y} \in \mathrm{C}^{6N-2m}$, and transform $\tilde{\boldsymbol{A}}$ to diagonal form by a unitary transformation $\boldsymbol{T}$. Thus, with $\boldsymbol{x} = \boldsymbol{T}\tilde{\boldsymbol{x}}$, we obtain a (complexified) system of type

$$\begin{aligned} \frac{d}{dt}\boldsymbol{x} &= \boldsymbol{A}\boldsymbol{x} + \epsilon^{-1}\,\boldsymbol{f}(\epsilon\boldsymbol{x},\boldsymbol{y},\epsilon)\,, \\ \frac{d}{dt}\boldsymbol{y} &= \qquad + \epsilon\,\boldsymbol{g}(\epsilon\boldsymbol{x},\boldsymbol{y},\epsilon)\,, \end{aligned}$$

with $\boldsymbol{x} \in \mathrm{C}^{2m}$ standing for the fast (internal) degrees of freedom and $\boldsymbol{y} = (\boldsymbol{Q}^T, \boldsymbol{P}^T)^T \in \mathrm{C}^{6N-2m}$.

The Benettin/Galgani/Giorgilli (BGG) Result

Let us assume that, on $\mathcal{B}_R\mathcal{K}$, $\mathcal{K} = \{\boldsymbol{0}\} \times \mathcal{V} \subset \mathrm{C}^{6N}$, $\mathcal{V} \subset \mathrm{R}^{6N-2m}$ a compact subset, we have

$$|\epsilon^{-1}\boldsymbol{f}(\epsilon)|_R \;\le\; \epsilon \quad \text{and} \quad |\epsilon\boldsymbol{g}(\epsilon)|_R \;\le\; \epsilon\,.$$

An appropriate norm $|.|_R$ will be defined in the following subsection. Let us also assume that $d\boldsymbol{x}/dt = \boldsymbol{A}\boldsymbol{x}$ decompose into $d \ge 1$ blocks of oscillators with equal frequency $\pm\omega_l$. Let us denote the corresponding vector of frequencies by $\boldsymbol{\omega} \in \mathrm{R}^d$. Then we assume that

$$|\boldsymbol{k}^T\boldsymbol{\omega}| \;\ge\; \frac{\delta}{K^{d-1}}\,, \qquad |\boldsymbol{k}| \in \mathrm{Z}^d_K\,.$$

This implies that the eigenvalues of $\boldsymbol{A}$ satisfy the estimate (1.3) with the resonance modules $\mathcal{M}_s$ defined appropriately. Benettin, Galgani & Giorgilli [13],[15] show then that there exists a coordinate transformation such that in the new coordinates

$$\begin{aligned} \frac{d}{dt}\bar{\boldsymbol{x}} &= \boldsymbol{A}\bar{\boldsymbol{x}} + \epsilon^{-1}\sum_i \boldsymbol{\Delta f}_i^r(\epsilon\bar{\boldsymbol{x}},\bar{\boldsymbol{y}},\epsilon)\,, \\ \frac{d}{dt}\bar{\boldsymbol{y}} &= \epsilon\,\boldsymbol{g}(\boldsymbol{0},\bar{\boldsymbol{y}}) + \epsilon\sum_i \boldsymbol{\Delta g}_i^r(\epsilon\bar{\boldsymbol{x}},\bar{\boldsymbol{y}},\epsilon) \end{aligned}$$

up to terms $\boldsymbol{T}(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}},\epsilon)$ satisfying an estimate of type (4.32), i.e.

$$||\,\boldsymbol{T}(\epsilon)\,||_{R/2} \;\le\; c_1\,\epsilon\,e^{-(c_2/\epsilon)^{1/d}}\,,$$

$c_1, c_2 > 0$ appropriate constants. To be more precise, Benettin, Galgani & Giorgilli show that the truncation error in the normal form expansion of the Hamiltonian can be made exponentially small. This, of course, implies a corresponding estimate for the equations of motion in normal form.

What are the consequences on the dynamics of the perturbed system? With each block of fast (internal) oscillators of equal frequency ω_l, we can associate the energy

$$E_l \;=\; \sum_i \frac{\omega_l^2}{2} p_i^2 + \frac{1}{2\epsilon^2} q_i^2$$

where the index i runs over all the oscillators in the block. The corresponding energy term $\bar{E}_l$ in normal form coordinates is a first integral of the truncated normal form system. Thus, E_l is an adiabatic invariant and is preserved up to fluctuations of size $\mathcal{O}(\epsilon^{1/d})$ over a time period $[0, T]$ with

$$T \;\leq\; c_1^{-1}\, e^{(c_2/\epsilon)^{1/d}}$$

(assuming that the solutions stay in $\mathcal{B}_{R/2}\mathcal{K}$). The energy exchange between the oscillators within one block is non-zero, in general, and is determined by the resonant terms in the normal form. Furthermore, the total internal vibrational energy is also an adiabatic invariant which is preserved over the same period of time T. Thus, for low initial internal vibrational energy, the solutions will stay close to the constraint manifold

$$\mathcal{M} \;:=\; \{\boldsymbol{q} \in \mathrm{R}^{3N} : \boldsymbol{g}(\boldsymbol{q}) = 0\,\}$$

and the external degrees of freedom move on a surface of constant energy. Thus the system effectively decouples; the fast degrees of motion can be replaced by rigid constraints and the slow external degrees of motion are approximately described by the Hamiltonian system

$$\begin{aligned}
\frac{d}{d\tau}\bar{\boldsymbol{Q}} &= \boldsymbol{M}_2(\boldsymbol{0}, \bar{\boldsymbol{Q}})\,\bar{\boldsymbol{P}}\,, \\
\frac{d}{d\tau}\bar{\boldsymbol{P}} &= -\boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} V(\boldsymbol{0}, \bar{\boldsymbol{Q}}) - \boldsymbol{\nabla}_{\bar{\boldsymbol{Q}}} \frac{\bar{\boldsymbol{P}}^T\, \boldsymbol{M}_2(\boldsymbol{0}, \bar{\boldsymbol{Q}})\,\bar{\boldsymbol{P}}}{2}\,.
\end{aligned}$$

Of course, in doing so we have neglected all terms of order $\mathcal{O}(\epsilon)$ or higher in the normal form expansion. But the reduced model is justified, for example, if the motion of the slow (external) degrees of motion is ergodic and ergodicity is robust with respect to small perturbations. The decoupling of the internal and external energy over an exponentially long period of time was already conjectured by Boltzmann and Jean.

A New Derivation of the BGG-Result

The proof by Benettin, Galgani & Giorgilli (BGG) is based on transforming a Hamiltonian to normal form. Since they do not use a modified Lie-Poisson bracket, the Hamiltonian and the normal form iteration are different. Here we outline a new proof based on the normal form theory developed in Chapter 4 and the discussion in Section 4.1.1.

We first introduce an appropriate norm for functions on $\mathcal{B}_R\mathcal{K}$, $\mathcal{K} = \{\boldsymbol{0}\} \times \mathcal{V} \subset \mathrm{C}^{6N}$, $\mathcal{V} \subset \mathrm{R}^{6N-2m}$ a compact subset. Let $f : \mathcal{B}_R\mathcal{K} \subset \mathrm{C}^{2m} \times \mathrm{C}^{6N-2m} \to \mathrm{C}$ be an analytic function. Then f can be written as

$$f(\boldsymbol{x},\boldsymbol{y}) \quad = \quad \sum_{\boldsymbol{k}\in\mathbb{I}^{2m}} f_{\boldsymbol{k}}(\boldsymbol{y})\boldsymbol{x}^{\boldsymbol{k}}\,,$$

$\boldsymbol{x}\in\mathcal{B}_R\{\boldsymbol{0}\}$, $\boldsymbol{y}\in\mathcal{B}_R\mathcal{V}$. Here $\boldsymbol{x}^{\boldsymbol{k}}$ denotes the monomial

$$\boldsymbol{x}^{\boldsymbol{k}} \quad = \quad x_1^{k_1}x_2^{k_2}\cdots x_{2m}^{k_{2m}}\,,$$

$\boldsymbol{x}=(x_1,x_2,\dots,x_{2m})^T$, $\boldsymbol{k}=(k_1,k_2,\dots,k_{2m})^T$, $\boldsymbol{k}\in\mathrm{I}^{2m}$, I the set of non-negative integers. For vector-valued functions $\boldsymbol{f}:\mathcal{B}_R\mathcal{K}\subset\mathrm{C}^{2m}\times\mathrm{C}^{6N-2m}\to\mathrm{C}^n$ the same expansion is applied to

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \quad = \quad \sum_{s=1}^{n} f_s(\boldsymbol{x},\boldsymbol{y})\,\boldsymbol{e}_s\,,$$

$\boldsymbol{e}_s$, $s=1,\dots,n$, basis vectors in R^n. The corresponding terms in the Taylor expansion are denoted by $f_{\boldsymbol{k},s}(\boldsymbol{y})$. The function space P_K is the space of all (bounded) analytic functions $\boldsymbol{f}:\mathcal{B}_R\mathcal{K}\subset\mathrm{C}^{2m}\times\mathrm{C}^{6N-2m}\to\mathrm{C}^n$ whose Taylor series representation contains only monomials $\boldsymbol{x}^{\boldsymbol{k}}$ with $|\boldsymbol{k}|\le K$, i.e.

$$\boldsymbol{k}\in\mathrm{I}_K^{2m} \quad := \quad \{\boldsymbol{k}\in\mathrm{I}^{2m}:|\boldsymbol{k}|\le K\,\}\,,$$

$|\boldsymbol{k}|=k_1+\cdots+k_{2m}$. For an analytic function $\boldsymbol{f}:\mathcal{B}_R\mathcal{K}\subset\mathrm{C}^{2m}\times\mathrm{C}^{6N-2m}\to\mathrm{C}^n$, $\boldsymbol{f}\in P_K$, we also introduce the norm

$$|\boldsymbol{f}|_r \quad := \quad \sup_{\boldsymbol{y}\in\mathcal{B}_r\mathcal{V}}\sum_s\sum_{\boldsymbol{k}\in\mathbb{I}_K^{2m}}|f_{\boldsymbol{k},s}(\boldsymbol{y})|\,r^{|\boldsymbol{k}|}\,,$$

$R\ge r>0$, on $\mathcal{B}_R\mathcal{K}=\mathcal{B}_R(\{\boldsymbol{0}\}\times\mathcal{V})$.

Let us now rewrite

$$\begin{aligned}\frac{d}{dt}\boldsymbol{x} \quad &= \quad \boldsymbol{A}\boldsymbol{x}+\epsilon^{-1}\,\boldsymbol{f}(\epsilon\boldsymbol{x},\boldsymbol{y},\epsilon)\,,\\ \frac{d}{dt}\boldsymbol{y} \quad &= \quad \qquad\;\;+\,\epsilon\,\boldsymbol{g}(\epsilon\boldsymbol{x},\boldsymbol{y},\epsilon)\end{aligned}$$

as

$$\frac{d}{dt}\boldsymbol{z} \quad = \quad \boldsymbol{A}\boldsymbol{z}+\epsilon^{-1}\boldsymbol{B}(\boldsymbol{z},\epsilon)\,, \tag{1.4}$$

$\boldsymbol{z}=(\boldsymbol{x}^T,\boldsymbol{y}^T)^T\in\mathrm{C}^{6N}$. The matrix $\boldsymbol{A}$ in (1.4) is the same as before except that it has been augmented by zeros so as to be in $\mathrm{C}^{6N\times 6N}$. The corresponding homological equation

$$[\boldsymbol{A},\boldsymbol{\Delta W}]+\boldsymbol{\Delta X}^n \quad = \quad \boldsymbol{0}$$

can be solved in terms of the Taylor expansions of $\boldsymbol{\Delta W}$ and $\boldsymbol{\Delta X}^n$.

At this point we have basically reduced the problem to a special case of systems near an equilibrium point as considered in Section 4.1.1. We only have to take into account that $6N - 2m$ eigenvalues of the matrix $\boldsymbol{A}$ are zero and that the corresponding eigenvectors are determined by the variable $\boldsymbol{y}$. This implies that we do not need the Taylor expansion with respect to the variable $\boldsymbol{y}$ to solve the homological equation.

Let λ_s, $s = 1, \ldots, 6N$, denote the eigenvalues of the matrix $\boldsymbol{A}$. Note that $\boldsymbol{A}$ has $6N - 2m$ zero eigenvalues. Then we define the resonance module $\mathcal{M}_s \subset \mathrm{I}^{2m}$, $s = 1, \ldots, 6N$, by

$$\mathcal{M}_s \quad := \quad \{\boldsymbol{k} \in \mathrm{I}^{2m} : \lambda_s = \boldsymbol{\lambda}^T \boldsymbol{k}\}\,.$$

Here $\boldsymbol{\lambda}^T \boldsymbol{k}$ contains only those eigenvalues of $\boldsymbol{A}$ that are non-zero. Thus $\boldsymbol{\lambda} \in \mathrm{C}^{2m}$ and $\boldsymbol{k} \in \mathrm{I}^{2m}$. We assume that the given $\boldsymbol{\Delta X}$ is an element of P_K, $K \geq 2$ with respect to its Taylor expansion in $\boldsymbol{x}$. Then the resonant part $\boldsymbol{\Delta X}^r$ of

$$\boldsymbol{\Delta X}(\boldsymbol{x}, \boldsymbol{y}) \quad = \quad \sum_s \sum_{\boldsymbol{k} \in \mathbb{I}_K^{2m}} \Delta X_{\boldsymbol{k},s}(\boldsymbol{y})\, \boldsymbol{x}^{\boldsymbol{k}} \boldsymbol{e}_s$$

is defined by

$$\boldsymbol{\Delta X}^r(\boldsymbol{x}, \boldsymbol{y}) \quad = \quad \sum_s \sum_{\boldsymbol{k} \in \mathcal{M}_s} \Delta X_{\boldsymbol{k},s}(\boldsymbol{y})\, \boldsymbol{x}^{\boldsymbol{k}} \boldsymbol{e}_s$$

and the non-resonant part by $\boldsymbol{\Delta X}^n = \boldsymbol{\Delta X} - \boldsymbol{\Delta X}^r$. The solution of the homological equations is now given by (in terms of the Taylor series coefficients):

$$\Delta W_{\boldsymbol{k},s}(\boldsymbol{y}) \quad = \quad \frac{-\Delta X_{\boldsymbol{k},s}(\boldsymbol{y})}{\lambda_s - \boldsymbol{\lambda}^T \boldsymbol{k}}\,, \qquad \boldsymbol{k} \in \mathrm{I}_K^{2m} \setminus \mathcal{M}_s\,,$$

$\Delta W_{\boldsymbol{k},s}(\boldsymbol{y}) = 0$ for $\boldsymbol{k} \in \mathcal{M}_s$. Note that $\boldsymbol{\Delta W} \in P_K$ with

$$\boldsymbol{\Delta W}(\boldsymbol{x}, \boldsymbol{y}) \quad = \quad \sum_s \sum_{\boldsymbol{k} \in \mathbb{I}_K^{2m}} \Delta W_{\boldsymbol{k},s}(\boldsymbol{y})\, \boldsymbol{x}^{\boldsymbol{k}} \boldsymbol{e}_s\,.$$

We assume that the estimate

$$|\boldsymbol{\lambda}^T \boldsymbol{k} - \lambda_s| \quad \geq \quad \frac{\delta}{(K-1)^{d-1}}\,, \qquad \boldsymbol{k} \in \mathrm{I}_K^{2m} \setminus \mathcal{M}_s$$

holds for all s and all $K \geq 2$ as well as

$$|\boldsymbol{\lambda}^T \boldsymbol{k} - \lambda_s| \quad \geq \quad \delta\,, \qquad \boldsymbol{k} \in \mathrm{I}_K^{2m} \setminus \mathcal{M}_s\,, \qquad K = 0, 1\,.$$

Here $\delta > 0$ and $d \geq 0$ are appropriate constants depending on the non-zero eigenvalues of $\boldsymbol{A}$. Then the solution $\boldsymbol{\Delta W}$ satisfies the estimate

$$|\boldsymbol{\Delta W}|_r \quad \leq \quad \gamma^{-1}\, |\boldsymbol{\Delta X}|_r\,, \qquad \gamma = \frac{\delta}{(K-1)^{d-1}}\,, \tag{1.5}$$

$R \geq r > 0$. With this, we can now start the normal form recursion (1.10)-(1.14). For the system (1.4), the normal form recursion yields vector fields $\boldsymbol{\Delta X}_i(\boldsymbol{x},\boldsymbol{y})$ that are elements of the space P_{i+1} and

$$\epsilon^i \,\boldsymbol{\Delta X}_i(\boldsymbol{x},\boldsymbol{y}) \quad = \quad \epsilon^{-1}\boldsymbol{\Delta X}_i(\epsilon\boldsymbol{x},\boldsymbol{y},\epsilon)\,.$$

With $\mathcal{K} = \{\boldsymbol{0}\} \times \mathcal{V}$ a compact set and $\epsilon^{-1}\boldsymbol{B}(\epsilon\boldsymbol{x},\boldsymbol{y},\epsilon)$ satisfying an estimate

$$\epsilon^{-1}|\boldsymbol{B}(\epsilon)|_R \quad \leq \quad \epsilon\,,$$

we could, in principle, apply Theorem 3.1. Again we have to be a bit carefully: Since $\boldsymbol{\Delta X}_i \in P_{i+1}$, the parameter γ in (1.5) depends on the iteration index i with $K = i+1$. In other words, we have to use the estimate (4.32) of Proposition 3.1 for the error in the normal form truncation

$$\bar{\boldsymbol{Y}}_{i_*}(\bar{\boldsymbol{z}},\epsilon) \quad = \quad \boldsymbol{A}\bar{\boldsymbol{z}} + \sum_{i=1}^{i_*} \epsilon^i \boldsymbol{\Delta X}_i^r(\bar{\boldsymbol{z}})\,, \qquad (\bar{\boldsymbol{z}} \in \mathcal{B}_{R/2}\mathcal{K})\,,$$

i.e., the truncation error $\boldsymbol{T}(\bar{\boldsymbol{z}},\epsilon)$ satisfies

$$||\,\boldsymbol{T}(\epsilon)\,||_{R/2} \quad \leq \quad c_1\,\epsilon\, e^{-(c_2/\epsilon)^{1/d}}\,,$$

$c_1, c_2 > 0$ appropriate constants and d proportional to the number of non-resonant eigenvalues of $\boldsymbol{A}$. Note that the linear differential equation

$$\frac{d}{dt}\boldsymbol{z} \quad = \quad \boldsymbol{A}\boldsymbol{z}$$

commutes with all the other terms in the truncated normal form. For Hamiltonian systems this implies that the corresponding energy in the fast degrees of motion commutes with the Hamiltonian of the system in normal form truncation and is, therefore, a first integral. This implies the adiabatic invariance of the internal vibrational energy of the Hamiltonian system over an exponentially long period of time (provided the solutions stay in $\mathcal{B}_{R/2}\mathcal{K}$).

A final word on the non-resonance condition

$$|\boldsymbol{\lambda}^T\boldsymbol{k} - \lambda_s| \quad \geq \quad \frac{\delta}{(K-1)^{d-1}}\,, \qquad \boldsymbol{k} \in \mathrm{I}_K^{2m} \setminus \mathcal{M}_s\,.$$

If the eigenvalues of the matrix $\hat{\boldsymbol{A}}$ are $\pm i\omega_l$, $l = 1,\ldots,m$, then this condition is equivalent to

$$|\boldsymbol{\omega}^T\boldsymbol{k}| \quad \geq \quad \frac{\delta}{K^{d-1}}\,, \qquad \boldsymbol{k} \in \mathrm{Z}_K^m \setminus \mathcal{M}$$

with

$$\mathcal{M} \quad := \quad \{\boldsymbol{k} \in \mathrm{Z}_K^m : \boldsymbol{\omega}^T\boldsymbol{k} = 0\,\}\,.$$

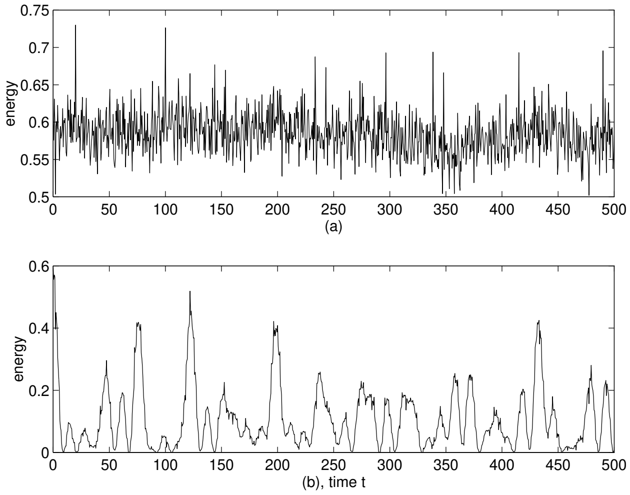


Figure 4.1: Time evolution of the total energy in all stiff degrees of freedom (a) and energy in one particular stiff degree of freedom (b).

Here $d \geq 1$ is the number of non-resonant frequencies in the spectrum of the matrix $\boldsymbol{A}$. For example, $d = 1$ for a system of identical diatomic molecules and $d = 2$ for water molecules.

Example 4.2. We consider a one-dimensional chain of $N = 21$ particles with equal mass $m = 1$. The particles interact alternating with their nearest neighbor on the one side via a stiff harmonic spring and with the nearest neighbor on the other side via a (soft) Lennard-Jones-type potential. In Fig. 4.1, it is shown that the total energy in the stiff harmonic degrees of freedom is an adiabatic invariant. We also show the evolution of the energy in one particular stiff degree of freedom. Due to the complete resonance of all the stiff degrees of freedom, this energy is not constant. For further numerical results see BENETTIN, GALGANI & GIORGILLI [13]. □

4.2 Numerical Methods

In this section we discuss three different approaches to the numerical integration of highly oscillatory mechanical systems that reduce in the limit to rigid bodies. Namely: (i) We derive an explicit symplectic method for rigid bodies. Our starting point is the following one: It is well-known that the motion of a rigid body is characterized by the superposition of a translation and a rotation and that the differential equations describing the translation of the center of mass and the rotation about the center of mass are Hamiltonian. Based on this Hamiltonian formulation, we derive an explicit symplectic scheme for rigid bodies moving under the influence of external conservative forces. The basic ideas were first published by the author in [101]. The special splitting method for water molecules is new. (ii) To take finite stiffness effects into account, the concept of soft-constraints was introduced in [126],[97],[99]. Here we suggest a different approach that modifies the force field instead of the constraint functions. An effective implementation is given. (iii) Inspired by the paper [42], a projected multiple-time-stepping method is suggested for the integration of mechanical systems with highly oscillatory internal vibrations. In contrast to formulations using constraints, multiple-time-stepping methods resolve the high frequency motions with a small step-size while the slowly varying components of the force field ("slow forces") are integrated with a much larger step-size. This is the standard multiple-time-stepping approach. To avoid resonance problems, projected multiple-time-stepping uses modified "slow forces" that are obtained from the original "slow forces" by projecting away the highly oscillatory solution components. This projection is done in a way similar to the SHAKE projection step [106].

4.2.1 Symplectic Integration of Rigid Bodies

Equations of Motion for a Single Rigid Body

In this subsection we give the Hamiltonian formulation for the equations of motion for a single rigid body moving in R^3 under the influence of an external force. This result naturally extends to unconstrained systems of rigid bodies moving in an external force field.

Consider a rigid body free to move in R^3. A reference configuration $\mathfrak{B}$ of the body is the closure of an open set in R^3. Points in $\mathfrak{B}$, denoted by $\boldsymbol{\xi} \in \mathfrak{B}$, are called material points. A configuration of $\mathfrak{B}$ is a mapping $\boldsymbol{\phi} : \mathfrak{B} \to \mathrm{R}^3$ which is smooth, orientation preserving, and invertible on its image. The points of the target space R^3 of $\boldsymbol{\phi}$ are called spatial points and denoted by $\boldsymbol{x} \in \mathrm{R}^3$. A motion of $\mathfrak{B}$ is a time dependent family of configurations, written $\boldsymbol{x}(t) = \boldsymbol{\phi}(\boldsymbol{\xi}, t) = \boldsymbol{\phi}_t(\boldsymbol{\xi})$. Rigidity of the body means that the distance between points of the body are fixed as the body moves. Let us assume that the center of mass of $\mathfrak{B}$ is at the origin and let us denote the motion of the center of mass by $\boldsymbol{q}(t) = \boldsymbol{\phi}_t(\mathbf{0})$. Then any motion of $\mathfrak{B}$ is the superposition of $\boldsymbol{q}(t)$ and a rotation about the center of mass; i.e.

$$\boldsymbol{x}(t) \quad = \quad \boldsymbol{Q}(t)\boldsymbol{\xi} + \boldsymbol{q}(t) \tag{2.6}$$

where $\boldsymbol{Q}(t) \in SO(3)$.

Let $m \in \mathrm{R}^+$ be the total mass of the rigid body and $\boldsymbol{M} \in \mathrm{R}^{3\times 3}$ the inertial tensor which, for simplicity, we assume to be diagonal. Furthermore, let a conservative force

$$\boldsymbol{F}(\boldsymbol{x}_o(t)) \quad = \quad -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{x}_o(t)) \tag{2.7}$$

$V : \mathrm{R}^3 \to \mathrm{R}$, act on the rigid body where

$$\boldsymbol{x}_o(t) \quad = \quad \boldsymbol{Q}(t)\boldsymbol{\xi}_o + \boldsymbol{q}(t)$$

and $\boldsymbol{\xi}_o \in \mathfrak{B}$ is a fixed point in $\mathfrak{B}$. Upon introducing canonical momenta $\boldsymbol{p} = m d\boldsymbol{q}/dt$ and $\boldsymbol{P} = d\boldsymbol{Q}/dt\boldsymbol{J}$ where the diagonal matrix $\boldsymbol{J} \in \mathrm{R}^{3\times 3}$ is defined by

$$M_{ii}^{-1} = \frac{1}{2}\sum_{i\neq j} J_{jj}^{-1},$$

the motion of the rigid body in R^3 is Hamiltonian with the Hamiltonian function

$$\begin{aligned} H_c(\boldsymbol{q},\boldsymbol{p},\boldsymbol{Q},\boldsymbol{P}) &= \frac{1}{2}(\boldsymbol{p}^T m^{-1}\boldsymbol{p}) + \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\boldsymbol{J}^{-1}\boldsymbol{P}^T) + \\ &= \qquad + V(\boldsymbol{q},\boldsymbol{Q}) + \frac{1}{2}\mathrm{tr}\left(\{\boldsymbol{Q}^T\boldsymbol{Q}-\boldsymbol{I}\}\boldsymbol{\Lambda}\right) \end{aligned} \tag{2.8}$$

where tr denotes the trace operator and $\boldsymbol{\Lambda} \in \mathrm{R}^{3\times 3}$ is a symmetric matrix implicitly determined by the holonomic constraint $\boldsymbol{Q} \in SO(3)$, i.e.

$$\mathbf{0} = \boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I}\,. \tag{2.9}$$

[82],[84]. Thus the motion for the center of mass is simply given by

$$\frac{d}{dt}\boldsymbol{q} = m^{-1}\boldsymbol{p} \tag{2.10}$$

$$\frac{d}{dt}\boldsymbol{p} = -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q},\boldsymbol{Q}) \tag{2.11}$$

where $V(\boldsymbol{q},\boldsymbol{Q})$ is defined by

$$V(\boldsymbol{q},\boldsymbol{Q}) = V(\boldsymbol{Q}\boldsymbol{\xi}_o + \boldsymbol{q})\,.$$

and the corresponding equations of motion for the variable $\boldsymbol{Q}$ are

$$\frac{d}{dt}\boldsymbol{Q} = \boldsymbol{P}\boldsymbol{J}^{-1} \tag{2.12}$$

$$\frac{d}{dt}\boldsymbol{P} = -\boldsymbol{\nabla}_{\boldsymbol{Q}} V(\boldsymbol{q},\boldsymbol{Q}) - \boldsymbol{Q}\boldsymbol{\Lambda} \tag{2.13}$$

$$\mathbf{0} = \boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I} \tag{2.14}$$

which define a Hamiltonian vector field on the phase space

$$\mathcal{M} = \{(\boldsymbol{Q},\boldsymbol{P}) : \boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I} = \mathbf{0},\, \boldsymbol{Q}^T\boldsymbol{P}\boldsymbol{J}^{-1} + \boldsymbol{J}^{-1}\boldsymbol{P}^T\boldsymbol{Q} = \mathbf{0}\}$$

Note that $\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q},\boldsymbol{Q}) = -\boldsymbol{F}(\boldsymbol{q},\boldsymbol{Q})$ and $\boldsymbol{\nabla}_{\boldsymbol{Q}} V(\boldsymbol{q},\boldsymbol{Q}) = -\boldsymbol{F}(\boldsymbol{q},\boldsymbol{Q})\boldsymbol{\xi}_o^T$.

Now $\mathcal{M} \neq T^*SO(3)$ in general. However, as shown by McLachlan and Scovel [84], the motion on $T^*SO(3)$ can be obtained from the motion on $\mathcal{M}$ by putting $\tilde{\boldsymbol{Q}} = \boldsymbol{Q}$ and $\tilde{\boldsymbol{P}} = \mathrm{Pr}(\boldsymbol{Q})\boldsymbol{P}$ where $\mathrm{Pr}(\boldsymbol{Q})$ is the orthogonal projector onto $T_Q^*SO(3) \subset \mathrm{R}^{3\times 3}$.

To make the equations (2.12)-(2.14) more transparent, let us rewrite them on $SO(3) \times \mathfrak{so}(3)^*$. Here $\mathfrak{so}(3)^*$ denotes the dual of the Lie algebra of skew-symmetric matrices [119]. This transformation can be achieved by introducing the body angular momentum $\boldsymbol{\Pi} = \boldsymbol{Q}^T\boldsymbol{P} \in \mathfrak{so}(3)^*$ [82]. Using the standard isomorphism between R^3 and $\mathfrak{so}(3)^*$ denoted by skew : $\mathrm{R}^3 \to \mathfrak{so}(3)^*$, we identify $\boldsymbol{\Pi} \in \mathfrak{so}(3)^*$ with $\boldsymbol{\pi} \in \mathrm{R}^3$, $[\boldsymbol{Q}^T\boldsymbol{F}(\boldsymbol{q},\boldsymbol{Q})\boldsymbol{\xi}_o^T - \boldsymbol{\xi}_o\boldsymbol{F}(\boldsymbol{q},\boldsymbol{Q})^T\boldsymbol{Q}]/2$ with $\boldsymbol{Q}^T\boldsymbol{F}(\boldsymbol{q},\boldsymbol{Q}) \times \boldsymbol{\xi}_o$, and $[\boldsymbol{J}^{-1}\boldsymbol{\Pi} + \boldsymbol{\Pi}\boldsymbol{J}^{-1}]/2$ with $\boldsymbol{M}^{-1}\boldsymbol{\pi}$ where $\boldsymbol{M}$ is the inertial tensor. Then we obtain the equations

$$\frac{d}{dt}\boldsymbol{\pi} = \boldsymbol{\pi} \times \boldsymbol{M}^{-1}\boldsymbol{\pi} + \{\boldsymbol{Q}^T\boldsymbol{F}(\boldsymbol{q},\boldsymbol{Q})\} \times \boldsymbol{\xi}_o \tag{2.15}$$

and

$$\frac{d}{dt}\boldsymbol{Q} = \boldsymbol{Q}\,\mathrm{skew}\,(\boldsymbol{M}^{-1}\boldsymbol{\pi}) \tag{2.16}$$

Note that for $\boldsymbol{F} = \boldsymbol{0}$, (2.15) becomes the standard Euler equation for the free rigid body [82].

In the following subsection we will discuss numerical methods for the symplectic integration of the constrained Hamiltonian system corresponding to (2.8)-(2.9).

Symplectic Integration of a Single Rigid Body

A Hamiltonian system on $\mathrm{R}^n \times \mathrm{R}^n$ with holonomic constraints of the form $\boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{0}$, where $\boldsymbol{g} : \mathrm{R}^n \to \mathrm{R}^m$ is a smooth map with full rank Jacobian in the neighborhood of its zero-set $\boldsymbol{g}^{-1}(\boldsymbol{0})$, is characterized by the equations

$$\frac{d}{dt}\boldsymbol{q} = \boldsymbol{\nabla}_{\boldsymbol{p}} H(\boldsymbol{q}, \boldsymbol{p}) \tag{2.17}$$

$$\frac{d}{dt}\boldsymbol{p} = -\boldsymbol{\nabla}_{\boldsymbol{q}} H(\boldsymbol{q}, \boldsymbol{p}) - \boldsymbol{G}(\boldsymbol{q})^T \boldsymbol{\lambda} \tag{2.18}$$

$$\boldsymbol{0} = \boldsymbol{g}(\boldsymbol{q}) \tag{2.19}$$

where $H : \mathrm{R}^n \times \mathrm{R}^n \to \mathrm{R}$ is the Hamiltonian of the unconstrained system, $\boldsymbol{G}(\boldsymbol{q}) = \boldsymbol{\partial}_{\boldsymbol{q}} \boldsymbol{g}(\boldsymbol{q})$, and

$$H_c = H + \boldsymbol{\lambda}^T \boldsymbol{g}$$

the Hamiltonian of the constrained system. The solutions of (2.17)-(2.19) define a symplectic flow [6] on the constraint manifold

$$\mathcal{M} = \{(\boldsymbol{q}, \boldsymbol{p}) : \boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{0},\, \boldsymbol{G}(\boldsymbol{q})\boldsymbol{\nabla}_{\boldsymbol{p}} H(\boldsymbol{q}, \boldsymbol{p}) = \boldsymbol{0}\}$$

Recently it has been shown by Jay [65] and Reich [98] that for arbitrary Hamiltonian functions H there exist discretizations of (2.17)-(2.19) which preserve the constraint manifold $\mathcal{M}$ and are symplectic. Even more recently, McLachlan & Scovel [84] extended this result to the symplectic integration of the corresponding constrained system on the cotangent manifold

$$T^*\mathcal{N} = \{(\boldsymbol{q}, \boldsymbol{p}) : \boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{0},\, \boldsymbol{G}(\boldsymbol{q})\boldsymbol{p} = \boldsymbol{0}\}$$

In case that the Hamiltonian H is separable, i.e. of the form

$$H(\boldsymbol{q}, \boldsymbol{p}) = T(\boldsymbol{p}) + V(\boldsymbol{q}),$$

we can, e.g., use the scheme RATTLE [3] which is a generalization of the Verlet scheme for unconstrained systems. It was shown to be symplectic and constraint-preserving by Leimkuhler & Skeel [76].

Since the Hamiltonian (2.8) is separable, the RATTLE discretization can be directly applied to the symplectic integration of the corresponding constrained Hamiltonian system. To see this, we have to use the correspondences $\boldsymbol{q} \mapsto (\boldsymbol{q}, \boldsymbol{Q})$ and $\boldsymbol{p} \mapsto (\boldsymbol{p}, \boldsymbol{P})$ and the constraint function $\boldsymbol{g}$ is given by (2.9). Thus we obtain the following algorithm:

SEMI-EXPLICIT RIGID BODY INTEGRATOR

Step 1.

$$\begin{aligned} \boldsymbol{q}_{n+1} &= \boldsymbol{q}_n + \Delta t m^{-1}\boldsymbol{p}_{n+1/2} \\ \boldsymbol{p}_{n+1/2} &= \boldsymbol{p}_n + (\Delta t/2)\boldsymbol{F}(\boldsymbol{q}_n, \boldsymbol{Q}_n) \end{aligned}$$

Step 2.

$$\begin{aligned} \boldsymbol{Q}_{n+1} &= \boldsymbol{Q}_n + \Delta t \boldsymbol{P}_{n+1/2}\boldsymbol{J}^{-1} \\ \boldsymbol{P}_{n+1/2} &= \boldsymbol{P}_n + (\Delta t/2)\left[\boldsymbol{F}(\boldsymbol{q}_n, \boldsymbol{Q}_n)\boldsymbol{\xi}_o^T + \boldsymbol{Q}_n\boldsymbol{\Lambda}_n\right] \\ \boldsymbol{0} &= \boldsymbol{Q}_{n+1}^T\boldsymbol{Q}_{n+1} - \boldsymbol{I} \end{aligned}$$

Step 3.

$$\begin{aligned} \boldsymbol{P}_{n+1} &= \boldsymbol{P}_{n+1/2} + (\Delta t/2)\left[\boldsymbol{F}(\boldsymbol{q}_{n+1}, \boldsymbol{Q}_{n+1})\boldsymbol{\xi}_o^T + \boldsymbol{Q}_{n+1}\boldsymbol{\mu}_{n+1}\right] \\ 0 &= \boldsymbol{Q}_{n+1}^T\boldsymbol{P}_{n+1} + \boldsymbol{P}_{n+1}^T\boldsymbol{Q}_{n+1} \end{aligned}$$

Step 4.

$$\boldsymbol{p}_{n+1} = \boldsymbol{p}_{n+1/2} + (\Delta t/2)\boldsymbol{F}(\boldsymbol{q}_{n+1}, \boldsymbol{Q}_{n+1})$$

where $\boldsymbol{F}(\boldsymbol{q}, \boldsymbol{Q}) = \boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q} + \boldsymbol{Q}\boldsymbol{\xi}_o)$.

Step 2 requires the solution of a nonlinear system of equations in 6 variables (since Λ is assumed to be symmetric). This can be avoided by applying recent results by McLachlan [83] and Reich [94] on the explicit symplectic integration of the Euler equations on the $\mathfrak{so}(3)^*$. Specifically:

Step 2 and 3 represent a second order symplectic integrator for the Hamiltonian

$$H_r(\boldsymbol{Q}, \boldsymbol{P}) = \frac{1}{2}\mathrm{tr}\,(\boldsymbol{P}\boldsymbol{J}^{-1}\boldsymbol{P}^T) + V(\boldsymbol{q}, \boldsymbol{Q}) + \frac{1}{2}\mathrm{tr}\,([\boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I}]\boldsymbol{\Lambda})$$

where $\boldsymbol{q}$ is treated as a parameter. Upon rewriting H_r as

$$H_r(\boldsymbol{Q}, \boldsymbol{P}) = H_e(\boldsymbol{Q}, \boldsymbol{P}, \boldsymbol{\Lambda}) + V(\boldsymbol{q}, \boldsymbol{Q}) \tag{2.20}$$

with

$$H_e(\boldsymbol{Q}, \boldsymbol{P}, \boldsymbol{\Lambda}) = \frac{1}{2}\mathrm{tr}\,(\boldsymbol{P}\boldsymbol{J}^{-1}\boldsymbol{P}^T) + \frac{1}{2}\mathrm{tr}\,([\boldsymbol{Q}^T\boldsymbol{Q} - \boldsymbol{I}]\boldsymbol{\Lambda})$$

we see that H_e is the Hamiltonian of a free rigid body fixed at the origin and the potential V corresponds to an external force acting on the body. Using the notation introduced in the previous subsection, the equations of motion of a free rigid body in terms of the body angular momentum are

$$\frac{dt}{dt}\boldsymbol{\pi} = \boldsymbol{\pi} \times \boldsymbol{M}^{-1}\boldsymbol{\pi} \tag{2.21}$$

where the Hamiltonian is now the kinetic energy of the body; i.e.

$$T(\boldsymbol{\pi}) \;=\; \frac{1}{2}\boldsymbol{\pi}^T \boldsymbol{M}^{-1}\boldsymbol{\pi}$$

As pointed out by McLachlan [83] and Reich [94], (2.21) can be integrated numerically by rewriting (2.21) as the sum of three Euler equations $d\boldsymbol{\pi}/dt = \boldsymbol{\pi} \times \nabla T_i(\boldsymbol{\pi})$ where $T_i(\boldsymbol{\pi}) = \frac{1}{2}\pi_i^2 M_{ii}^{-1}$. For each of these equations we have $\pi_i = const.$ and therefore each equation can be solved exactly. The solutions, denoted by

$$\boldsymbol{\Phi}_{t,T_i}(\boldsymbol{\pi}) \;=\; \mathbf{exp}(t\boldsymbol{X}_{T_i}) \cdot \boldsymbol{\pi}\,,$$

are rotations about the π_i-axis with constant angular velocity $\pi_i M_{ii}^{-1}$, i.e. $\mathbf{exp}(t\boldsymbol{X}_{T_i}) \in SO(3)$. For example, for $i = 1$, we have to solve the system of linear differential equations

$$\begin{aligned}
\dot{\pi}_1 &= 0 \\
\dot{\pi}_2 &= +\pi_1 M_{11}^{-1}\,\pi_3 \\
\dot{\pi}_3 &= -\pi_1 M_{11}^{-1}\,\pi_2
\end{aligned}$$

By applying the Baker-Campbell-Hausdorff formula [119], one can show that the scheme

$$\boldsymbol{\pi}_{n+1} \;=\; \boldsymbol{A}_{\Delta t}(\boldsymbol{\pi}_n)\,\boldsymbol{\pi}_n$$

with

$$\begin{aligned}
\boldsymbol{A}_{\Delta t}(\boldsymbol{\pi}) \;=\; & \mathbf{exp}(\Delta t/2\boldsymbol{X}_{T_1}) \cdot \mathbf{exp}(\Delta t/2\boldsymbol{X}_{T_2}) \cdot \mathbf{exp}(\Delta t\boldsymbol{X}_{T_3}) \cdot \\
& \cdot \mathbf{exp}(\Delta t/2\boldsymbol{X}_{T_2}) \cdot \mathbf{exp}(\Delta t/2\boldsymbol{X}_{T_1})
\end{aligned}$$

is of second order in the step-size Δt and naturally preserves the Lie-Poisson structure of the Euler equation [82]. Note that each $\boldsymbol{\Phi}_{\Delta t,T_i}$ requires the evaluation of a sine and cosine function. This can be avoided by using the Cayley transformation to obtain an approximate rotation matrix $\boldsymbol{R}_i(\Delta t)$ with

$$\mathbf{exp}(\Delta t\boldsymbol{X}_{T_i}) \;=\; \boldsymbol{R}_i(\Delta t) + \mathcal{O}(\Delta t^3)\,.$$

Let us now return to the symplectic integration of the corresponding Hamiltonian system on $T^*SO(3)$. We make use of the same splitting of the kinetic energy $T(\boldsymbol{\pi})$. But now we have to solve for each $T_i(\boldsymbol{\pi})$ the system of linear differential equations

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{\pi} &= \boldsymbol{\pi} \times \boldsymbol{\nabla}_{\boldsymbol{\pi}} T_i(\boldsymbol{\pi}) \\
\frac{d}{dt}\boldsymbol{Q} &= \boldsymbol{Q}\,\mathrm{skew}\left(\boldsymbol{\nabla}_{\boldsymbol{\pi}} T_i(\boldsymbol{\pi})\right)
\end{aligned}$$

Composing the single steps as done above for the Euler equation, we finally obtain the second-order symplectic scheme

$$\boldsymbol{Q}_{n+1} \;=\; \boldsymbol{Q}_n\,[\boldsymbol{A}_{\Delta t}(\boldsymbol{\pi}_n)]^T \qquad (2.22)$$

$$\boldsymbol{\pi}_{n+1} \;=\; \boldsymbol{A}_{\Delta t}(\boldsymbol{\pi}_n)\,\boldsymbol{\pi}_n \qquad (2.23)$$

with $\boldsymbol{A}_{\Delta t}$ as above.

The motion due to the potential V in (2.20) is given by the differential equation

$$\frac{d}{dt}\boldsymbol{\pi} = [\boldsymbol{Q}^T \boldsymbol{F}(\boldsymbol{q}, \boldsymbol{Q})] \times \boldsymbol{\xi}_o \tag{2.24}$$

$$\frac{d}{dt}\boldsymbol{Q} = \boldsymbol{0} \tag{2.25}$$

This differential equation can be solved exactly and Step 2 and 3 in the previous algorithm can now be replaced by a proper composition of (2.22)-(2.23) and the exact time-Δt-flow of the differential equation (2.24)-(2.25). Thus we obtain the following explicit algorithm:

<u>Explicit Rigid Body Integrator</u>

Step 1.

$$\begin{aligned} \boldsymbol{q}_{n+1} &= \boldsymbol{q}_n + \Delta t m^{-1} \boldsymbol{p}_{n+1/2} \\ \boldsymbol{p}_{n+1/2} &= \boldsymbol{p}_n + (\Delta t/2)\boldsymbol{F}(\boldsymbol{q}_n, \boldsymbol{Q}_n) \end{aligned}$$

Step 2.

$$\bar{\boldsymbol{\pi}}_n = \boldsymbol{\pi}_n + (\Delta t/2)[\boldsymbol{Q}_n^T \boldsymbol{F}(\boldsymbol{q}_n, \boldsymbol{Q}_n)] \times \boldsymbol{\xi}_o$$

$$\begin{aligned} \boldsymbol{Q}_{n+1} &= \boldsymbol{Q}_n [\boldsymbol{A}_{\Delta t}(\bar{\boldsymbol{\pi}}_n)]^T \\ \bar{\boldsymbol{\pi}}_{n+1} &= \boldsymbol{A}_{\Delta t}(\bar{\boldsymbol{\pi}}_n)\,\bar{\boldsymbol{\pi}}_n \end{aligned}$$

$$\boldsymbol{\pi}_{n+1} = \bar{\boldsymbol{\pi}}_{n+1} + (\Delta t/2)[\boldsymbol{Q}_{n+1}^T \boldsymbol{F}(\boldsymbol{q}_{n+1}, \boldsymbol{Q}_{n+1})] \times \boldsymbol{\xi}_o$$

Step 3.

$$\boldsymbol{p}_{n+1} = \boldsymbol{p}_{n+1/2} + (\Delta t/2)\boldsymbol{F}(\boldsymbol{q}_{n+1}, \boldsymbol{Q}_{n+1})$$

If necessary, $\boldsymbol{P}_{n+1}$ can be computed from $\boldsymbol{Q}_{n+1}$ and $\boldsymbol{\pi}_{n+1}$ by means of

$$\boldsymbol{P}_{n+1} = \boldsymbol{Q}_{n+1}\boldsymbol{\Pi}_{n+1}$$

where, using the standard isomorphism between R^3 and $\mathfrak{so}(3)^*$, we transform $\boldsymbol{\pi}_{n+1} \in \mathrm{R}^3$ back into $\boldsymbol{\Pi}_{n+1} \in \mathfrak{so}(3)^*$. Let us denote the resulting scheme by

$$\begin{pmatrix} \boldsymbol{q}_{n+1} \\ \boldsymbol{Q}_{n+1} \\ \boldsymbol{p}_{n+1} \\ \boldsymbol{P}_{n+1} \end{pmatrix} = \boldsymbol{\Psi}_{\Delta t} \begin{pmatrix} \boldsymbol{q}_n \\ \boldsymbol{Q}_n \\ \boldsymbol{p}_n \\ \boldsymbol{P}_n \end{pmatrix} \tag{2.26}$$

Note that this scheme is now explicit in $\boldsymbol{Q}$ and $\boldsymbol{P}$. The scheme is of second order. This can be seen from the time reversibility of the scheme. Furthermore, by construction, the scheme is symplectic (also if the exact rotations are replaced by the Cayley transformations) and preserves the constraint $\boldsymbol{Q} \in SO(3)$. Systems of rigid bodies can be treated by applying the scheme to each rigid body.

Remark 4.3. Both algorithms have been implemented in codes for simulation of molecular systems. See A. Kol, B. Laird & B. Leimkuhler [70] and A. Dullweber, B. Leimkuhler & R. McLachlan [34] for details. □

Water Molecules

A water molecule is typically modeled as a planar rigid body. For planar rigid bodies the moments of inertia M_{ii}, $i = 1, 2, 3$, satisfy

$$M_{33} = M_{11} + M_{22}$$

(provided the body's reference configuration is appropriately placed in the x-y plane). Thus the rotational kinetic energy is

$$\begin{aligned}
T(\boldsymbol{\pi}) &= \sum_i M_{ii}^{-1} \pi_i^2 \,, \\
&= \frac{M_{22}(M_{11}+M_{22})\pi_1^2 + M_{11}(M_{11}+M_{22})\pi_2^2 + M_{11}M_{22}\pi_3^2}{M_{11}M_{22}(M_{11}+M_{22})} \,, \\
&= \frac{M_{11}M_{22}\boldsymbol{\pi}^T\boldsymbol{\pi}}{M_{11}M_{22}(M_{11}+M_{22})} + \frac{M_{22}\pi_1^2}{M_{11}(M_{11}+M_{22})} + \frac{M_{11}\pi_2^2}{M_{22}(M_{11}+M_{22})} \,, \\
&= \frac{M_{22}}{M_{11}M_{33}}\pi_1^2 + \frac{M_{11}}{M_{22}M_{33}}\pi_2^2 + \frac{1}{M_{33}}\boldsymbol{\pi}^T\boldsymbol{\pi} \,, \\
&=: \tilde{T}_1(\boldsymbol{\pi}) + \tilde{T}_2(\boldsymbol{\pi}) + \tilde{T}_3(\boldsymbol{\pi}) \,.
\end{aligned}$$

Each of the corresponding systems

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{\pi} &= \boldsymbol{\pi} \times \boldsymbol{\nabla}_{\boldsymbol{\pi}} \tilde{T}_i(\boldsymbol{\pi}) \,, \\
\frac{d}{dt}\boldsymbol{Q} &= \boldsymbol{Q}\, \mathrm{skew}\,(\boldsymbol{\nabla}_{\boldsymbol{\pi}} \tilde{T}_i(\boldsymbol{\pi})) \,,
\end{aligned}$$

$i = 1, 2, 3$ can be solved exactly. In particular, $\boldsymbol{\pi}(t) = const.$ for the Hamiltonian $\tilde{T}_3$! A second-order integrator can be obtained as before by the appropriate composition of the corresponding flow maps. Again, the exact rotations can be replaced by the corresponding Cayley transformations. Upon using this in the explicit rigid body integrator, a very efficient method for simulation of water has been derived.

4.2.2 Soft Constraints and Modified Force Fields

The approximation of a flexible system by a constrained system (rigid bodies)

$$\begin{aligned} \frac{d}{d\tau}\bar{\boldsymbol{q}} &= \boldsymbol{M}^{-1}\bar{\boldsymbol{p}}\,, \\ \frac{d}{d\tau}\bar{\boldsymbol{p}} &= -\boldsymbol{\nabla}_{\bar{\boldsymbol{q}}}V(\bar{\boldsymbol{q}}) - \boldsymbol{\nabla}_{\bar{\boldsymbol{q}}}\boldsymbol{g}(\bar{\boldsymbol{q}})^T\boldsymbol{\lambda}\,, \\ \boldsymbol{0} &= \boldsymbol{g}(\bar{\boldsymbol{q}}) \end{aligned}$$

neglects contributions of order $\mathcal{O}(\epsilon^2)$ [97]. The idea of soft constraints is to partially include those terms (see Brooks, Zhou & Reich [126], Reich [97], Reich [99]). (Note that this modification is close to what is discussed by Kopell in [71].) However, the modified constraint functions are costly to implement. Here we suggest a different approach: Instead of modifying the constraint functions we modify the potential energy function such that the modified potential energy function is more accurate then $V(\bar{\boldsymbol{q}})$ for $\epsilon > 0$. The basic idea is to introduce a near to the identity transformation

$$\tilde{\boldsymbol{q}} \;:=\; \boldsymbol{\phi}(\bar{\boldsymbol{q}})$$

by means of

$$\begin{aligned} \tilde{\boldsymbol{q}} &:= \bar{\boldsymbol{q}} + \boldsymbol{M}^{-1}\,\boldsymbol{\nabla}_{\bar{\boldsymbol{q}}}\boldsymbol{g}(\bar{\boldsymbol{q}})\,\boldsymbol{\mu}\,, \\ \boldsymbol{0} &= \boldsymbol{\partial}_{\bar{\boldsymbol{q}}}\boldsymbol{g}(\bar{\boldsymbol{q}})\,\boldsymbol{M}^{-1}\left[\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}V(\tilde{\boldsymbol{q}}) + \epsilon^{-2}\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}\boldsymbol{g}(\tilde{\boldsymbol{q}})\,\boldsymbol{g}(\tilde{\boldsymbol{q}})\right]. \end{aligned} \tag{2.27}$$

Note that $\bar{\boldsymbol{q}}$ satisfies $\boldsymbol{g}(\bar{\boldsymbol{q}}) = \boldsymbol{0}$. Thus $||\tilde{\boldsymbol{q}} - \bar{\boldsymbol{q}}|| = \mathcal{O}(\epsilon^2)$.

The modified potential energy function is defined by

$$\begin{aligned} W(\tilde{\boldsymbol{q}}) &:= V(\tilde{\boldsymbol{q}}) + \frac{1}{2\epsilon^2}\boldsymbol{g}(\tilde{\boldsymbol{q}})^T\boldsymbol{g}(\tilde{\boldsymbol{q}})\,, \\ &= W(\boldsymbol{\phi}(\bar{\boldsymbol{q}}))\,. \end{aligned}$$

The mapping $\boldsymbol{\phi}$ can be understood as a partial approximation to the $\mathcal{O}(\epsilon^2)$ term in the normal form expansion. In other words, instead of enforcing

$$\boldsymbol{r} \;:=\; \boldsymbol{g}(\boldsymbol{q}) \;=\; \boldsymbol{0}\,,$$

we note that the unconstrained system will oscillate about the minimum of the total potential energy with respect to the variable $\boldsymbol{r}$, i.e., will oscillate about the manifold defined by

$$\begin{aligned} \boldsymbol{\nabla}_{\boldsymbol{r}}W(\tilde{\boldsymbol{q}}) &\approx \boldsymbol{\partial}_{\bar{\boldsymbol{q}}}\boldsymbol{g}(\bar{\boldsymbol{q}})\,\boldsymbol{M}^{-1}\left[\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}V(\tilde{\boldsymbol{q}}) + \epsilon^{-2}\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}\boldsymbol{g}(\tilde{\boldsymbol{q}})\,\boldsymbol{g}(\tilde{\boldsymbol{q}})\right], \\ &= \boldsymbol{0} \end{aligned}$$

(upon neglecting velocity dependent contributions). But this is what has been used in (2.27). The trick is to keep the evaluation of the corresponding gradient

$$\boldsymbol{\nabla}_{\bar{\boldsymbol{q}}}W(\tilde{\boldsymbol{q}}) \;=\; [\boldsymbol{\partial}_{\bar{\boldsymbol{q}}}\boldsymbol{\phi}(\bar{\boldsymbol{q}})]^T\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}\left[\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}V(\tilde{\boldsymbol{q}}) + \epsilon^{-2}\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}\boldsymbol{g}(\tilde{\boldsymbol{q}})\,\boldsymbol{g}(\tilde{\boldsymbol{q}})\right]$$

cheap. It is indeed easily checked that the evaluation of the gradient does not require the computation of the Hessian of V but only needs the computation of the second derivative of $\boldsymbol{g}$. In other words

$$\begin{aligned} \boldsymbol{d\tilde{q}} &:= \boldsymbol{\partial_{\bar{q}}\phi(\bar{q})\, d\bar{q}}\,, \\ &= \boldsymbol{d\bar{q}} + \boldsymbol{M}^{-1}\,\boldsymbol{\nabla_{\bar{q}} g(\bar{q})\, d\mu} + \boldsymbol{M}^{-1}\sum_i \mu_i\,\boldsymbol{\partial^2_{\bar{q}} g}^i(\boldsymbol{\bar{q}})\,\boldsymbol{d\bar{q}}\,, \end{aligned}$$

but $\boldsymbol{d\mu}$ is not needed because of

$$\boldsymbol{\partial_{\bar{q}} g(\bar{q})\, M}^{-1}\left[\boldsymbol{\nabla_{\tilde{q}}} V(\boldsymbol{\tilde{q}}) + \epsilon^{-2}\boldsymbol{\nabla_{\tilde{q}} g(\tilde{q})\, g(\tilde{q})}\right] = \boldsymbol{0}\,.$$

RATTLE Algorithm with Modified Force Field

Step 1.

$$\begin{aligned} \boldsymbol{\tilde{q}}_n &= \boldsymbol{\phi}(\boldsymbol{\bar{q}}_n)\,, \\ \boldsymbol{F}_n &= -[\boldsymbol{\partial_{\bar{q}}\phi}(\boldsymbol{\bar{q}}_n)]^T\,\boldsymbol{\nabla_{\tilde{q}}} W(\boldsymbol{\tilde{q}}_n)\,. \end{aligned}$$

Step 2.

$$\begin{aligned} \boldsymbol{\bar{q}}_{n+1} &= \boldsymbol{\bar{q}}_n + \Delta t \boldsymbol{M}^{-1}\boldsymbol{\bar{p}}_{n+1/2}\,, \\ \boldsymbol{\bar{p}}_{n+1/2} &= \boldsymbol{\bar{p}}_n + \frac{\Delta t}{2}\left[\boldsymbol{F}_n - \boldsymbol{\nabla_{\bar{q}} g}(\boldsymbol{\bar{q}}_n)\,\boldsymbol{\lambda}_n\right]\,, \\ \boldsymbol{0} &= \boldsymbol{g}(\boldsymbol{\bar{q}}_{n+1})\,. \end{aligned}$$

Step 3.

$$\begin{aligned} \boldsymbol{\tilde{q}}_{n+1} &= \boldsymbol{\phi}(\boldsymbol{\bar{q}}_{n+1})\,, \\ \boldsymbol{F}_{n+1} &= -[\boldsymbol{\partial_{\bar{q}}\phi}(\boldsymbol{\bar{q}}_{n+1})]^T\,\boldsymbol{\nabla_{\tilde{q}}} W(\boldsymbol{\tilde{q}}_{n+1})\,. \end{aligned}$$

Step 4.

$$\begin{aligned} \boldsymbol{\bar{p}}_{n+1} &= \boldsymbol{\bar{p}}_{n+1/2} + \frac{\Delta t}{2}\left[\boldsymbol{F}_{n+1} - \boldsymbol{\nabla_{\bar{q}} g}(\boldsymbol{\bar{q}}_{n+1})\,\boldsymbol{\lambda}_{n+1}\right]\,, \\ \boldsymbol{0} &= \boldsymbol{\partial_{\bar{q}} g}(\boldsymbol{\bar{q}}_{n+1})\,\boldsymbol{M}^{-1}\,\boldsymbol{\bar{p}}_{n+1}\,. \end{aligned}$$

Note that the modified forces could also be used in the rigid body integrators described in the previous section. We also like to point out that the modified force field requires additional force field evaluations. However, these additional force evaluations can be restricted to nearest neighborhood interactions. For a more detailed description of the modified force field approach see [103]. Below we report about simulation results for water molecules and fluctuating charge force fields.

Example 4.3. We simulated the collision of two water molecules. The force field was taken from the CHARMM package [27]. Initial conditions were chosen such

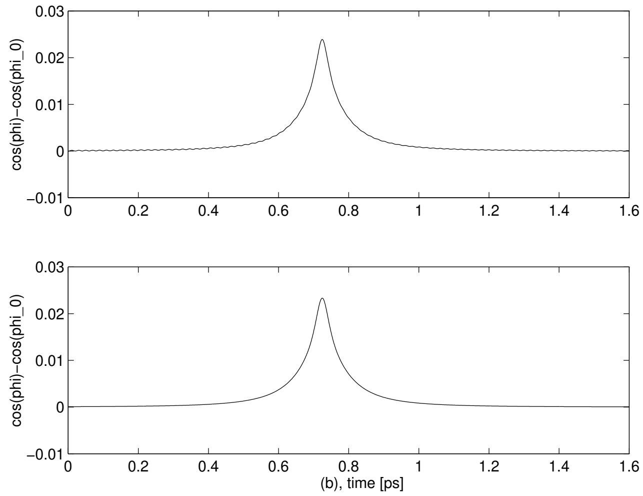


Figure 4.2: Time evolution of $\cos\phi - \cos\phi_0$ for free dynamics (a) and dynamics with modified force field/soft constraints (b).

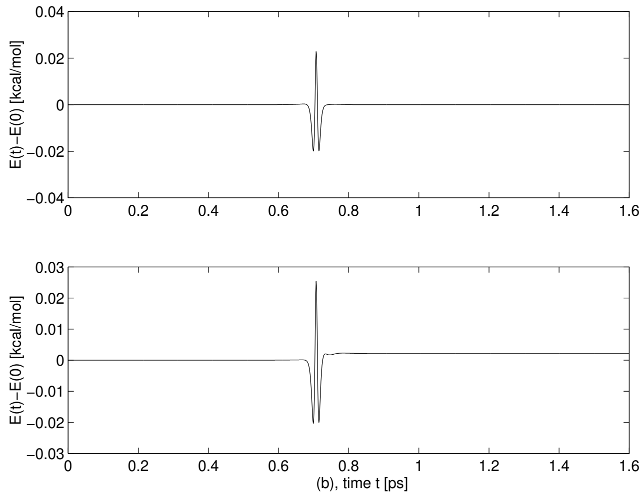


Figure 4.3: Time evolution of the total energy for "correct" modified force field (a) and "simplified" modified force field (b).

that no internal vibrations were excited. Fig. 4.2 gives a comparison of the free dynamics and the dynamics with soft constraints in the bond-angle of one of the water molecules. Note the excellent agreement (standard constrained dynamics would yield $\cos\phi - \cos\phi_0 = 0$). In Fig. 4.3, we demonstrate the importance of the correct modification of the force field. As shown in (b), the simplified force field

$$\boldsymbol{F}_n = -\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}} W(\tilde{\boldsymbol{q}}_n)$$

leads to a drift in energy after one collision. □

Example 4.4. Various models to include polarizability in classical MD simulations have been suggested. Here we will discuss the approach due to Rick, Stuart & Berne [104]. In their approach the charges $\boldsymbol{Q}$ are considered as dynamical variables and the resulting equations of motion are

$$\begin{aligned} \epsilon^2 \frac{d^2}{dt^2}\boldsymbol{Q} &= -\boldsymbol{J}(\boldsymbol{q})\,\boldsymbol{Q} - \boldsymbol{c}\,, \\ \boldsymbol{M}\frac{d^2}{dt^2}\boldsymbol{q} &= -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}) - \boldsymbol{\nabla}_{\boldsymbol{q}} \frac{\boldsymbol{Q}^T \boldsymbol{J}(\boldsymbol{q})\boldsymbol{Q}}{2}\,. \end{aligned}$$

Introducing conjugate momenta $\boldsymbol{P}$ and $\boldsymbol{p}$, the equations are Hamiltonian with Hamiltonian

$$H = \frac{\boldsymbol{P}^T\boldsymbol{P}}{2\epsilon^2} + \frac{\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}}{2} + V(\boldsymbol{q}) + \boldsymbol{c}^T\boldsymbol{Q} + \frac{\boldsymbol{Q}^T\boldsymbol{J}(\boldsymbol{q})\boldsymbol{Q}}{2}\,.$$

We assume that the symmetric matrix $\boldsymbol{J}(\boldsymbol{q})$ and the vector $\boldsymbol{c}$ are chosen such that the total charge of the system is preserved, i.e.

$$\boldsymbol{1}^T\left(\boldsymbol{J}(\boldsymbol{q})\boldsymbol{Q} + \boldsymbol{c}\right) = 0\,,$$

$\boldsymbol{1}^T = (1,1,\ldots,1)$. The equations of motion are highly oscillatory in the charges $\boldsymbol{Q}$. In particular, the charges will oscillate about their equilibrium value

$$\boldsymbol{Q}(\boldsymbol{q}) = -[\boldsymbol{J}(\boldsymbol{q})]^{-1}\boldsymbol{c}\,.$$

Thus the elimination of the high-frequency oscillations can be achieved by means of the following modified Hamiltonian

$$\begin{aligned} \tilde{H} &= \frac{\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}}{2} + V(\boldsymbol{q}) - \frac{\boldsymbol{c}^T[\boldsymbol{J}(\boldsymbol{q})]^{-1}\boldsymbol{c}}{2}\,, \\ &= \frac{\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}}{2} + V(\boldsymbol{q}) - \frac{\boldsymbol{Q}(\boldsymbol{q})^T\boldsymbol{J}(\boldsymbol{q})\boldsymbol{Q}(\boldsymbol{q})}{2}\,! \end{aligned}$$

To derive the corresponding equations of motion in $(\boldsymbol{q},\boldsymbol{p})$, let us take a different point of view: We define $\boldsymbol{Q}(\boldsymbol{q})$ by

$$\boldsymbol{J}(\boldsymbol{q})\,\boldsymbol{Q}(\boldsymbol{q}) + \boldsymbol{c} = \boldsymbol{0}\,.$$

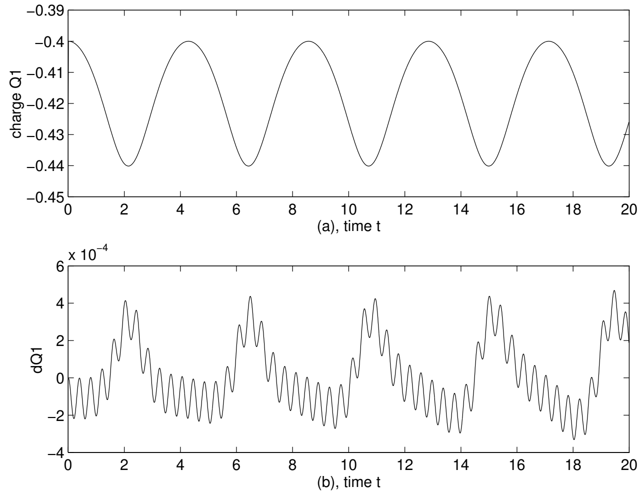


Figure 4.4: Time evolution of the partial charge Q_1 for the reduced model (a) and difference in the partial charge Q_1 between the dynamic fluctuating charge model and our reduced model (b).

Then the electrostatic force on the atomic positions is given by

$$\begin{aligned}\boldsymbol{F}_{elec}(\boldsymbol{q}) &= \boldsymbol{\nabla}_{\boldsymbol{q}}\frac{\boldsymbol{Q}(\boldsymbol{q})^T\boldsymbol{J}(\boldsymbol{q})\boldsymbol{Q}(\boldsymbol{q})}{2}\,, \\ &= \boldsymbol{\nabla}_{\boldsymbol{q}}\frac{\boldsymbol{Q}^T\boldsymbol{J}(\boldsymbol{q})\boldsymbol{Q}}{2} + \boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{Q}^T\boldsymbol{J}\boldsymbol{Q}(\boldsymbol{q})\end{aligned}$$

where suppressed arguments mean that we do not take the gradient with respect to those expressions. Now

$$\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{J}(\boldsymbol{q})\,\boldsymbol{Q} + \boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{J}\,\boldsymbol{Q}(\boldsymbol{q}) \;=\; \boldsymbol{0}$$

and, thus,

$$\boldsymbol{F}_{elec}(\boldsymbol{q}) \;=\; -\boldsymbol{\nabla}_{\boldsymbol{q}}\frac{\boldsymbol{Q}^T\boldsymbol{J}(\boldsymbol{q})\boldsymbol{Q}}{2}\,.$$

This is "equivalent" to the electrostatic force we would obtain from the fluctuating charge model. The equations of motion are now

$$\begin{aligned}\boldsymbol{M}\frac{d^2}{dt^2}\boldsymbol{q} &= -\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}) + \boldsymbol{F}_{elec}(\boldsymbol{q},\boldsymbol{Q})\,, \\ \boldsymbol{0} &= \boldsymbol{J}(\boldsymbol{q})\,\boldsymbol{Q} + \boldsymbol{c}\,, \qquad \boldsymbol{1}^T\boldsymbol{Q} = \text{ const.}\end{aligned}$$

which are equivalent to the classical equations of motion plus a linear system of equations.

In a numerical experiment, we considered two charged particles with mass $m = 1$ interacting through a Lennard-Jones potential

$$U_{LJ}(r) \quad := \quad 4\,\delta\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right],$$

$\delta = 0.15$, $\sigma = 3.5$, and a Coulomb potential

$$U_C(r) \quad := \quad J_{12}\frac{Q_1\,Q_2}{r},$$

$J_{12} = 10$. The partial charges Q_1, Q_2 are created via

$$E(Q_i) \quad := \quad \pm\xi\,Q_i + \frac{1}{2}J\,Q_i^2,$$

$J = 5$, $\xi = 1$. Thus the partial charges are given by the linear system

$$\begin{aligned} 0 &= +\xi + J\,Q_1 + \frac{J_{12}}{r}\,Q_2\,, \\ 0 &= -\xi + J\,Q_2 + \frac{J_{12}}{r}\,Q_1\,. \end{aligned}$$

Note that $Q_1 + Q_2 = 0$ is automatically satisfied. We implemented the dynamic fluctuating charge model with $\epsilon = 0.1$. The resulting partial charges were compared to our reduced model. See Fig. 4.4. □

4.2.3 Projected Multiple-Time-Stepping

Let us come back to highly oscillatory Hamiltonian systems of type

$$\begin{aligned} \frac{d}{dt}\boldsymbol{q} &= \boldsymbol{M}^{-1}\boldsymbol{p}\,, \\ \frac{d}{dt}\boldsymbol{p} &= -\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}) - \epsilon^{-2}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{g}(\boldsymbol{q})\,, \end{aligned}$$

$\boldsymbol{q},\boldsymbol{p} \in \mathrm{R}^{3N}$, $\boldsymbol{g} : \mathrm{R}^{3N} \to \mathrm{R}^m$, $m < 3N$. We assume that the $m \times m$ matrix $\boldsymbol{\partial_q g}(\boldsymbol{q})\boldsymbol{M}^{-1}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})$ is invertible. The Hamiltonian is

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

As in standard multiple-time-stepping [19],[118], we define the fast system by

$$\epsilon^{-1}\boldsymbol{A}(\boldsymbol{q},\boldsymbol{p}) \quad = \quad \begin{pmatrix} \boldsymbol{M}^{-1}\boldsymbol{p} \\ -\epsilon^{-2}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{g}(\boldsymbol{q}) \end{pmatrix}$$

and the "slow" vector field $\boldsymbol{B}$ by

$$\boldsymbol{B}(\boldsymbol{q}) \quad = \quad \begin{pmatrix} \boldsymbol{0} \\ -\boldsymbol{\nabla}_{\boldsymbol{q}}V(\boldsymbol{q}) \end{pmatrix}.$$

This leads to the following multiple-time-stepping scheme:

<u>STANDARD MULTIPLE-TIME-STEPPING</u>

Step 1.

$$\bar{\boldsymbol{p}}_n = \boldsymbol{p}_n - \frac{\Delta t}{2}\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}_n)$$

Step 2.

Integrate the fast system

$$\begin{aligned} \frac{d}{dt}\boldsymbol{q} &= \boldsymbol{M}^{-1}\boldsymbol{p} \\ \frac{d}{dt}\boldsymbol{p} &= -\epsilon^{-2}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{g}(\boldsymbol{q}) \end{aligned}$$

using Verlet with a step-size $\delta t = \Delta t/N$, $N \gg 1$, and initial conditions $(\boldsymbol{q}_n, \bar{\boldsymbol{p}}_n)$. Denote the result by $(\boldsymbol{q}_{n+1}, \bar{\boldsymbol{p}}_{n+1})$.

Step 3.

$$\boldsymbol{p}_{n+1} = \bar{\boldsymbol{p}}_{n+1} - \frac{\Delta t}{2}\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{q}_{n+1})$$

This formulation suffers from resonance induced instabilities [19],[21]. In [42], GARCÍA-ARCGILLA, SANZ-SERNA & SKEEL suggested to combine averaging with multiple-time-stepping. Here we use informations on the analytical solution behavior to define an approximation to the average $\bar{\boldsymbol{B}}$ of $\boldsymbol{B}$ along solution curves of $\epsilon^{-1}\boldsymbol{A}$. Note that, by energy considerations, $\boldsymbol{g}(\boldsymbol{q}) = \mathcal{O}(\epsilon)$. Furthermore, the motion is (almost) harmonic and highly oscillatory in $\boldsymbol{r} := \boldsymbol{g}(\boldsymbol{q})$. Thus we approximate the averaged vector field $\bar{\boldsymbol{B}}$ by

$$\bar{\boldsymbol{B}}(\boldsymbol{q}) = \begin{pmatrix} \boldsymbol{0} \\ -\boldsymbol{\nabla}_{\boldsymbol{q}} V(\boldsymbol{\rho}(\boldsymbol{q})) \end{pmatrix}.$$

The function $\boldsymbol{\rho}$ is defined by the nonlinear system of equations

$$\begin{aligned} \boldsymbol{\rho}(\boldsymbol{q}) &= \boldsymbol{q} + \boldsymbol{M}^{-1}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{\mu}\,, \\ \boldsymbol{0} &= \boldsymbol{g}(\boldsymbol{\rho}(\boldsymbol{q})) \end{aligned}$$

in the variable $\boldsymbol{\mu} \in \mathrm{R}^m$. Note that $\boldsymbol{\rho}$ basically projects the $\boldsymbol{r} = \boldsymbol{g}(\boldsymbol{q})$ solution component away. This choice of $\bar{\boldsymbol{B}}$ introduces an error of size $\mathcal{O}(\epsilon)$. For some of the components of the potential energy function V this might be not tolerable. Then we simply split V into two parts V_1 and V_2, include the "troublesome" V_1 in the fast part $\epsilon^{-1}\boldsymbol{A}$ that is solved with a small step-size δt, and only keep V_2 in the vector field $\boldsymbol{B}$. Again we have $[\epsilon^{-1}\boldsymbol{A}, \bar{\boldsymbol{B}}] = \mathcal{O}(1)$, and the corresponding modified multiple-time-stepping method can be used with a macro step-size $\epsilon \ll \Delta t \ll 1$. To be more precise, the step-size Δt is only determined by the slowly varying solution components and is independent of ϵ. One final remark: To implement our approach, we need the

Jacobian $\boldsymbol{\partial}_{\boldsymbol{q}}\boldsymbol{\rho}$ of $\boldsymbol{\rho}$. This requires the computation of the second derivative of $\boldsymbol{g}$ and the solution of a linear system of equations, i.e.,

$$\begin{aligned} \boldsymbol{d\tilde{q}} &= \boldsymbol{dq} + \boldsymbol{M}^{-1}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{d\mu} + \boldsymbol{M}^{-1}\sum_{i=1}^{m}\mu_i\,\boldsymbol{\partial}_{\boldsymbol{q}}\boldsymbol{g}^i(\boldsymbol{q})\,\boldsymbol{dq}\,, \\ \boldsymbol{0} &= \boldsymbol{\partial}_{\tilde{\boldsymbol{q}}}\boldsymbol{g}(\tilde{\boldsymbol{q}})\,\boldsymbol{d\tilde{q}}\,, \end{aligned}$$

with $\tilde{\boldsymbol{q}} = \boldsymbol{\rho}(\boldsymbol{q})$ and $\boldsymbol{d\tilde{q}} = \boldsymbol{\partial}_{\boldsymbol{q}}\boldsymbol{\rho}(\boldsymbol{q})\boldsymbol{dq}$, or, in other words,

$$\boldsymbol{\partial}_{\boldsymbol{q}}\boldsymbol{\rho}(\boldsymbol{q}) = \left[\boldsymbol{I} - \boldsymbol{M}^{-1}\,\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{N}\,\boldsymbol{\partial}_{\tilde{\boldsymbol{q}}}\boldsymbol{g}(\tilde{\boldsymbol{q}})\right]\left[\boldsymbol{I} + \boldsymbol{M}^{-1}\sum_{i=1}^{m}\mu_i\,\boldsymbol{\partial}^2_{\boldsymbol{q}}\boldsymbol{g}^i(\boldsymbol{q})\right].$$

with

$$\boldsymbol{N} := [\boldsymbol{\partial}_{\tilde{\boldsymbol{q}}}\boldsymbol{g}(\tilde{\boldsymbol{q}})\,\boldsymbol{M}^{-1}\,\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})]^{-1}\,.$$

For a system with Hamiltonian

$$H(\boldsymbol{q},\boldsymbol{p}) = \frac{\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p}}{2} + V_1(\boldsymbol{q}) + V_2(\boldsymbol{q}) + \frac{\boldsymbol{g}(\boldsymbol{q})^T\boldsymbol{g}(\boldsymbol{q})}{2\epsilon^2}\,,$$

where the gradient of V_2 is much more expensive to compute than the other forces, we suggest the following scheme:

PROJECTED MULTIPLE-TIME-STEPPING

Step 1.

$$\begin{aligned}\tilde{\boldsymbol{q}}_n &= \boldsymbol{\rho}(\boldsymbol{q}_n)\,,\\ \boldsymbol{F}_n &= -[\boldsymbol{\partial}_{\boldsymbol{q}}\boldsymbol{\rho}(\boldsymbol{q}_n)]^T\,\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}V_2(\tilde{\boldsymbol{q}}_n)\end{aligned}$$

Step 2.

$$\bar{\boldsymbol{p}}_n = \boldsymbol{p}_n + \frac{\Delta t}{2}\boldsymbol{F}_n$$

Step 3.

Integrate the fast system

$$\begin{aligned}\frac{d}{dt}\boldsymbol{q} &= \boldsymbol{M}^{-1}\boldsymbol{p}\\ \frac{d}{dt}\boldsymbol{p} &= -\epsilon^{-2}\boldsymbol{\nabla}_{\boldsymbol{q}}\boldsymbol{g}(\boldsymbol{q})\,\boldsymbol{g}(\boldsymbol{q}) - \boldsymbol{\nabla}_{\boldsymbol{q}}V_1(\boldsymbol{q})\end{aligned}$$

using Verlet with a step-size $\delta t = \Delta t/N$, $N \gg 1$, and initial conditions $(\boldsymbol{q}_n, \bar{\boldsymbol{p}}_n)$. Denote the result by $(\boldsymbol{q}_{n+1}, \bar{\boldsymbol{p}}_{n+1})$.

Step 4.

$$\begin{aligned}\tilde{\boldsymbol{q}}_{n+1} &= \boldsymbol{\rho}(\boldsymbol{q}_{n+1})\,,\\ \boldsymbol{F}_{n+1} &= -[\boldsymbol{\partial}_{\boldsymbol{q}}\boldsymbol{\rho}(\boldsymbol{q}_{n+1})]^T\,\boldsymbol{\nabla}_{\tilde{\boldsymbol{q}}}V_2(\tilde{\boldsymbol{q}}_{n+1})\end{aligned}$$

Step 5.

$$\boldsymbol{p}_{n+1} = \bar{\boldsymbol{p}}_{n+1} + \frac{\Delta t}{2}\boldsymbol{F}_{n+1}$$

This symplectic (!) scheme should avoid the resonance problems typically encountered in standard multiple-time-stepping and should be useful whenever the evaluation of $\boldsymbol{\nabla}_{\boldsymbol{q}}V_2(\boldsymbol{q})$ (long-range forces) is much more expensive than the evaluation of $\boldsymbol{\nabla}_{\boldsymbol{q}}V_1(\boldsymbol{q})$.

Example 4.5. The modified multiple-time-stepping method of GARCIA-ARCHILLA, SANZ-SERNA & SKEEL as well as our projected multiple-time-stepping method have been successfully tested for a box of water. Both methods allow one to increase the step-size Δt from $1-2$ femtoseconds to $5-7$ femtoseconds without any additional evaluation of the long-range forces. However, the projected multiple-time-stepping method seems more robust (less drift in total energy) [63]. □

Bibliography

[1] Adams, M., Ratiu, T., and Schmid, R., The Lie Group Structure of Diffeomorphism Groups and Invertible Fourier Integrals Operators with Applications, in: *Infinite Dimensional Groups with Applications*, Kac, V. (editor), Springer Verlag, New York, 1985.

[2] Allen, M.P. and Tildesley, D.J., *Computer Simulations of Liquids*, Clarendon Press, Oxford, 1987.

[3] Anderson, H.C., Rattle: A 'Velocity' Version of the Shake Algorithm for Molecular Dynamics Calculations, *J. Comp. Phys.* **52**, 24–34, 1983.

[4] Anosov, D.V., Averaging in Systems of Ordinary Differential Equations with Rapidly Oscillating Solutions, *Izv. Akad. Nauk SSSR* **24**, 721–742, 1960.

[5] Arnold, V.I., *Geometrische Methoden in der Theorie der gewöhnlichen Differentialgleichungen*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1987.

[6] Arnold, V.I., *Mathematische Methoden der klassischen Mechanik.* VEB Deutscher Verlag der Wissenschaften, Berlin, 1988.

[7] Arnold, V.I., Kozlov, V.V., and Neishtadt, A.I., *Mathematical Aspects of Classical and Celestial Mechanics*, second edition, Springer-Verlag, New York, 1997.

[8] Ascher, U. and Reich, S., The Midpoint Scheme and Variants for Hamiltonian Systems: Advantages and Pitfalls, *SIAM J. Sci. Comput.*, to appear, 1998.

[9] Auerbach, S.P. and Friedman, A., Long-time Behaviour of Numerically Computed Orbits: Small and Intermediate Time-Step Analysis of One-Dimensional Systems, *J. Comput. Phys.* **93**, 189–223, 1991.

[10] Barth, E. and Leimkuhler, B., Symplectic Methods for Conservative Multibody Systems, *Fields Instit. Commun.* **10**, 25-44, 1996.

[11] Barth, E., Leimkuhler, B., and Reich, S., A Semi-Explicit, Variable Stepsize, Time-Reversible Integrator for Constrained Dynamics, *SIAM J. Sci. Comput.*, to appear, 1998.

[12] Benettin, G., Carati, A., and Gallavotti, G., A Rigorous Implementation of the Jeans-Landau-Teller Approximation for Adiabatic Invariants, *Nonlinearity* **10**, 479–505, 1997.

[13] Benettin, G., Galgani, L., and Giorgilli, A., Exponential Law for the Equipartition Times Among Translational and Vibrational Degrees of Freedom, *Physica Letters A* **120**, 23–27, 1987.

[14] Benettin, G., Galgani, and L., Giorgilli, A., Realization of Holonomic Constraints and Freezing of High Frequency Degrees of Freedom in the Light of Classical Perturbation Theory. Part I. *Commun. Math. Phys.* **113**, 87–103, 1987.

[15] Benettin, G., Galgani, L., and Giorgilli, A., Realization of Holonomic Constraints and Freezing of High Frequency Degrees of Freedom in the Light of Classical Perturbation Theory. Part II. *Commun. Math. Phys.* **113**, 557–601, 1989.

[16] Benettin, G. and Giorgilli, A., On the Hamiltonian Interpolation of Near to the Identity Symplectic Mappings with Application to Symplectic Integration Algorithms. *J. Statist. Phys.* **74**, 1117–1143, 1994.

[17] Berry, M.V., Quantum Phase Factors Accompanying Adiabatic Change, *Porc. Roy. Soc.* **A392**, 45, 1984.

[18] Beyn, W.-J., Numerical Methods for Dynamical Systems, in *Advances in Numerical Analysis* Vol. I, Clarendon Press, Oxford, 1991.

[19] Biesiadecki, J.J. and Skeel, R.D., Dangers of Multiple-Time-Step Methods, *J. Comput. Phys.* **109**, 318–328, 1993.

[20] Birkhoff, G.D., *Dynamical Systems*, A.M.S. Publications, Providence, 1927.

[21] Bishop, T, Skeel, R.D., Schulten, K., Difficulties with Multiple Timestepping and the Fast Multipole Algorithm in Molecular Dynamics, *J. Comput. Chem.* **18**, 1785–1791, 1997.

[22] Bogolyubov, N.N. and Mitropolskij, Y.A., *Asymptotic Methods in the Theory of Nonlinear Oscillations*, Gordon and Breach Science Publ., NY, 1996.

[23] Born, M. and Fock, V., Beweis des Adiabatensatzes, *Z. Phys.* **51**, 165–180, 1928.

[24] Bornemann, F. and Schütte, Ch., Homogenization of Hamiltonian Systems with a Strong Constraining Potential, *Physica D* **102**, 57–77, 1997.

[25] Bornemann, F. and Schütte, Ch., A Mathematical Approach to Smoothed Molecular Dynamics: Correcting Potentials for Freezing Bond Angles, Preprint SC 95-30, Konrad-Zuse-Zentrum Berlin, 1995.

[26] Bornemann, F. and Schütte, Ch., On the Singular Limit of the Quantum-Classical Molecular Dynamics Model, *SIAM J. Appl. Math.*, to appear, 1998.

[27] Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M., CHARMM: A Program for Macromolecular Energy, Minmization, and Dynamics Cacluations, *J. Comput. Chem.* **4**, 187–217, 1983.

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

[29] de Almeida, A.M.O., *Hamiltonian Systems: Chaos and Quantization*. Cambridge University Press, Cambridge, 1988.

[30] Deitrick, G.L., Scriven, L.E., and Davis, H.T., A New Method of Error Analysis for Molecular Simulations, *Computer Physics Communications* **62**, 327–335, 1991.

[31] Delshams, A. and Gutiérrez, P., Estimates on Invariant Tori near an Elliptic Equilibrium Point of a Hamiltonian System, *J. Diff. Eqs.* **131**, 277–303, 1996.

[32] Deuflhard, P. and Bornemann, F., *Numerische Mathematik*, Vol. II, Walter de Gryter, Berlin, 1994.

[33] Dittrich, W. and Reuter, M., *Classical and Quantum Dynamics*, second edtion, Springer-Verlag, 1994.

[34] Dullweber, A., Leimkuhler, B., and McLachlan, R., Split-Hamiltonian Methods for Rigid Body Molecular Dynamics, *J. Chem. Phys.* **107**, 5840–5852, 1997.

[35] Eirola, T., Aspects of Backward Error Analysis of Numerical ODEs, *J. Comput. Appl. Math.* **45**, 65–73, 1993.

[36] Fassò, F., Lie Series Method for Vector Fields and Hamiltonian Perturbation Theory, *ZAMP* **41**, 843–854, 1990.

[37] Feng, K., Formal Power Series and Numerical Algorithms for Dynamical Systems. Proceedings of International Conference on Scientific Computation, Hangzhou, China, Eds. Tony Chan & Zhong-Ci Shi, *Series on Appl. Math* **1**, 28–35, 1991.

[38] Fiedler, B. and Scheurle, J., Discretization of Homoclinic Orbits and Invisible Chaos, *Memoirs Amer. Math. Soc.* **570**, 1996.

[39] Fixman, M., Classical Statistical Mechanics of Constraints: A Theorem and Application to Polymers, *Proc. Nat. Acad. Sci.* **71**, 2635–2638, 1974.

[40] Garay, B.M., Hyperbolic Structures in ODE's and Their Discretization. in: *Nonlinear Analysis and Boundary Value Problems for Ordinary Differential Equations*, F. Zanolin (ed.), 1996.

[41] Garay, B.M., The Discretized Flow on Domains of Attraction: A Structural Stability Result, submitted, 1997.

[42] Garcia-Archilla, B., Sanz-Serna, J.M., and Skeel, R.D., Long-Time-Step Methods for Oscillatory Differential Equations, *SIAM J. Sci. Comput.*, to appear, 1998.

[43] Ge, Z., Symplectic Difference Schemes and Generating Functions, *Physica D* **49**, 376–386, 1991.

[44] Ge, Z. and Marsden, J.E., Lie-Poisson Hamilton-Jacobi Theory and Lie-Poisson Integrators, *Physics Letters A* **133**, 134–139, 1988.

[45] Giacaglia, G.E.O., Perturbation Methods in Non-Linear Systems, Springer-Verlag, NY, 1972.

[46] Giorgilli, A., Delshams, A., Fontich, E., Galgani, L., and Simó, C., Effective Stability for a Hamiltonian System Near an Elliptic Equilibrium Point with an Application to the Restricted Three Body Problem, *J. Diff. Eqs.* **77**, 167–198, 1989.

[47] Giorgilli, A. and Galgani, L., Rigorous Estimates for the Series Expansions of Hamiltonian Perturbation Theory, *Cel. Mech.* **37**, 95–112, 1985.

[48] Griffiths, D.F. and Sanz-Serna, J.M., On the Scope of the Modified Equations. *J. Sci. Statist. Comput.* **7**, 994–1008, 1986.

[49] Gonzales, O. and Stuart, A., Remarks on the Qualitative Properties of Modified Equations, in: *Foundations of Computational Mathematics*, F. Cucker and M. Shub (eds.), Springer-Verlag, New York, 1997.

[50] Grubmüller, H., Heller, H., Windemuth, A., and Schulten, K., Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-Range Interactions, *Molecular Simulations* **6**, 121–142, 1991.

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

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

[53] Hairer, E., Backward Analysis of Numerical Integrators and Symplectic Methods, *Annals of Numerical Mathematics* **1**, 107–132, 1994.

[54] Hairer, E., Variable Time Step Integration with Symplectic Methods, *Appl. Numer. Math.* **25**, 219–227, 1997.

[55] Hairer, E. and Lubich, Ch., The Life-Span of Backward Error Analysis for Numerical Integrators, *Numer. Math.* **76**, 441–462, 1997.

[56] Hairer, E., Nørsett, S.P., and Wanner, G., *Solving Ordinary Differential Equations*, Vol. I., second revised edition, Springer Verlag, 1993.

[57] Hairer, E. and Stoffer, D., Reversible Long-Term Integration With Variable Step Sizes. *SIAM J. Sci. Comput.* **18**, 257–269, 1997.

[58] Hamilton, R., The Inverse Function Theorem of Nash and Moser, *Bull. Am. Math. Soc.* **7**, 65-222, 1982.

[59] Helfand, E., Flexible vs. Rigid Constraints in Statistical Mechanics, *J. Chem. Phys.* **71**, 5000–5007, 1979.

[60] Hinch, E.J., *Perturbation Methods*, Cambridge University Press, Cambridge, 1991.

[61] Hori, G., Theory of General Perturbations with Unspecified Variables, *Astron. Soc. Japan* **18**, 287, 1966.

[62] Huang, W. and Leimkuhler, B., The Adaptive Verlet Method, *SIAM J. Sci. Comput.* **18**, 239–256, 1997.

[63] Izaguirre, J., Reich, S., and Skeel, R.D., Longer Time-Steps for Molecular Dynamics, submitted, 1998.

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

[65] Jay, L., Symplectic Partitioned Runge-Kutta Methods for Constrained Hamiltonian Systems, *SIAM J. Num. Anal.* **33**, 368–387, 1996.

[66] Joye, A., Kunz, H., and Pfister, Ch.-Ed., Exponential Decay and Geometric Aspects of Transition Probabilities in the Adiabatic Limit, *Ann. Phys.* **208**, 299–332, 1991.

[67] Kamel, A.A., Perturbation Methods in the Theory of Nonlinear Oscillations, *Cel. Mech.* **3**, 90–106, 1970.

[68] Kato, T., *Perturbation Theory for Linear Operators*, 2nd edition, Springer-Verlag, 1976.

[69] Kirchgraber, U., and Stoffer, D., On the Definition of Chaos, *ZAMM* **69**, 175–185, 1989.

[70] Kol, A., Laird, B., and Leimkuhler, B., A Symplectic Method for Rigid-Body Molecular Simulation, *J. Chem. Phys.* **107**, 2580–2588, 1997.

[71] Kopell, N., Invariant Manifolds and the Initialization Problem for some Atmospheric Equations, *Physica D* **14**, 203–215, 1985.

[72] Krylov, N.M. and Bogoliubov, N.N., *Introduction to Nonlinear Mechanics*, Princeton University Press, Princeton, 1947.

[73] Landau, L.D. and Lifschitz, E.M., *Lehrbuch der Theoretischen Physik*, Vol. I, Akademie-Verlag, Berlin, 1987.

[74] Landau, L.D. and Lifschitz, E.M., *Lehrbuch der Theoretischen Physik*, Vol. III, Akademie-Verlag, Berlin, 1986.

[75] Landau, L.D. and Teller, E., *Physik. Z. Sowietunion* **10**, 34, 1936.

[76] Leimkuhler, B. and Skeel, R.D., Symplectic Numerical Integrators in Constrained Hamiltonian Systems, *J. Comp. Phys.* **112**, 117–125, 1994.

[77] Leimkuhler, B., Skeel, R.D., and Reich, S., Integration Methods for Molecular Dynamics, in: *Mathematical Approaches to Biomolecular Structure and Dynamics*, Schulten, K. and Mesirov, J.P. (eds.), 161–186, Springer-Verlag, New York, 1996.

[78] Lichtenberg, A.J. and Lieberman, M.A., *Regular and Chaotic Dynamics*, second edition, Springer-Verlag, New York, 1992.

[79] Liverani, G., Decay of correlation, *Ann. of Math.* **142**, 239–301, 1995.

[80] Lochak, P. and Meunier, C., *Multiphase Averaging for Classical Systems*, Springer-Verlag, 1988.

[81] Mackey, M.C., The Dynamical Origin of Increasing Entropy, *Review of Modern Physics* **61**, 981–1015, 1989.

[82] Marsden, J.R. and Ratiu, T., *An Introduction to Mechanics and Symmetry*, Springer-Verlag, 1994.

[83] McLachlan, R.I., Explicit Lie-Poisson Integration and the Euler Equations, *Phys. Rev. Lett.* **71**, 3043–3046, 1993.

[84] McLachlan, R.I., Scovel, C., Equivariant Constrained Symplectic Integration, *J. Nonlinear Science* **5**, 233–256, 1995.

[85] Moser, J., Lectures on Hamiltonian Systems, *Mem. Am. Math. Soc.* **81**, 1–60, 1968.

[86] Moser, J., *Stable and Random Motion in Dynamical Systems.* Princeton University Press, Princeton, 1973.

[87] Neishtadt, A.I., Estimates in the Kolmogorov Theorem on the Conservation of Conditionally Periodic Motions, *PMM* **45**, 1981.

[88] Neishtadt, A.I., The Separation of Motions in Systems with Rapidly Rotating Phase. *J. Appl. Math. Mech.* **48**, 133–139, 1984.

[89] Nekhoroshev, N.N., Exponentially Estimate for the Stability Time of Near-Integrable Hamiltonian Systems, *Russ. Math. Surv.* **32**, 1-65, 1977.

[90] Olver, P., *Applications of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1986.

[91] Perry, A.D. and Wiggins, S., KAM Tori are Very Sticky: Rigorous Lower Bounds on the Time to Move Away from an Invariant Lagrangian Torus with Linear Flow, *Physica D* **71**, 102–121, 1994.

[92] Poincaré, H., *Les Méthodes Nouvelles de la Mécanique Céleste*, Vol. I–III, Gauthier-Villars, Paris, 1899.

[93] Pöschel, J., Nekhoroshev Estimates for Quasi-Convex Hamiltonian Systems, *Math. Z.* **213**, 187–216, 1993.

[94] Reich, S., Numerical Integration of Generalized Euler Equations, preprint, 1993.

[95] Reich, S., Momentum Conserving Symplectic Integrators, *Physica D* **76**, 375–383, 1994.

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

[97] Reich, S., Smoothed Dynamics of Highly Oscillatory Hamiltonian Systems, *Physica D* **89**, 28–42, 1995.

[98] Reich, S., Symplectic Integration of Constrained Hamiltonian Systems by Composition Methods, *SIAM J. Numer. Anal.* **33**, 475–491, 1996.

[99] Reich, S., Torsion Dynamics of Molecular Systems, *Phys. Rev. E* **53**, 4176–4181, 1996.

[100] Reich, S., Smoothed Langevin Dynamics of Highly Oscillatory Systems, submitted, 1996.

[101] Reich, S., Symplectic Integrators for Systems of Rigid Bodies, *Fields Institute Communications* **10**, 181–191, 1996.

[102] Reich, S., Preservation of Adiabatic Invariants under Symplectic Discretization, *Appl. Numer. Math.*, to appear, 1998.

[103] Reich, S., Modified Potential Energy Functions for Constrained Molecular Dynamics, *Numerical Algorithms*, to appear, 1998.

[104] Rick, S.W., Stuart, S.J., Berne, B.J., Dynamical Fluctuating Charge Force Fields: Application to Liquid Water, *J. Chem. Phys.* **101**, 6141–6156, 1994.

[105] Rubin, H. and Ungar, P., Motion Under a Strong Constraining Force, *Comm. Pure Appl. Math.* **10**, 65–87, 1957.

[106] Ryckaert, J.P., Ciccotti, G., Berendsen, H.J.C., Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes, *J. Comput. Phys.* **23**, 327–342, 1977.

[107] Sanders, J.A. and Verhulst, F., *Averaging Methods in Nonlinear Dynamical Systems*, Springer-Verlag, Berlin, 1985.

[108] Sanz-Serna, J.M., Symplectic Integrators for Hamiltonian Problems: An Overview, *Acta Numerica* **1**, 243–286, 1992.

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

[110] Sauer, T. and York, J.A., Rigorous Verification of Trajectories for the Computer Simulation of Dynamical Systems, *Nonlinearity* **4**, 961–979, 1994.

[111] Skeel, R.D. and Biesiadecki, J.J., Symplectic Integration with Variable Stepsize, *Annals of Numer. Math.* **1**, 191–198, 1994.

[112] Shimada, M. and Yoshida, H., Long-Term Conservation of Adiabatic Invariants by Using Symplectic Integrators. *Publ. Astron. Soc. Japan* **48**, 147–155, 1996.

[113] Siegel, C.L., Über die Existenz einer Normalform analytischer Hamiltonscher Differentialgleichungen in der Nähe einer Gleichgewichtslösung, *Math. Ann.* **128**, 144–170, 1942.

[114] Stoffer, D. and Nipp, K., Invariant Curves for Variable Step Size Integrators, *BIT* **31**, 169–180, 1991.

[115] Sun, G., Symplectic Partitioned Runge-Kutta Methods, *J. Comput. Math.* **11**, 365–372, 1993.

[116] Takens, F., Motion Under the Influence of a Strong Constraining Force, in: Global Theory of Dynamical Systems, *Lecture Notes Math.* **819**, 425–445, 1980.

[117] Toda, M., Kubo, R., Saito, N., *Statistical Physics I*, second edition, Springer-Verlag, 1992.

[118] Tuckerman, M., Berne, B.J., and Martyna, G.J., Reversible Multiple Time Scale Molecular Dynamics, *J. Chem. Phys.* **97**, 1990–2001, 1992.

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

[120] Verlet, L., Computer Experiments on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules, *Phys. Rev.* **159**, 1029–1039, 1967.

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

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

[123] Warming, R.F. and Hyett, B.J., The Modified Equation Approach to the Stability and Accuracy of Finite-Difference Methods, *J. Comp. Phys.* **14**, 159–179, 1974.

[124] Yoshida, H., Construction of Higher Order Symplectic Integrators, *Phys. Lett. A* **150**, 262–268, 1990.

[125] Zare, K. and Szebehely, V., Time Transformations for the Extended Phase Space, *Celestial Mechanics* **11**, 469–482, 1975.

[126] Zhou, J., Reich, S., and Brooks, B.R., Elastic Molecular Dynamics with Self-Consistent Flexible Constraints, submitted, 1998.