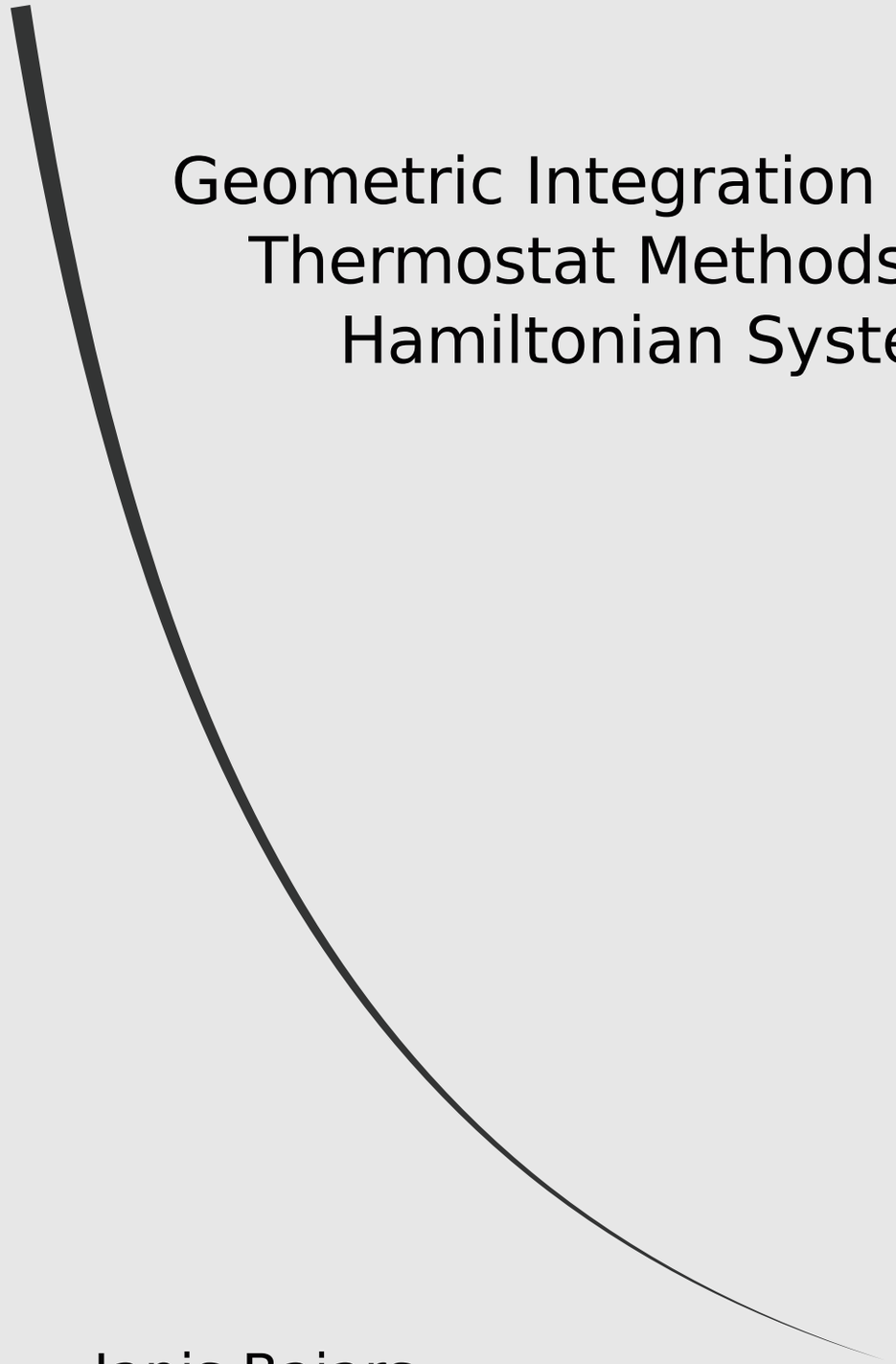
Geometric Integration and Thermostat Methods for Hamiltonian Systems

Janis Bajars

# Geometric Integration and Thermostat Methods for Hamiltonian Systems



Janis Bajars

# Geometric Integration and Thermostat Methods for Hamiltonian Systems

Printed by Ipskamp Drukkers.

ISBN 978-94-6191-488-0

# Geometric Integration and Thermostat Methods for Hamiltonian Systems

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 7 december 2012, te 14:00 uur

door

**Janis Bajars**

geboren te Riga, Letland

**Promotiecommissie**

| | |
|---|---|
| Promotor: | prof. dr. ir. J.E. Frank |
| Overige leden: | dr. J.H. Brandts |
| | prof. dr. B. Leimkuhler |
| | prof. dr. L.R.M. Maas |
| | prof. dr. M.R.H. Mandjes |
| | prof. dr. ir. C.W. Oosterlee |
| | prof. dr. R.P. Stevenson |

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 613.000.552 and carried out at the Centrum Wiskunde & Informatica (CWI) in Amsterdam.

*To my parents*

*Jānis and Iveta*

# Contents

# Preface

Geometric numerical integration has been an active research area for the last three decades. The subject has redefined numerical analysis. Attention has turned to the development of numerical methods for particular classes of equations such as, Hamiltonian dynamical systems, Hamiltonian and multisymplectic partial differential equations, Poisson systems, Euler-Lagrange equations, etc. Geometric methods have found their applications in celestial mechanics, rigid body dynamics, molecular dynamics, geophysical fluid dynamics and statistical mechanics, among others.

Thermostats, deterministic and stochastic, are built upon Hamiltonian structure to allow trajectories to sample a (possibly modified) Gibbs measure and have been used as modelling devises in molecular dynamics simulations with great success. As such, thermostat methods can be viewed as model reduction techniques and may be applicable to stochastic modelling of unresolved scales or used for development of closure models with applications in geophysical fluid dynamics and fluid dynamics in general.

For the interested reader this thesis may serve as motivation, from the application point of view, for the use of geometric numerical methods and give some insights on the particular research areas conducted under this thesis, that is, structure preserving discretization of the Euler-Boussinesq equations and study of wave attractors in a confined stratified fluid, stochastic-dynamical thermostat methods applied to Hamiltonian systems with holonomic constraints and weakly coupled heat bath models for nonlinear wave equations.

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 613.000.552 and conducted in the Modelling, Analysis and Computing (MAC) department of the Centrum Wiskunde & Informatica (CWI) in Amsterdam.

Chapters 2 through 4 of this thesis have appeared as published or submitted journal articles:

**Chapter 2** J. Bajars, J. Frank & L.R.M. Maas, "On the appearance of internal wave attractors due to an initial or parametrically excited disturbance", Journal of Fluid Mechanics, in press.

**Chapter 3** J. Bajars, J. Frank & B. Leimkuhler, "Stochastic-dynamical thermostats for constraints and stiff restraints", The European Physical Journal - Special Topics **200** (2011), 131-152.

**Chapter 4** J. Bajars, J. Frank & B. Leimkuhler, "Weakly coupled heat bath models for Gibbs-like invariant states in nonlinear wave equations", submitted, 2012.

Chapter 1 provides an introduction to geometric numerical integration and thermostated dynamics, and collects much of the background material needed to read the rest of the thesis.

Janis Bajars
*Amsterdam, August 2012*

# Chapter 1

# Introduction to Geometric Integration and Thermostats for Hamiltonian Systems

## 1.1 Geometric numerical integration

In this thesis we are concerned with geometric numerical integration of wave equations and thermostat methods with applications to molecular dynamics and geophysical fluid dynamics. Under geometric numerical integration we understand the structure preserving numerical methods for the ordinary and partial differential equations, (ODEs) and (PDEs), respectively. In particular: Hamiltonian dynamical systems and Hamiltonian PDEs, with extensions to the Hamiltonian systems with holonomic constraints and stochastic differential equations (SDEs), such as thermostated dynamics. The choice of the geometric integrators in this thesis is directly related to the underlying structure of Hamiltonian dynamics, i.e. conserved quantities, symplecticity, volume preservation and time reversibility. Our objective is to preserve these properties under the numerical integration, in space and time.

We begin our introduction with a motivating example in Section 1.1.1. The main unifying mathematical concept in this thesis is Hamiltonian dynamics, which we present in Section 1.1.2. In Section 1.1.3 we consider canonical Hamiltonian systems from the perspective of classical mechanics. A key property of canonical Hamiltonian systems, i.e. symplecticity, is described in Section 1.1.4. Extension to Poisson systems and definition of the Poisson bracket are presented in Section 1.1.5. In Section 1.1.6 we describe Hamiltonian PDEs. Examples of structure preserving numerical methods for Hamiltonian PDEs are given in Section 1.1.7. In Section 1.1.8 we describe geometric integrators for Hamiltonian systems and semi-discretized Hamiltonian PDEs.

Most of the material presented in this section can be found in the following references: Hairer et al. [37], Leimkuhler & Reich [63], Sanz-Serna & Calvo [105], Arnold [3], Olver [91], Swaters [107], Golub & van Loan [35], Ortega [93], Durran

[23], Trefethen [110], Iserles [47], Leveque [66, 67].

### 1.1.1 Motivation: discrete vs. continuous dynamics

In this subsection we discus and show the importance of structure preserving methods for conservative dynamical systems and semi-discretized wave equations. Let us illustrate with a simple example how different numerical integrators effect the qualitative nature of the original dynamical system. We consider one of the classical motivational examples for use of geometric numerical integrators, i.e. the harmonic oscillator equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = y, \tag{1.1}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\omega^2 x, \tag{1.2}$$

where $x(t), y(t) : \mathbf{R} \to \mathbf{R}$ and $\omega \in \mathbf{R}_+$ is a frequency. The system of differential equations (1.1)–(1.2) is a linear *autonomous* dynamical system subject to the initial conditions $(x(0), y(0)) = (x_0, y_0)$. For any initial condition $(x_0, y_0) \in \mathbf{R}^2$ the analytical solution of (1.1)–(1.2) reads:

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{bmatrix} \cos(\omega t) & \frac{1}{\omega}\sin(\omega t) \\ -\omega \sin(\omega t) & \cos(\omega t) \end{bmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \tag{1.3}$$

System (1.1)–(1.2) is derived from Newton's $2^{nd}$ law of motion and describes the motion of a bob (point of mass) attached to the elastic spring in a frictionless environment. Functions $x(t)$ and $y(t) := \frac{\mathrm{d}x}{\mathrm{d}t}$ stand for the bob's displacement and the velocity, respectively, from its equilibrium state $(x, y) \equiv (0, 0)$. This equilibrium state is also a stationary point of the dynamical system (1.1)–(1.2), i.e. if $(x_0, y_0) \equiv 0$ then $(x(t), y(t)) \equiv 0$ for all times $t$. In fact, the origin $(0, 0)$ is a unique stationary point of (1.1)–(1.2) and a *center*, since eigenvalues of the system matrix of (1.1)–(1.2) are purely imaginary, i.e. $\lambda = \pm\omega i$. This implies that the solutions are periodic which can be seen from (1.3) and dynamics is constrained to the periodic orbits in the *phase space* $\mathbf{R}^2$ of $(x, y)$. Indeed, the function $H(x, y)$ (the total energy of the system (1.1)–(1.2)) defined by

$$H(x, y) = \frac{1}{2}y^2 + \frac{1}{2}\omega^2 x^2 \tag{1.4}$$

is invariant under the motion of (1.1)–(1.2), i.e.

$$\frac{\mathrm{d}H}{\mathrm{d}t} = y\frac{\mathrm{d}y}{\mathrm{d}t} + \omega^2 x\frac{\mathrm{d}x}{\mathrm{d}t} = -\omega^2 yx + \omega^2 xy = 0,$$

and for each value of $H(x_0, y_0) > 0$ defines the equation for an ellipse in $(x, y)$ coordinates.

For simplicity we take $\omega = 1$ such that the solution (1.3) is $2\pi$-periodic, i.e. $(x(t+2\pi), y(t+2\pi)) = (x(t), y(t))$ for all $t$, and the periodic orbit in phase space is a circle with center at origin $(0, 0)$ and radius $R = \sqrt{2H(x_0, y_0)}$. We divide time segment

$[0, 2\pi]$ in 10 evenly time intervals $[t^n, t^{n+1}]$ of length $\tau = 2\pi/10$ (time step) such that $t^n = n\tau$ for $n = 0, \ldots, 10$ and with $(x^n, y^n)$ we identify the discrete function values at time $t^n$, i.e. $(x^n, y^n) = (x(n\tau), y(n\tau))$ for $n = 0, \ldots, 10$. We solve system (1.1)–(1.2) in time till $t = 2\pi$ with three different iterative time stepping methods, explicit Euler method (ExE):

$$x^{n+1} = x^n + \tau y^n, \tag{1.5}$$

$$y^{n+1} = y^n - \tau\omega^2 x^n, \tag{1.6}$$

implicit Euler method (ImE):

$$x^{n+1} = x^n + \tau y^{n+1}, \tag{1.7}$$

$$y^{n+1} = y^n - \tau\omega^2 x^{n+1}, \tag{1.8}$$

and Störmer-Verlet method (StV):

$$y^* = y^n - \frac{\tau}{2}\omega^2 x^n, \tag{1.9}$$

$$x^{n+1} = x^n + \tau y^*, \tag{1.10}$$

$$y^{n+1} = y^* - \frac{\tau}{2}\omega^2 x^{n+1}. \tag{1.11}$$

We choose 5 different initial conditions $(x^0, y^0) = (x_0, y_0)$, depicted in Figure 1.1, which when connected with lines form a star. Different line widths of the stars indicate solutions at different times, with increasing time the line width decreases. The boldest star indicates the configuration of the initial conditions. Additionally with dashed circles we indicate the associated periodic orbits for each initial condition. We plot results every other time step. The analytical solution (1.3) in time is shown in Figure 1.1(a). The motion of the star is clockwise. Note that the exact solution at the computational final time, i.e. $t = 2\pi$, coincides with the initial condition due to the periodicity and each vertex of the star stays on the associated periodic orbit, a circle. In Figure 1.1(b) we plot the solutions of the ExE method (1.5)–(1.6) and the results are disappointing. Solutions grow in time and the vertices of the star do not stay on the associated constrained circles. The situation is no better for the ImE method (1.7)–(1.8), see Figure 1.1(c). The solutions contract towards a single point, $(0, 0)$. On the contrary, solutions of the StV method (1.9)–(1.11) in Figure 1.1(d) stay close to the associated periodic orbits, except, at the final computational time the numerical results do not match exactly with the initial conditions. There is a good explanation for this and it will become clear from the following discussion.

All three numerical methods presented above are one step methods and can be expressed in general form:

$$\begin{pmatrix} x^{n+1} \\ y^{n+1} \end{pmatrix} = A(\omega, \tau) \begin{pmatrix} x^n \\ y^n \end{pmatrix},$$

where matrix $A(\omega, \tau) \in \mathbf{R}^{2\times 2}$ depends on the frequency $\omega$ and time step $\tau$. By iteration it follows that the solution at any time $t^n$ is given by

$$\begin{pmatrix} x^n \\ y^n \end{pmatrix} = A(\omega, \tau)^n \begin{pmatrix} x^0 \\ y^0 \end{pmatrix}. \tag{1.12}$$

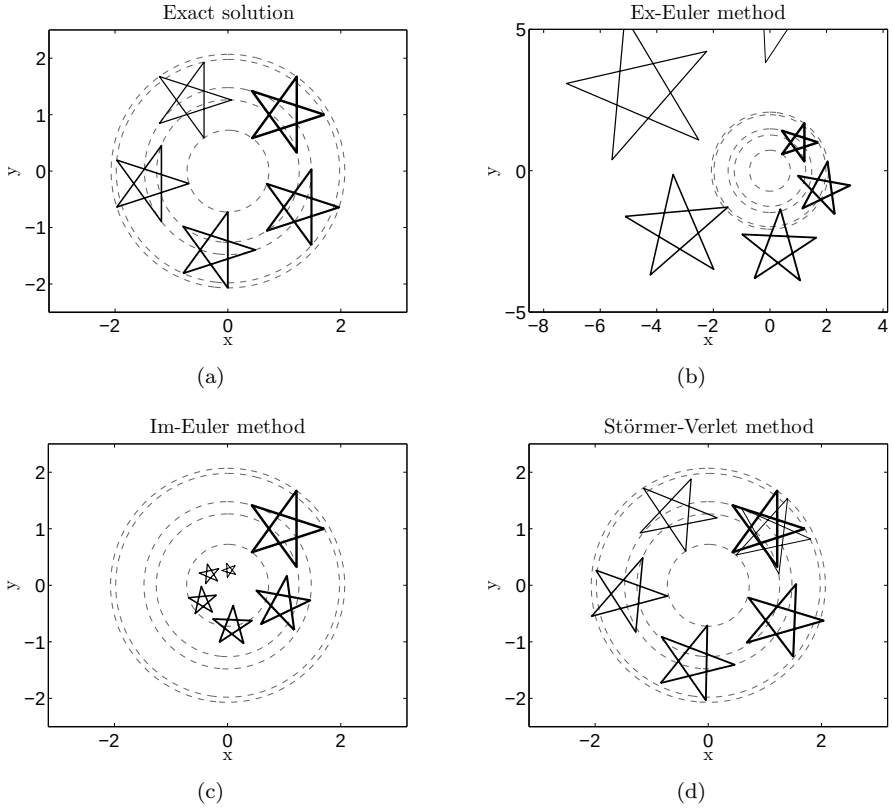



Figure 1.1: Solutions of the harmonic oscillator equations (1.1)–(1.2) with $\omega$=1. (a) exact solution, (b) numerical solution with the explicit Euler method (1.5)–(1.6), (c) numerical solution with the implicit Euler method (1.7)–(1.8), (d) numerical solution with the Störmer-Verlet method (1.9)–(1.11). The progression in time is indicated with decreasing line widths of the star.

Hence the long time solution of the iterative system (1.12) can be interiorly characterized by the eigenvalues of the matrix $A(\omega, \tau)$. We find that the eigenvalues of the matrix $A(\omega, \tau)$ for the ExE method (1.5)–(1.6) are $\lambda = 1 \pm \tau\omega i$. Since $|\lambda| > 1$ for $\tau, \omega > 0$, the ExE method is unconditionally unstable method for any time step $\tau$. This explains why solutions in Figure 1.1(b) grow in time and do not stay on periodic orbits. The eigenvalues of the matrix $A(\omega, \tau)$ for the ImE method (1.7)–(1.8) are $\lambda = (1 \pm \tau\omega i)/(1 + \tau^2\omega^2)$. Since $|\lambda| < 1$ for $\tau, \omega > 0$, the origin $(0, 0)$ is an asymptotically stable point of the method. Hence the ImE method is unconditionally stable method for any value of $\tau$ but solutions will always tend towards the origin $(0, 0)$. That is what we see in Figure 1.1(c). The eigenvalues for the StV method (1.9)–(1.11) are $\lambda = a \pm \tau\omega\sqrt{-(1 + a)/2}$ where $a = 1 - \tau^2\omega^2/2$. As long as $|\tau\omega| \leq 2$ for $\tau, \omega > 0$, the eigenvalues are of modulus one, i.e. $|\lambda| = 1$. This implies that the StV method is stable and the magnitude of the solutions do not

grow nor decay, they stay bounded for all times. This explains results in Figure 1.1(d). Note that the condition $|\tau\omega| \leq 2$ was satisfied in our computations with $\omega = 1$ and $\tau = 2\pi/10$.

Inspection of the eigenvalues of the matrix $A(\omega, \tau)$ has shown that both methods, the explicit Euler method (1.5)–(1.6) and the implicit Euler method (1.7)–(1.8), have changed the dynamical property of the stationary point of the original dynamical system (1.1)–(1.2). The stationary center point has been changed to a *source* or *sink*, respectively. On the contrary, the Störmer-Verlet method (1.9)–(1.11) has preserved this property. This will become evident from the following analysis.

With the ansatz of a single frequency $\hat{\omega} \in \mathbf{R}_+$ wave solution

$$(x(t), y(t)) = \mathrm{Re}(ae^{-i\hat{\omega}t}, be^{-i\hat{\omega}t}),$$

where $a, b \in \mathbf{C}$, we derive a so called *dispersion relation* for the harmonic oscillator equations (1.1)–(1.2):

$$\hat{\omega} = \omega, \qquad b = -i\hat{\omega}a. \tag{1.13}$$

This is exactly what we would expect from the analytical solution (1.3). With the condition $|\tau\omega| \leq 2$ and the ansatz

$$(x^n, y^n) = \mathrm{Re}(ae^{-i\hat{\omega}n\tau}, be^{-i\hat{\omega}n\tau})$$

we derive a real discrete dispersion relation for the StV method (1.9)–(1.11):

$$\hat{\omega} = \frac{\arccos\left(1 - \frac{\tau^2\omega^2}{2}\right)}{\tau} \underset{\tau\to 0}{\to} \omega, \qquad b = -ia\frac{\sin(\hat{\omega}\tau)}{\tau} \underset{\tau\to 0}{\to} -i\hat{\omega}a. \tag{1.14}$$

In the limit when $\tau \to 0$ we recover the continuous dispersion relation (1.13). From the discrete dispersion relation (1.14) follows that $\hat{\omega} \geq \omega$ for any $\tau > 0$ satisfying $|\tau\omega| \leq 2$. Hence the solutions of the StV method (1.9)–(1.11) oscillate faster than the original solution (1.3). This explains the mismatch between the exact and the discrete solutions at the final computational time in Figure 1.1(d). In fact, we can write down the exact solution of the StV method (1.9)–(1.11) for a fixed value of $\tau$:

$$x(t) = x_0 \cos(\hat{\omega}t) + y_0 \frac{\tau}{\sin(\hat{\omega}\tau)} \sin(\hat{\omega}t), \tag{1.15}$$

$$y(t) = -x_0 \frac{\sin(\hat{\omega}\tau)}{\tau} \sin(\hat{\omega}t) + y_0 \cos(\hat{\omega}t), \tag{1.16}$$

which in the limit when $\tau \to 0$ converges to the analytical solution (1.3). The analytical solution (1.15)–(1.16) is understood in the sense that $(x^n, y^n) \equiv (x(n\tau), y(n\tau))$ for all $n = 0, 1, 2, \ldots$ and for a fixed value $\tau$ satisfies the modified harmonic oscillator equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\hat{\omega}\tau}{\sin(\hat{\omega}\tau)} y, \tag{1.17}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\hat{\omega}\frac{\sin(\hat{\omega}\tau)}{\tau} x, \tag{1.18}$$

with a modified energy function

$$\hat{H}(x,y,\hat{\omega}(\tau),\tau) = \frac{1}{2}\frac{\hat{\omega}\tau}{\sin(\hat{\omega}\tau)}y^2 + \frac{1}{2}\hat{\omega}\frac{\sin(\hat{\omega}\tau)}{\tau}x^2 \tag{1.19}$$
$$\underset{\tau\to 0}{\to} H(x,y) = \frac{1}{2}y^2 + \frac{1}{2}\omega^2 x^2.$$

Hence the numerical solution of the Störmer-Verlet method (1.9)–(1.11) preserves the characteristic property of the original dynamical system (1.1)–(1.2), i.e. the characteristic of the stationary point. Complimentary, the modified energy function (1.19) is invariant under the motion of (1.17)–(1.18) and implies that periodic orbits in phase space of the StV method (1.9)–(1.11) are ellipses.

For further reference we describe the Störmer-Verlet method for the general partitioned differential equation system:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(y), \tag{1.20}$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = g(x), \tag{1.21}$$

where $x, y \in \mathbf{R}^n$ and $f, g : \mathbf{R}^n \to \mathbf{R}^n$. The Störmer-Verlet method with time step $\tau$ applied to (1.20)–(1.21) reads:

$$y^* = y^n + \frac{\tau}{2}g(x^n), \tag{1.22}$$
$$x^{n+1} = x^n + \tau f(y^*), \tag{1.23}$$
$$y^{n+1} = y^* + \frac{\tau}{2}g(x^{n+1}). \tag{1.24}$$

Note that if evaluation of the function $f(y)$ is cheaper compared to the evaluation of the function $g(x)$, then equivalently one can exchange equation for $y$ with the equation for $x$, and vice versa, in method (1.22)–(1.24).

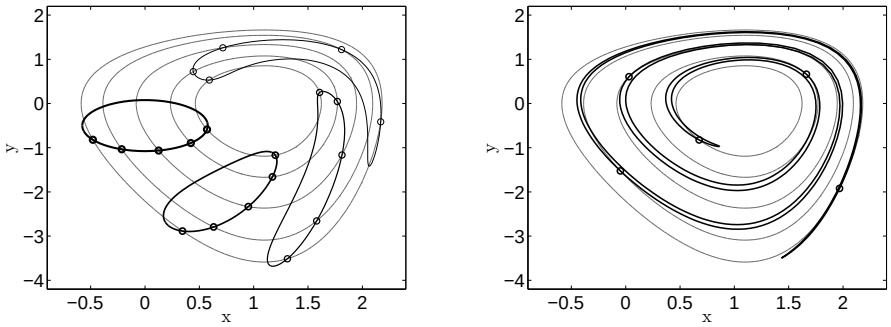
Preservation of the dynamical properties by the StV method applied to the harmonic oscillator equations (1.1)–(1.2) gives a good motivation to study structure preserving numerical methods. But do these results extend also to nonlinear dynamical systems and semi-discretized PDEs or is it just an artifact of solving linear differential equations? The answer is positive: yes, and it will become evident from the following two examples.

We give additional motivation by considering nonlinear autonomous dynamical system in $\mathbf{R}^2$:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = 1 - e^y, \tag{1.25}$$
$$\frac{\mathrm{d}y}{\mathrm{d}t} = e^x - 3, \tag{1.26}$$

which are transformed equations of *Lotka-Volterra* model in logarithmic coordinates. Lotka-Volterra models are considered in mathematical biology to model the growth of animal species. The dynamical system (1.25)–(1.26) has invariant of motion:

$$H(x,y) = y - e^y + 3x - e^x, \tag{1.27}$$




(a) (b)

Figure 1.2: Periodic orbits and numerical solutions of (1.25)–(1.26) with the Störmer-Verlet method (1.22)–(1.24), $\tau = 0.01$. (a) solutions at times $t = 0, 1, 2, 3$. (b) solutions at time $t = 50$. Progression in time is indicated by decrease in width of a closed curve connecting the solutions.

i.e.

$$\frac{\mathrm{d}H}{\mathrm{d}t} = (1 - e^y)\,\frac{\mathrm{d}y}{\mathrm{d}t} + (3 - e^x)\,\frac{\mathrm{d}x}{\mathrm{d}t} = (1 - e^y)\,(e^x - 3) + (3 - e^x)\,(1 - e^y) = 0.$$

The Jacobian matrix of the right hand side vector field of (1.25)–(1.26) and the Hessian matrix of (1.27) are

$$J = \begin{bmatrix} 0 & -e^y \\ e^x & 0 \end{bmatrix}, \qquad \nabla\nabla H(x, y) = \begin{bmatrix} -e^x & 0 \\ 0 & -e^y \end{bmatrix}, \tag{1.28}$$

respectively. The system of equations (1.25)–(1.26) has a unique stationary point $(\ln 3, 0)$ which locally is a center, since at this point the Jacobian matrix in (1.28) has purely imaginary eigenvalues $\lambda = \pm\sqrt{3}$. In fact, the Jacobian matrix has purely imaginary eigenvalues at each point $(x, y) \in \mathbf{R}^2$. Since all eigenvalues of the Hessian matrix in (1.28) are real and negative for each value of $(x, y) \in \mathbf{R}^2$, the invariant of motion (1.27) is a concave function and defines periodic orbits in phase space $\mathbf{R}^2$ for each given value of $H(x_0, y_0)$.

We solve the system of equations (1.25)–(1.26) in time with the StV method (1.22)–(1.24) and set the time step to $\tau = 0.01$. We choose a set of initial conditions that lie on a circle with center $(0, -0.5)$ and radius $1/\sqrt{3}$. In Figure 1.2(a) we plot the initial condition and three numerical solutions after each unit of time, i.e. at times $t = 1, 2, 3$. The solution propagates anticlockwise. The progression in time is indicated by decreasing width of the closed curve connecting the solutions. Additionally we pick five random initial conditions, indicated with small circles, and draw the associated periodic orbits to each of these initial conditions. Periodic orbits were computed from (1.27). In Figure 1.2(b) we show solutions at time $t = 50$. Notice that in both Figures 1.2(a) and 1.2(b) each solution indicated by a small circle stays close to the associated periodic orbit. We saw similar results in

Figure 1.1(d). In fact, the area enclosed by the curves is preserved in time. We give rigorous mathematical proof for this in Section 1.1.8.

As for the final example of this motivational subsection we consider the semi-linear 1D wave equations defined on an open interval $(0,1)$, the sine-Gordon equations:

$$\frac{\partial u}{\partial t} = v, \tag{1.29}$$

$$\frac{\partial v}{\partial t} = \frac{\partial^2 u}{\partial x^2} - \sin(u), \tag{1.30}$$

$$u(0,x) = f(x), \quad v(0,x) = g(x), \tag{1.31}$$

$$u(t,0) = u(t,1) = 0, \quad v(t,0) = v(t,1) = 0. \tag{1.32}$$

The initial boundary value problem (1.29)–(1.32) has a conserved quantity along solutions of the system: the total energy

$$\mathcal{H} = \int_0^1 \left( \frac{1}{2} v^2 + \frac{1}{2} \left( \frac{\partial u}{\partial x} \right)^2 - \cos(u) \right) \mathrm{d}x. \tag{1.33}$$

Straightforward calculations show that

$$\begin{aligned} \frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} &= \int_0^1 \left( \frac{\partial v}{\partial t} v + \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial x t} + \sin(u) \frac{\partial u}{\partial t} \right) \mathrm{d}x \\ &= \int_0^1 \left( \frac{\partial v}{\partial t} v - \frac{\partial^2 u}{\partial x^2} v + \sin(u) v \right) \mathrm{d}x = 0, \end{aligned}$$

where the boundary terms from integration by parts drop out due to the homogeneous boundary conditions (1.32).

Our objective is to consider the semi-discretized equations of (1.29)–(1.32) that preserve the discrete approximation of functional (1.33) and then integrate these in time with the StV method (1.22)–(1.24) to see if we can preserve this invariant of motion during a long time simulation. Consider $N+1$ equally spaced grid points $x_i$ on a segment $[0,1]$ and the grid size $\Delta x = 1/N$ such that $x_i = i\Delta x$ for $i = 0, \ldots, N$. The time dependent discrete values of functions $u$ and $v$ at each grid point are defined by $u_i = u(t, i\Delta x)$ and $v_i = v(t, i\Delta x)$, respectively, and their initial values are defined by $u_i^0 = u(0, i\Delta x) = f(i\Delta x)$, $v_i^0 = v(0, i\Delta x) = g(i\Delta x)$ for each $i = 0, \ldots, N$. Dirichlet boundary conditions (1.32) imply that $u_0 = u_N = 0$ and $v_0 = v_N = 0$ for all times. Note that the functions $f(x)$ and $g(x)$ should be consistent with the boundary conditions (1.32), i.e. $f(0) = f(1) = 0$ and $g(0) = g(1) = 0$. With $\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v} \in \mathbf{R}^{N-1}$ we define a vector of the grid values $x_i$ and vectors of the discrete function values $u_i$, $v_i$, respectively, for $i = 1, \ldots, N-1$. The semi-discretized sine-Gordon equations in a vector form read:

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \boldsymbol{v}, \tag{1.34}$$

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = -D_x^T D_x \boldsymbol{u} - \sin(\boldsymbol{u}), \tag{1.35}$$

$$\boldsymbol{u}^0 = f(\boldsymbol{x}), \quad \boldsymbol{v}^0 = g(\boldsymbol{x}), \tag{1.36}$$

where $f(\boldsymbol{x})$, $g(\boldsymbol{x})$ and $\sin(\boldsymbol{u})$ are understood in pointwise manner. The matrix $D_x \in \mathbf{R}^{N\times N-1}$ is a backward finite difference approximation matrix of the first order spatial derivative $\frac{\partial}{\partial x}$, defined by

$$(D_x\boldsymbol{u})_1 = \frac{u_1}{\Delta x}, \qquad (D_x\boldsymbol{u})_i = \frac{u_i - u_{i-1}}{\Delta x}, \quad i = 2,\ldots,N-1,$$

where minus its transpose, $-D_x^T \in \mathbf{R}^{(N-1)\times N}$, defines a forward finite difference approximation matrix such that

$$\left(-D_x^T\boldsymbol{u}\right)_i = \frac{u_{i+1} - u_i}{\Delta x}, \quad i = 1,\ldots,N-2, \qquad \left(-D_x^T\boldsymbol{u}\right)_{N-1} = \frac{-u_{N-1}}{\Delta x}.$$

Notice that we have included the zero boundary conditions $u_0 = u_N = 0$ into the definition of the matrix $D_x$. The product of two matrices $D_{xx} := -D_x^T D_x \in \mathbf{R}^{(N-1)\times(N-1)}$ leads to the classical three point finite difference approximation matrix of the second order spatial derivative $\frac{\partial^2}{\partial x^2}$, i.e.

$$\begin{aligned}(D_{xx}\boldsymbol{u})_i &= \frac{u_{i-1} - 2u_i + u_{i+1}}{\Delta x^2}, \quad i = 2,\ldots,N-2,\\ (D_{xx}\boldsymbol{u})_1 &= \frac{-2u_1 + u_2}{\Delta x^2}, \qquad (D_{xx}\boldsymbol{u})_{N-1} = \frac{u_{N-2} - 2u_{N-1}}{\Delta x^2}.\end{aligned}$$

The matrix $D_{xx}$ is symmetric and negative definite, and hence possesses an orthogonal basis of eigenvectors, i.e. $D_{xx} = QDQ^T$, where $Q^TQ = QQ^T = I_{N-1}$, $Q \in \mathbf{R}^{(N-1)\times(N-1)}$, $I_{N-1} \in \mathbf{R}^{(N-1)\times(N-1)}$ is an identity matrix and $D \in \mathbf{R}^{(N-1)\times(N-1)}$ is a diagonal matrix with negative entries. If we would drop the nonlinear term $\sin(\boldsymbol{u})$ from the equation (1.35) and consider the semi-discretized linear wave equations

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \boldsymbol{v}, \tag{1.37}$$

$$\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} = D_{xx}\boldsymbol{u}, \tag{1.38}$$

then in new variables $\hat{\boldsymbol{u}} := Q^T\boldsymbol{u}$ and $\hat{\boldsymbol{v}} := Q^T\boldsymbol{v}$ system (1.37)–(1.38) would reduce to the decoupled system of harmonic oscillators (1.1)–(1.2), i.e.

$$\frac{\mathrm{d}\hat{\boldsymbol{u}}}{\mathrm{d}t} = \hat{\boldsymbol{v}}, \tag{1.39}$$

$$\frac{\mathrm{d}\hat{\boldsymbol{v}}}{\mathrm{d}t} = D\hat{\boldsymbol{u}}. \tag{1.40}$$

This is exactly what we would expect in the continuous case if we were solving a linear wave equation with the method of separation of variables.

In Section 1.1.7 we explain why the discrete approximation function of the energy functional (1.33) by the quadrature rule:

$$\begin{aligned}H(\boldsymbol{u},\boldsymbol{v}) &= \left(\frac{1}{2}\boldsymbol{v}^T\boldsymbol{v} + \frac{1}{2}(D_x\boldsymbol{u})^T(D_x\boldsymbol{u}) - \cos(\boldsymbol{u})^T\mathbf{1}\right)\Delta x \\ &= \left(\frac{1}{2}\boldsymbol{v}^T\boldsymbol{v} - \frac{1}{2}\boldsymbol{u}^T(D_{xx}\boldsymbol{u}) - \cos(\boldsymbol{u})^T\mathbf{1}\right)\Delta x,\end{aligned} \tag{1.41}$$

where $\mathbf{1} \in \mathbf{R}^{N-1}$ is a vector of ones, appears to be the conserved quantity along the solution of the semi-discretized sine-Gordon equations (1.34)–(1.36). From the symmetry property of the matrix $D_{xx}$ it follows that

$$\begin{aligned}\frac{\mathrm{d}H}{\mathrm{d}t} &= \left(\boldsymbol{v}^T\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} - (D_{xx}\boldsymbol{u})^T\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} + \sin(\boldsymbol{u})^T\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}\right)\Delta x \\ &= \boldsymbol{v}^T\left(\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} - D_{xx}\boldsymbol{u} + \sin(\boldsymbol{u})\right)\Delta x = 0.\end{aligned}$$

Note that in the linear case the discrete energy (1.41) can be decoupled into the sum of the associated energies of the harmonic oscillators (1.39)–(1.40), i.e.

$$H(\hat{\boldsymbol{u}}, \hat{\boldsymbol{v}}) = \frac{1}{2}\left(\hat{\boldsymbol{v}}^T\hat{\boldsymbol{v}} - \hat{\boldsymbol{u}}^T D\hat{\boldsymbol{u}}\right)\Delta x,$$

which defines a multidimensional ellipsoid in the phase space $\mathbf{R}^{2(N-1)}$.

We study the conservation of energy (1.41) under long time integration with the Störmer-Verlet method (1.22)–(1.24). We choose $N = 100$ such that $\Delta x = 0.01$ and take $\tau = 0.01$. We consider smooth initial conditions: $f(x) = \sin(4\pi x)e^{-x}$, $g(x) = x(x-1)$, and perform $10^6$ time steps. In Figure 1.3(a) we plot in time the absolute value of the relative error of the discrete energy function (1.41). With $H^0 := H(\boldsymbol{u}^0, \boldsymbol{v}^0)$ we indicate the initial value of the energy. Evidently, the energy is not exactly conserved in time by the StV method but the errors are small and, remarkably, stay bounded during the whole computation. In Figure 1.3(b) we plot the maximum value of the absolute value of the relative error of the energy (1.41) from the simulations with different time steps $\tau$. For each simulation we keep the same computational time window, i.e. $t \in [0, 10^4]$. Figure 1.3(b) shows that the relative error of the energy decreases by factor 2 with respect to the time step $\tau$. Hence the energy (1.41) is conserved in time to second order accuracy, i.e.

$$H(t) - H^0 = O(\tau^2)$$

for long times $t$. We will address this property in Section 1.1.8.

To explain the long time approximate energy conservation by the StV method applied to the semi-discretized sine-Gordon equations (1.34)–(1.36), we require the mathematical theory that we discus in Section 1.1.8. On the contrary, the analysis of the StV method (1.9)–(1.11) extends straightforwardly to the linear semi-discretized wave equations (1.37)–(1.38). The StV method (1.22)–(1.24) applied to (1.37)–(1.38) in the vector form reads:

$$\begin{pmatrix}\boldsymbol{u}^{n+1}\\ \boldsymbol{v}^{n+1}\end{pmatrix} = A(D_{xx}, \tau)\begin{pmatrix}\boldsymbol{u}^{n}\\ \boldsymbol{v}^{n}\end{pmatrix},$$

$$A(D_{xx}, \tau) = \begin{bmatrix} I_{N-1} + \frac{\tau^2}{2}D_{xx} & \tau I_{N-1} \\ \frac{\tau}{2}D_{xx}\left(2I_{N-1} + \frac{\tau^2}{2}D_{xx}\right) & I_{N-1} + \frac{\tau^2}{2}D_{xx}\end{bmatrix},$$

where $A(D_{xx}, \tau) \in \mathbf{R}^{2(N-1)\times 2(N-1)}$ and $I_{N-1} \in \mathbf{R}^{(N-1)\times(N-1)}$ is an identity matrix. From $D_{xx} = QDQ^T$ and $QQ^T = Q^TQ = I_{N-1}$ follows

$$A(D_{xx}, \tau) = \begin{bmatrix} Q & 0 \\ 0 & Q\end{bmatrix} A(D, \tau)\begin{bmatrix} Q^T & 0 \\ 0 & Q^T\end{bmatrix},$$

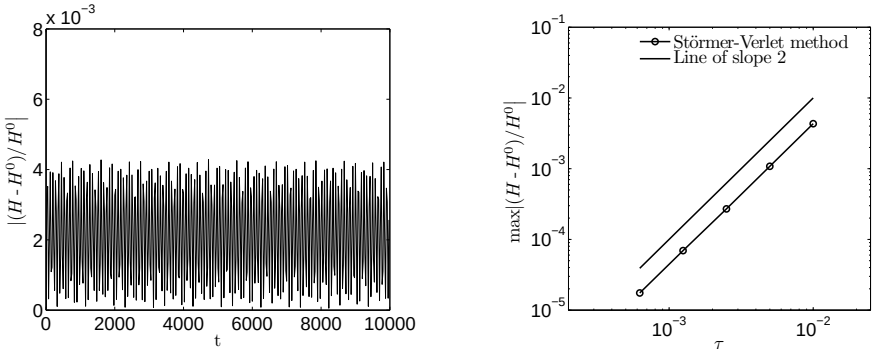



Figure 1.3: Long time integration of the semi-discretized sine-Gordon equations (1.34)–(1.36) with the Störmer-Verlet method (1.22)–(1.24), $\Delta x = 0.01$. (a) absolute value of the relative error of the energy (1.41) over the time window $[0, 10^4]$ with time step $\tau = 0.01$. (b) maximum value of the relative error of the energy (1.41) over the computational time window $[0, 10^4]$ for the different values of the time step $\tau$.

$$A(D,\tau) = \begin{bmatrix} I_{N-1} + \frac{\tau^2}{2} D & \tau I_{N-1} \\ \frac{\tau}{2} D \left( 2 I_{N-1} + \frac{\tau^2}{2} D \right) & I_{N-1} + \frac{\tau^2}{2} D \end{bmatrix},$$

where $A(D,\tau) \in \mathbf{R}^{2(N-1)\times 2(N-1)}$ is a block matrix of diagonal matrices. Since

$$A(D_{xx},\tau)^n = \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix} A(D,\tau)^n \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix},$$

in new variables $\hat{\boldsymbol{u}}^n := Q^T \boldsymbol{u}^n$ and $\hat{\boldsymbol{v}}^n := Q^T \boldsymbol{v}^n$ the StV method (1.22)–(1.24) applied to (1.37)–(1.38) reduces to the StV method applied to the decoupled system of harmonic oscillators (1.39)–(1.40), i.e.

$$\begin{pmatrix} \hat{\boldsymbol{u}}^n \\ \hat{\boldsymbol{v}}^n \end{pmatrix} = A(D,\tau)^n \begin{pmatrix} \hat{\boldsymbol{u}}^0 \\ \hat{\boldsymbol{v}}^0 \end{pmatrix}.$$

The analysis of the StV method (1.9)–(1.11) follows for each pair $(\hat{\boldsymbol{u}}_i, \hat{\boldsymbol{v}}_i)$ with stability condition

$$\left| \tau \max_i \left\{ \sqrt{|d_i|} \right\} \right| \leq 2, \quad d_i = \frac{2}{\Delta x^2} (\cos(\pi i \Delta x) - 1), \quad i = 1, \ldots, N-1,$$

where $d_i$ is the $i^{th}$ diagonal element of the matrix $D$.

This completes the motivational subsection where we considered three conservative model equations: the harmonic oscillator equations (1.1)–(1.2), the transformed Lotka-Volterra model (1.25)–(1.26) and the semi-discretized sine-Gordon equations (1.34)–(1.36). For the harmonic oscillator equations we showed that numerical methods can destroy or preserve the characteristic properties of the dynamical system

and this motivates us to study and use structure preserving numerical methods for general class(es) of equations, e.g. Hamiltonian systems and Hamiltonian PDEs. With the transformed Lotka-Volterra model we extended our discussion to nonlinear dynamical systems and showed that the method of choice, the Störmer-Verlet method (1.22)–(1.24), captured very well the characteristic properties of the nonlinear dynamical system. The motivation for using the Störmer-Verlet method will become evident in Section 1.1.8. With the final example, the semi-discretized sine-Gordon equations, we extended the discussion to structure preserving methods for conservative PDEs and illustrated its importance in numerical simulation.

We remark that the explicit and implicit Euler methods applied to the transformed Lotka-Volterra model (1.25)–(1.26) and the semi-discretized sine-Gordon equations (1.34)–(1.36) would lead to the disappointing results of the same character as for the harmonic oscillator equations.

### 1.1.2 Hamiltonian dynamics

In this thesis we are concerned with Hamiltonian dynamics of the general form:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = J\,\nabla H(X), \qquad X(0) = X_0, \tag{1.42}$$

where $X : \mathbf{R} \to \mathbf{R}^n$, $t$ is time, $J \in \mathbf{R}^{n\times n}$ is a constant skew-symmetric matrix and $H(X) : \mathbf{R}^n \to \mathbf{R}$ is the Hamiltonian function. Differential equation (1.42) is an *autonomous* dynamical system with initial condition $X_0 \in \mathbf{R}^n$. We assume that the Hamiltonian function $H(X)$ is at least twice continuously differentiable on a connected nonempty open set $\Omega \subset \mathbf{R}^n$, *phase space* of $X$, such that the standard existence and uniqueness theorems apply to the corresponding initial value problem (1.42) in the open neighborhood of $(X_0, 0) \in \Omega \times \mathbf{R}$. With the product $\Omega \times \mathbf{R}$ we identify the *extended phase space* of equation (1.42).

Since matrix $J$ is skew-symmetric, from equation (1.42) follows two very important properties of the Hamiltonian dynamics. The first is the conservation of the Hamiltonian function $H(X)$ along the solution of the system (1.42), i.e.

$$\frac{\mathrm{d}H(X)}{\mathrm{d}t} = \nabla H(X)^T\,\frac{\mathrm{d}X}{\mathrm{d}t} = \nabla H(X)^T\,J\,\nabla H(X) \equiv 0.$$

This implies that the Hamiltonian function $H(X)$ is a *first integral* of the system (1.42). Hamiltonian function $H(X)$ may not be the only conserved quantity of (1.42). Thus for the function $I(X) : \mathbf{R}^n \to \mathbf{R}$ to be a first integral of the system (1.42), the following relation must hold:

$$\nabla I(X)^T\,J\,\nabla H(X) = 0. \tag{1.43}$$

It is easy to check that if $I_1(X), I_2(X) : \mathbf{R}^n \to \mathbf{R}$ are two first integrals of the system then also a function $I_3(X) : \mathbf{R}^n \to \mathbf{R}$ defined by

$$I_3(X) = \nabla I_1(X)^T\,J\,\nabla I_2(X)$$

is also a first integral of the system (1.42). We address related questions to the construction of the first integral preserving numerical schemes for the Hamiltonian systems in Sections 1.1.7–1.1.8 and in Chapters 2 and 4.

The second property shows that the right hand side vector field of (1.42) is divergence free, i.e.

$$\nabla \cdot (J \nabla H(X)) = \operatorname{trace}(J \nabla_{XX} H(X)) \equiv 0, \tag{1.44}$$

where $\nabla_{XX} H(X)$ is a symmetric Hessian matrix of Hamiltonian $H(X)$. This result follows from the property that the trace of the product of a symmetric and a skew-symmetric matrix is equal to zero.

The divergence free property (1.44) implies volume preservation in the phase space $\Omega$ by the flow map $\Phi_H^t$ of the Hamiltonian system (1.42), and vice versa. As long as the solution of (1.42) exists at time $t$, it is defined by

$$X = \Phi_H^t(X_0), \qquad X_0 = \Phi_H^0(X_0).$$

By definition $\Phi_H^t$ defines a transformation from $X_0$ to $X$ and maps the phase space $\Omega$ into itself. Additionally, flow maps $\Phi_H^t$ of time $t$ as a one-parameter operator family define a commutative group. Under volume preservation by the flow map $\Phi_H^t$ we understand that for any bounded subset $\mathcal{U} \subset \Omega$ for which $\Phi_H^t(\mathcal{U})$ exists, volumes and orientation of $\mathcal{U}$ and $\Phi_H^t(\mathcal{U})$ are the same, i.e.

$$\int_{\mathcal{U}} \mathrm{d}X_0 = \int_{\Phi_H^t(\mathcal{U})} \mathrm{d}X.$$

As a standard rule for the change of variables under the integral sign, for the transformation to be volume preserving, the determinant of the Jacobian matrix of $\Phi_H^t(X_0)$ must be equal to 1:

$$\left| \frac{\partial \Phi_H^t(X_0)}{\partial X_0} \right| = 1, \qquad \forall\, t,\, X_0. \tag{1.45}$$

The associated matrix-valued variational equation of (1.42) is

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = A(t)Y, \qquad Y(0) = I_n, \tag{1.46}$$

where $Y(t) = \frac{\partial \Phi_H^t(X_0)}{\partial X_0}$, $A(t) = J \nabla_{XX} H(X)$ at $X = \Phi_H^t(X_0)$ and $I_n \in \mathbf{R}^{n \times n}$ is an identity matrix. From the *Abel-Liouville-Jacobi-Ostrogradskii identity* and equation (1.44) follow that

$$\frac{\mathrm{d}}{\mathrm{d}t}|Y| = \operatorname{trace}(A(t))|Y| = \nabla \cdot (J \nabla H(X))|Y| \equiv 0, \qquad \forall\, t,\, X_0. \tag{1.47}$$

This implies the statement (1.45) and proves the following lemma:

**Lemma 1.1.2.1.** *The flow map $\Phi_H^t(X_0)$ of the Hamiltonian system (1.42) is volume preserving, statement (1.45), if and only if $\nabla \cdot (J \nabla H(X)) = 0$ for all $X$.*

The divergence free property of the right hand side vector field of the Hamiltonian system (1.42), or of any autonomous dynamical system, plays an important role for the statistical mechanics. We address these questions in Section 1.2.1 and Chapters 3 and 4.

Both properties described above directly follow from the specific form of the equation (1.42) and that matrix $J$ is skew-symmetric. Here we mention another property of (1.42) under the following conditions. We call Hamiltonian system (1.42) *time reversible* under the linear coordinate transformation:

$$\hat{t} = -t, \tag{1.48}$$

$$\hat{X} = SX, \tag{1.49}$$

where $S \in \mathbf{R}^{n\times n}$ is a nonsingular matrix, if the following conditions hold:

$$H(X) = H(\hat{X}), \qquad J = -SJS^T. \tag{1.50}$$

These conditions imply that the linear transformation (1.48)–(1.49) does not alter the dynamical system (1.42), i.e.

$$\frac{\mathrm{d}\hat{X}}{\mathrm{d}\hat{t}} = -SJ\,\nabla H(X) = -SJS^T\,\nabla H(\hat{X}) = J\,\nabla H(\hat{X}).$$

As an example we show time reversibility property with respect to the *involution* $S$, i.e. $SS = I_n$, for the block skew-symmetric matrix $J$ and quadratic Hamiltonian function:

$$S = \begin{bmatrix} I_m & 0 \\ 0 & -I_k \end{bmatrix}, \qquad J = \begin{bmatrix} 0 & K \\ -K^T & 0 \end{bmatrix}, \qquad H(X) = \frac{1}{2}X^TX,$$

where $I_m \in \mathbf{R}^{m\times m}$, $I_k \in \mathbf{R}^{k\times k}$ are identity matrices, $m + k = n$ and $K \in \mathbf{R}^{m\times k}$ is some arbitrary matrix. Simple calculations yield:

$$H(\hat{X}) = \frac{1}{2}(SX)^TSX = \frac{1}{2}X^TS^TSX = \frac{1}{2}X^TX = H(X)$$

and

$$SJS^T = \begin{bmatrix} 0 & -I_mKI_k \\ I_nK^TI_m & 0 \end{bmatrix} = \begin{bmatrix} 0 & -K \\ K^T & 0 \end{bmatrix} = -J.$$

Hence the conditions (1.50) are satisfied and the associated Hamiltonian system is time reversible with respect to the involution $S$.

Recall that any skew-symmetric matrix of even dimension $2m$ with full rank is invertible and has $m$-pairs of nonzero purely complex conjugate eigenvalues. On the contrary, any odd dimensional skew-symmetric matrix is singular. Let us assume that the system matrix $J$ of (1.42) has rank $2m$ and $n = 2m + k$ where $k$ is odd. Hence $J$ is a singular matrix with $m$-pairs of purely complex conjugate eigenvalues and $k$ zero eigenvalues. This leads to the existence of $k$ linear *distinguished functions*, *Casimirs*:

$$C(X) = CX, \qquad CJ \equiv 0, \tag{1.51}$$

where $C \in \mathbf{R}^{k\times n}$. From the relation (1.43) follows that Casimir functions (1.51) are first integrals of the system (1.42).

The following derivation is a special result of the *Darboux-Lie* theorem. Consider two skew-symmetric matrices defined by

$$\hat{J} = MJM^T$$

and

$$J_{Id} = \begin{bmatrix} 0 & I_m \\ -I_m & 0 \end{bmatrix}, \tag{1.52}$$

where $I_m \in \mathbf{R}^{m\times m}$ is an identity matrix and $M \in \mathbf{R}^{2m\times n}$ is an arbitrary matrix with rank $2m$ such that the matrix $\hat{J} \in \mathbf{R}^{2m\times 2m}$ is skew-symmetric of even dimension with rank $2m$. We can always construct such a matrix $M$ by setting columns of $M$ associated to the linearly independent columns of matrix $J$ to the unit vectors of $\mathbf{R}^{2m}$ and by setting the rest of the columns of matrix $M$ to zero. For example, if matrix $J \in \mathbf{R}^{2m\times 2m}$ is of full rank then $M = I_{2m}$.

Real orthogonal decompositions of skew-symmetric matrices $\hat{J}$ and $J_{Id}$ are given by

$$\hat{J} = U\Lambda U^T, \qquad J_{Id} = V\Sigma V^T,$$

where $U, \Lambda, V, \Sigma \in \mathbf{R}^{2m\times 2m}$, $UU^T = U^TU = I_{2m}$, $VV^T = V^TV = I_{2m}$. Matrices $\Lambda = -\Lambda^T$ and $\Sigma = -\Sigma^T$ are block diagonal skew-symmetric matrices containing the imaginary parts of the eigenvalues of the matrices $\hat{J}$ and $J_{Id}$, respectively. Note that the following relation holds:

$$\Sigma = \left(\Lambda^T\Lambda\right)^{-\frac{1}{4}} \Lambda \left(\Lambda^T\Lambda\right)^{-\frac{1}{4}T},$$

where matrix $\Lambda^T\Lambda$ is diagonal with positive entries.

With the linear transformation

$$\begin{pmatrix} Z \\ z \end{pmatrix} = SX, \qquad S = \begin{bmatrix} \hat{M}M \\ C \end{bmatrix}, \qquad \hat{M} = V\left(\Lambda^T\Lambda\right)^{-\frac{1}{4}} U^T, \tag{1.53}$$

where $Z \in \mathbf{R}^{2m}$, $z \in \mathbf{R}^k$, $\hat{M} \in \mathbf{R}^{2m\times 2m}$ and matrix $C$ defines linear Casimir functions (1.51), system (1.42) transforms into

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} Z \\ z \end{pmatrix} &= SJS^T \begin{pmatrix} \nabla_Z H(Z,z) \\ \nabla_z H(Z,z) \end{pmatrix} = \begin{bmatrix} \hat{M}M \\ C \end{bmatrix} J \begin{bmatrix} M^T\hat{M}^T & C^T \end{bmatrix} \begin{pmatrix} \nabla_Z H(Z,z) \\ \nabla_z H(Z,z) \end{pmatrix} \\ &= \begin{bmatrix} \hat{M}\hat{J}\hat{M}^T & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \nabla_Z H(Z,z) \\ \nabla_z H(Z,z) \end{pmatrix} = \begin{bmatrix} J_{Id} & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} \nabla_Z H(Z,z) \\ \nabla_z H(Z,z) \end{pmatrix}, \end{aligned}$$

since

$$\hat{M}\hat{J}\hat{M}^T = V\left(\Lambda^T\Lambda\right)^{-\frac{1}{4}} U^T \left(U\Lambda U^T\right) U \left(\Lambda^T\Lambda\right)^{-\frac{1}{4}T} V^T = V\Sigma V^T = J_{Id}.$$

Hence any odd, $n = 2m + k$, dimensional Hamiltonian system (1.42) of rank $2m$ can be transformed into the even dimensional $2m$ Hamiltonian system, plus $k$ trivial dynamics. Since $z \equiv const$, we can formally neglect $z$ and write transformed Hamiltonian system with respect to $Z$ only, i.e.

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = J_{Id}\,\nabla H(Z). \tag{1.54}$$

We call system (1.42) with structure matrix (1.52), i.e. equation (1.54), a *canonical* Hamiltonian dynamical system and *noncanonical* otherwise. Evidently, the derivation above applies to any Hamiltonian system (1.42) with nonzero skew-symmetric system matrix $J$.

This completes the proof of the following proposition:

**Proposition 1.1.2.1.** *Any Hamiltonian system (1.42) with a nonzero constant skew-symmetric system matrix $J$ can be transformed in the canonical form (1.54), possibly on a reduced phase space.*

We illustrate the analysis above for the simple example in $\mathbf{R}^3$. Consider the Hamiltonian system (1.42) with matrix $J$, the associated Casimir function's row matrix $C$ and transformations matrix $S$:

$$J = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}, \qquad S = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

respectively. Then the system transforms into a canonical Hamiltonian system, plus one trivial dynamics

$$SJS^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} S^T = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

### 1.1.3 Canonical Hamiltonian systems

In this subsection we discuss canonical Hamiltonian systems, i.e. equation (1.42) with canonical matrix (1.52), from the perspective of classical mechanics. In classical mechanics we are concerned with $2n$-dimensional canonical Hamiltonian systems where $X = (q, p)^T$:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \nabla_p H(q, p), \tag{1.55}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla_q H(q, p), \tag{1.56}$$

subject to the initial conditions $X_0 = (q_0, p_0)^T$. Variables $q, p : \mathbf{R} \to \mathbf{R}^n$ are generalized coordinates and momenta, respectively, and $H(q, p) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ is the Hamiltonian function. Note that two examples considered in Section 1.1.1, i.e. harmonic oscillator equations (1.1)–(1.2) and transformed Lotka-Volterra model (1.25)–(1.26) can be written in canonical Hamiltonian form (1.55)–(1.56) with Hamiltonian functions (1.4) and (1.27), respectively. Additionally, this implies that the flow maps of the both equations are volume preserving.

In passing we mention that equations (1.55)–(1.56) can be derived from the Euler-Lagrange equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} \nabla_{\dot{q}} L(q, \dot{q}) - \nabla_q L(q, \dot{q}) = 0, \tag{1.57}$$

where $\dot{q} = \frac{\mathrm{d}q}{\mathrm{d}t}$ and $L(q, \dot{q}) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ is the Lagrange function. Here we assume that Lagrange function $L(q, \dot{q})$ does not explicitly depend on time $t$. Equation (1.57) is derived from the action integral:

$$S[q] = \int_{t_0}^{t_1} L(q, \dot{q}) \, \mathrm{d}t,$$

applying Hamilton's principle, i.e. $\frac{\delta S}{\delta q} = 0$, where $\frac{\delta S}{\delta q}$ is a variational derivative of $S[q]$. By using $p$ as an independent variable instead of $\dot{q}$ and applying a Legendre transformation, Hamilton's principle applied to the action integral:

$$S_H[q, p] = \int_{t_0}^{t_1} \left( H(q, p) - p \cdot \frac{\mathrm{d}q}{\mathrm{d}t} \right) \mathrm{d}t,$$

yields system (1.55)–(1.56). With the dot we indicate the Euclidean inner product of two vectors. Straightforward calculations show that

$$\begin{aligned} \delta S_H[q, p] &= \int_{t_0}^{t_1} \left( \nabla_q H(q, p) \cdot \delta q + \nabla_p H(q, p) \cdot \delta p - p \cdot \frac{\mathrm{d}\delta q}{\mathrm{d}t} - \frac{\mathrm{d}q}{\mathrm{d}t} \cdot \delta p \right) \mathrm{d}t \\ &= \int_{t_0}^{t_1} \left( \nabla_q H(q, p) \cdot \delta q + \nabla_p H(q, p) \cdot \delta p + \frac{\mathrm{d}p}{\mathrm{d}t} \cdot \delta q - \frac{\mathrm{d}q}{\mathrm{d}t} \cdot \delta p \right) \mathrm{d}t, \end{aligned}$$

where we used integration by parts and boundary terms drop out, since $\delta q = \delta p = 0$ on the boundary. Hence

$$\begin{aligned} \frac{\delta}{\delta p} S_H[q, p] &= \nabla_p H(q, p) - \frac{\mathrm{d}q}{\mathrm{d}t} = 0, \\ \frac{\delta}{\delta q} S_H[q, p] &= \nabla_q H(q, p) + \frac{\mathrm{d}p}{\mathrm{d}t} = 0, \end{aligned}$$

and we recover the system of equations (1.55)–(1.56).

When the Hamiltonian function is directly related to the total energy of the system, e.g. Hamiltonian function (1.4) of the harmonic oscillator equations (1.1)–(1.2), we often deal with separable Hamiltonian functions, i.e.

$$H(q, p) = K(p) + V(q),$$

where

$$K(p) = \frac{1}{2} p^T M^{-1} p$$

is a kinetic energy with symmetric and positive definite mass matrix $M \in \mathbf{R}^{n \times n}$, and $V(q) : \mathbf{R}^n \to \mathbf{R}$ is a potential energy function. In this case (1.55)–(1.56) reduces to

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1} p, \tag{1.58}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla V(q). \tag{1.59}$$

The system of equations (1.58)–(1.59) naturally arises from the equations of Newton's $2^{nd}$ law, i.e. mass times acceleration equals to force:

$$M\frac{\mathrm{d}^2q}{\mathrm{d}t^2} = F(q), \tag{1.60}$$

with conservative force $F(q) = -\nabla V(q)$. With $p = M\frac{\mathrm{d}q}{\mathrm{d}t}$ and nonsingular matrix $M$ equation (1.60) can be cast in form (1.58)–(1.59). Since the kinetic energy $K(p)$ is quadratic with respect to $p$, system (1.58)–(1.59) is time reversible with respect to the involution matrix $S$, considered in the previous subsection, by taking $m = k = n$. For the general system (1.55)–(1.56) to be time reversible the sufficient condition is: $H(q,p) = H(q,-p)$. From this condition follows that the harmonic oscillator equations (1.1)–(1.2) are time reversible but the equations of the transformed Lotka-Volterra model (1.25)–(1.26) are not.

As an example we consider mathematical pendulum equation:

$$\frac{\mathrm{d}^2q}{\mathrm{d}t^2} = -\sin(q), \tag{1.61}$$

where mass of the bob (point of mass), the length of the rod and the acceleration of gravity are set to unity. Then with $p = \frac{\mathrm{d}q}{\mathrm{d}t}$, $K(p) = p^2/2$ and $V(q) = -\cos(q)$ equation (1.61) can be written in the Hamiltonian form (1.58)–(1.59), i.e.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = p, \tag{1.62}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\sin q. \tag{1.63}$$

Note that in the mathematical pendulum equations (1.62)–(1.63) variable $q$ describes the rotation angle of the pendulum. By introducing transformation (parametrization):

$$x = \sin(q), \tag{1.64}$$

$$y = -\cos(q), \tag{1.65}$$

the system of equations (1.62)–(1.63) can be derived from the mathematical pendulum equations in Cartesian coordinates $(x,y) \in \mathbf{R}^2$:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -2x\lambda, \tag{1.66}$$

$$\frac{\mathrm{d}^2y}{\mathrm{d}t^2} = -1 - 2y\lambda, \tag{1.67}$$

$$0 = x^2 + y^2 - 1, \tag{1.68}$$

where $\lambda \in \mathbf{R}$ is Lagrange multiplier. System (1.66)–(1.68) belongs to the class of Hamiltonian (Newtonian) dynamical systems with holonomic constraints. The augmented Hamiltonian function is given by

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

Then the system of equations read:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \tag{1.69}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla V(q) - \nabla g(q)^T\lambda, \tag{1.70}$$

$$0 = g(q), \tag{1.71}$$

where the function $g : \mathbf{R}^n \to \mathbf{R}^m$ (at least twice continuously differentiable) defines the configurational manifold $\mathcal{M}$ of co-dimension $m$:

$$\mathcal{M} = \{q \in \mathbf{R}^n \,|\, g(q) = 0\},$$

and Lagrange multiplier $\lambda \in \mathbf{R}^m$ is introduced to enforce the constraint (1.71). The time derivative of (1.71), i.e. $\nabla g(q)M^{-1}p = 0$, implies that momentum $p$ belongs to the tangent plane of the constraint manifold at position $q$. The tangent space for given $q \in \mathcal{M}$ is defined by

$$\mathcal{T}_q\mathcal{M} = \left\{p \in \mathbf{R}^n \,\middle|\, \nabla g(q)M^{-1}p = 0\right\}.$$

Hence the associated phase space of system (1.69)–(1.71) is the tangent bundle denoted by

$$\mathcal{T}\mathcal{M} = \left\{q, p \in \mathbf{R}^n \,\middle|\, q \in \mathcal{M}, \nabla g(q)M^{-1}p = 0\right\}.$$

Additionally, since $\tilde{H}(q,p) = \tilde{H}(q,-p)$, system (1.69)–(1.71) is time reversible.

Example equations (1.66)–(1.68) can be written in the form (1.69)–(1.71) with augmented Hamiltonian

$$\tilde{H}\left(x, y, \frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \lambda\right) = \frac{1}{2}\left(\frac{\mathrm{d}x}{\mathrm{d}t}^2 + \frac{\mathrm{d}y}{\mathrm{d}t}^2\right) + y + (x^2 + y^2 - 1)\lambda$$

such that $q = (x, y)^T$, $p = (\frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t})^T$ and $\lambda \in \mathbf{R}$.

Similarly to system (1.55)–(1.56), the system of equations (1.69)–(1.71) can be derived from the Euler-Lagrange equations with holonomic constraints. In Chapter 3 we discuss Hamiltonian dynamics with holonomic constraints in depth. There we also introduce a canonical parametrization of the associated phase space of (1.69)–(1.71). Note that the parametrization (1.64)–(1.65) is a canonical mapping for the mathematical pendulum equations (1.66)–(1.68), i.e. with (1.64)–(1.65) we can transform system (1.66)–(1.68) with constraints into the canonical Hamiltonian system (1.62)–(1.63) without constraints.

### 1.1.4 Symplecticity

In this subsection we describe symplecticity property of the canonical Hamiltonian system (1.55)–(1.56), i.e. system (1.42) with canonical matrix $J = J_{Id}$ and $X = (q,p)^T$. Symplecticity is a characteristic property of solutions to the Hamiltonian system rather than a property of the specific form of the Hamiltonian equations. We say that the flow map $\Phi_H^t(X_0)$ of a differential equation is canonically symplectic if

$$\frac{\partial \Phi_H^t(X_0)}{\partial X_0}^T J^{-1} \frac{\partial \Phi_H^t(X_0)}{\partial X_0} = J^{-1} \tag{1.72}$$

holds for any value of $t$ and $X_0$ for which the map is defined. From $Y(t) = \frac{\partial \Phi_H^t(X_0)}{\partial X_0}$, which satisfies the variational equation (1.46), we see that the condition (1.72) is true at $t = 0$. Hence we are left to show that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(Y^T J^{-1} Y\right) = 0.$$

We find that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(Y^T J^{-1} Y\right) &= Y^T J^{-1} \frac{\mathrm{d}Y}{\mathrm{d}t} + \frac{\mathrm{d}Y}{\mathrm{d}t}^T J^{-1} Y \\ &= Y^T J^{-1} \left(J \nabla_{XX} H(X) Y\right) + \left(Y^T \nabla_{XX} H(X) J^T\right) J^{-1} Y \\ &= Y^T \nabla_{XX} H(X) Y - Y^T \nabla_{XX} H(X) Y = 0,\end{aligned}$$

where $X = \Phi_H^t(X_0)$. This completes the proof of the following theorem:

**Theorem 1.1.4.1.** *The flow map $\Phi_H^t(X_0)$ of canonical Hamiltonian system (1.55)–(1.56) with $X_0 = (q_0, p_0)^T$ is symplectic for any value of $t$ and $X_0 \in \Omega$ for which the map is defined.*

As a remark we state that with *consistent initial values* $X_0 = (q_0, p_0) \in \mathcal{TM}$, the symplecticity property of the flow map of the constrained Hamiltonian system (1.69)–(1.71) can be shown.

Note that the symplecticity property (1.72) implies volume preservation (1.45). Since matrix $J$ is nonsingular, taking determinants of both sides of (1.72) yields:

$$\begin{aligned}\left|\frac{\partial \Phi_H^t(X_0)}{\partial X_0}^T J^{-1} \frac{\partial \Phi_H^t(X_0)}{\partial X_0}\right| &= \left|J^{-1}\right|, \\ \left|\frac{\partial \Phi_H^t(X_0)}{\partial X_0}\right| &= \pm 1.\end{aligned}$$

Since at time $t = 0$ the determinant is equal to one and satisfies the equation (1.47) for all $t$ and $X_0$, the value $-1$ can be excluded. This proves the statement.

Consider the flow map of (1.55)–(1.56) expressed in the following form:

$$\begin{aligned}q &= Q_H^t(q_0, p_0), \\ p &= P_H^t(q_0, p_0).\end{aligned}$$

Then it can be shown that symplecticity property (1.72) equivalently can be stated in the following form:

$$\mathrm{d}q \wedge \mathrm{d}p = \mathrm{d}q_0 \wedge \mathrm{d}p_0, \tag{1.73}$$

where $\mathrm{d}q$ and $\mathrm{d}p$ are differential 1-forms in vector representation, i.e.

$$\mathrm{d}q = \frac{\partial Q_H^t(q_0, p_0)}{\partial q_0}\mathrm{d}q_0 + \frac{\partial Q_H^t(q_0, p_0)}{\partial p_0}\mathrm{d}p_0, \tag{1.74}$$

$$\mathrm{d}p = \frac{\partial P_H^t(q_0, p_0)}{\partial q_0}\mathrm{d}q_0 + \frac{\partial P_H^t(q_0, p_0)}{\partial p_0}\mathrm{d}p_0. \tag{1.75}$$

The bilinear skew-symmetric wedge product $\wedge$ of two differential 1-forms gives a new entity called a differential 2-form. Equation (1.73) is a conservation law of differential 2-form under the flow of the Hamiltonian system (1.55)–(1.56).

Equation (1.73) expresses the following. Consider the oriented *two-dimensional* surfaces $\mathcal{D}$ in $2n$-dimensional phase space $\Omega$. Here, $\mathcal{D}_i$, $i = 1, \ldots, n$ indicate the projections onto the $n$ two-dimensional planes of the variables $(q_i, p_i)$. Then the sum of these two-dimensional oriented areas of these projections is expressed with differential 2-form $\mathrm{d}q \wedge \mathrm{d}p$ and conserved in time. A direct consequence of (1.73) is that when $n = 1$ the symplecticity property (1.72) coincides with area preservation (1.45), since $\mathrm{d}q \wedge \mathrm{d}p$ represents oriented area in two-dimensional phase space.

Differential forms provide an elegant way to check symplecticity of Hamiltonian systems and their numerical approximations, see Section 1.1.8. As an example we show that the canonical Hamiltonian system (1.55)–(1.56) considered in the general form (1.42) conserves differential 2-form $\mathrm{d}q \wedge \mathrm{d}p$. By differentiating equation (1.42) we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{d}X = J\,\nabla_{XX}H(X)\mathrm{d}X.$$

Note the resemblance to the matrix-valued variational equation (1.46). Inverting matrix $J$ and taking the wedge product with $\mathrm{d}X$ we find that

$$\mathrm{d}X \wedge J^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{d}X = \mathrm{d}X \wedge (\nabla_{XX}H(X)\mathrm{d}X) = 0,$$

where we used the wedge product property that for any symmetric matrix $A$ the wedge product $\mathrm{d}X \wedge A\mathrm{d}X = 0$. Hence

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{d}X \wedge J^{-1}\mathrm{d}X\right) = 0.$$

With $\mathrm{d}X = (\mathrm{d}q, \mathrm{d}p)^T$

$$\mathrm{d}X \wedge J^{-1}\mathrm{d}X = -\mathrm{d}q \wedge \mathrm{d}p + \mathrm{d}p \wedge \mathrm{d}q = -2\mathrm{d}q \wedge \mathrm{d}p,$$

which proves the statement.

Symplecticity is a characteristic property of flow maps of canonical Hamiltonian systems. In Section 1.1.2 we saw that any general Hamiltonian system (1.42) can be transformed into the canonical Hamiltonian system (1.54) on a reduced phase space. So far we have neglected the possibility for the matrix $J$ to be explicitly dependent on $X$. Hence in the following subsection we address related questions to the noncanonical Hamiltonian systems with $X$ dependent matrices $J$, i.e. Poisson systems.

### 1.1.5 Poisson bracket

The Poisson bracket of any two smooth functions $F(X), G(X) : \mathbf{R}^n \to \mathbf{R}$ is defined by

$$\{F, G\} = \nabla F(X)^T J\,\nabla G(X), \tag{1.76}$$

where $J$ is a system matrix of Hamiltonian system (1.42). Due to the skew-symmetry of matrix $J$ and the standard rules of calculus the Poisson bracket (1.76) is bilinear, skew-symmetric, i.e. $\{F, G\} = -\{G, F\}$, it satisfies *Leibniz'* rule

$$\{FG, H\} = F\{G, H\} + G\{F, H\} \tag{1.77}$$

and the *Jacobi identity*

$$\{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} = 0, \tag{1.78}$$

where $F(X), G(X), H(X) : \mathbf{R}^n \to \mathbf{R}$ are three arbitrary smooth functions of $X$. The definition of the Poisson bracket is motivated by the fact that time derivative of any smooth function $F(X) : \mathbf{R}^n \to \mathbf{R}$ along the solution of the Hamiltonian system (1.42) is

$$\frac{\mathrm{d}F(X)}{\mathrm{d}t} = \nabla F(X)^T J \nabla H(X) = \{F, H\}.$$

Hence in a vector notation with function $F(X) = X$ we recover Hamiltonian system (1.42), i.e.

$$\frac{\mathrm{d}X}{\mathrm{d}t} = \{X, H\} = J \nabla H(X).$$

The conservation property of the Hamiltonian function follows from the skew-symmetry of the Poisson bracket, i.e. $\{H, H\} = 0$, and the condition (1.43) for the function $I(X) : \mathbf{R}^n \to \mathbf{R}$ to be the first integral of the system reduces to $\{I, H\} = 0$. The Jacobi identity (1.78) implies that if $I_1(X), I_2(X) : \mathbf{R}^n \to \mathbf{R}$ are two first integrals, then their Poisson bracket $\{I_1, I_2\}$ is again a first integral. The divergence free property $\nabla \cdot \{X, H\} = 0$ follows from the skew-symmetry property of the constant matrix $J$. Here we see that the basic properties of the Hamiltonian system (1.42) considered in Section 1.1.2 are shared with properties of the Poisson bracket (1.76). Additionally, with the canonical transformation (1.53) we can transform the Poisson bracket (1.76) to its canonical form:

$$\{F, G\} = \nabla F(Z)^T J_{Id} \nabla G(Z). \tag{1.79}$$

The system

$$\frac{\mathrm{d}X}{\mathrm{d}t} = J(X) \nabla H(X), \qquad X(0) = X_0, \tag{1.80}$$

where $X \in \mathbf{R}^n$, is called a Poisson system if the associated Poisson bracket:

$$\{F, G\} = \nabla F(X)^T J(X) \nabla G(X), \tag{1.81}$$

is bilinear, skew-symmetric and satisfies Leibniz' rule and the Jacobi identity. It can be shown that these properties are satisfied as long as the following condition for the skew-symmetric matrix $J(X)$ holds for all $i, j, k = 1, \ldots, n$:

$$\sum_{l=1}^{n} \left( \frac{\partial J_{ij}(X)}{\partial X_l} J_{lk}(X) + \frac{\partial J_{jk}(X)}{\partial X_l} J_{li}(X) + \frac{\partial J_{ki}(X)}{\partial X_l} J_{lj}(X) \right) = 0.$$

Note that these conditions are not sufficient to guarantee the divergence free property $\nabla \cdot \{X, H\} = 0$. We illustrate this with an example at the end of this subsection.

The fact that the Poisson bracket (1.81) and the canonical Poisson bracket (1.79) satisfy the same properties leads to the celebrated *Darboux-Lie* theorem, which states that every Poisson system (1.80) can be at least locally written in canonical Hamiltonian form (1.54) after the suitable change of coordinates, hence it is at least locally symplectic and volume preserving in the new coordinates. Proposition 1.1.2.1 is a special case of the Darboux-Lie theorem for the constant system matrix $J$. Importantly, Darboux-Lie theorem implies existence of the *Casimir* functions $C(X)$, not necessary linear, such that $\{C, F\} = 0$ for all differentiable functions $F(X)$.

Flow map $\Psi_H^t(X_0)$ of the Poisson system (1.80) when it is defined satisfies

$$\frac{\partial \Psi_H^t(X_0)}{\partial X_0} J(\Psi_H^t(X_0)) \frac{\partial \Psi_H^t(X_0)}{\partial X_0}^T = J(\Psi_H^t(X_0)). \tag{1.82}$$

The proof follows from the Darboux-Lie theorem. When the matrix $J$ is a canonical matrix (1.52), the condition (1.82) is equivalent to the symplecticity condition (1.72). By taking the inverse of (1.82) and multiplying with the Jacobian matrix of $\Psi_H^t(X_0)$ from the right and with its transpose from the left we recover (1.72). It is evident that the volume preservation does not follow from the condition (1.82), since matrix $J(X)$ may be singular and computation of determinants then would be meaningless. Hence in general volume preservation is at least a local property of the transformed Poisson system due to the Darboux-Lie theorem.

We illustrate the result of the Darboux-Lie theorem with an example of a Poisson system on the phase space $\mathbf{R}^3_+$: Lotka-Volterra model

$$\frac{\mathrm{d}\hat{z}}{\mathrm{d}t} = \begin{pmatrix} z_1(z_2+z_3) \\ z_2(z_1-z_3+1) \\ z_3(z_1+z_2+1) \end{pmatrix} = \begin{bmatrix} 0 & z_1z_2 & z_1z_3 \\ -z_1z_2 & 0 & -z_2z_3 \\ -z_1z_3 & z_2z_3 & 0 \end{bmatrix} \nabla H(\hat{z}), \tag{1.83}$$

$$H(\hat{z}) = -z_1 + z_2 + z_3 + \ln z_2 - \ln z_3, \qquad C(\hat{z}) = -\ln z_1 - \ln z_2 + \ln z_3,$$

where $\hat{z} = (z_1, z_2, z_3)^T$ and $C(\hat{z})$ is the Casimir function of system (1.83). The system of equations (1.83) is not volume preserving, since the right hand side vector field is not divergence free, and demonstrates that volume preservation is not a general property of the Poisson system (1.80). Nevertheless, the Lotka-Volterra model (1.83) with transformation:

$$\begin{aligned} z_1 &= e^x, \\ z_2 &= e^y, \\ z_3 &= e^{x+y+z}, \end{aligned}$$

which constitutes a global change of coordinates, can be transformed into canonical Hamiltonian form (1.54) on a reduced phase space $\mathbf{R}^2$:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= e^y + e^{x+y+z} = \nabla_y H(x,y,z), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= e^x - e^{x+y+z} + 1 = -\nabla_x H(x,y,z), \end{aligned}$$

with Hamiltonian function $H(x,y,z) = -e^x + e^y + e^{x+y+z} - x - z$ and $z \equiv const.$ The dynamics on $\mathbf{R}^2$ is area preserving.

### 1.1.6 Hamiltonian PDEs

We begin introducing Hamiltonian PDEs by defining the Poisson bracket. Consider the inner product space $\mathcal{U}^d$ of smooth functions defined on a simply connected bounded open set $\mathcal{D} \in \mathbf{R}^d$ with boundary $\partial\mathcal{D}$. Then for any two functionals $\mathcal{F}, \mathcal{G} : \mathcal{U}^d \to \mathbf{R}$ and a matrix differential operator $\mathcal{J}(u)$, that is skew-symmetric with respect to the inner product on $\mathcal{U}^d$, the Poisson bracket (1.81) generalizes to the integral

$$\{\mathcal{F}, \mathcal{G}\} = \int_{\mathcal{D}} \frac{\delta \mathcal{F}}{\delta u} \mathcal{J}(u) \frac{\delta \mathcal{G}}{\delta u} \, \mathrm{d}x, \tag{1.84}$$

where $u(x) \in \mathcal{U}^d$, $x \in \mathcal{D}$ and $\frac{\delta}{\delta u}$ denotes variational derivative defined by

$$\left( \frac{\delta \mathcal{F}}{\delta u}, v \right) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[ \mathcal{F}(u + \epsilon v) - \mathcal{F}(u) \right], \quad \forall\, v \in \mathcal{U}^d$$

with an inner product $(\cdot, \cdot) : \mathcal{U}^d \times \mathcal{U}^d \to \mathbf{R}$ on $\mathcal{U}^d$. Bracket (1.84) is a skew-symmetric, bilinear form acting on functionals on $\mathcal{U}^d$ and is said to be a Poisson bracket if it satisfies the Jacobi identity (1.78). If the matrix differential operator $\mathcal{J}$ does not explicitly depend on function $u$ then Jacobi identity is automatically satisfied. This follows from the skew-symmetry property of $\mathcal{J}$ and the standard rules of calculus. Since there is no well-defined multiplication between functionals, we have neglected the requirement for the Poisson bracket (1.84) to satisfy Leibniz's rule (1.77).

We call a partial differential equation a Hamiltonian PDE if it can be written in the following form:

$$\frac{\partial u}{\partial t} = \mathcal{J}(u) \frac{\delta \mathcal{H}}{\delta u}, \tag{1.85}$$

where $\mathcal{H} : \mathcal{U}^d \to \mathbf{R}$ is a Hamiltonian functional, and the associated Poisson bracket (1.84) satisfies all above mentioned properties of the bracket. Then any functional $\mathcal{F} : \mathcal{U}^d \to \mathbf{R}$ under the dynamics of the Hamiltonian PDE (1.85) obeys the integral equation

$$\frac{\partial \mathcal{F}}{\partial t} = \{\mathcal{F}, \mathcal{H}\}.$$

From the properties of the Poisson bracket (1.84) follow conservation of the Hamiltonian functional $\mathcal{H}$, i.e. $\{\mathcal{H}, \mathcal{H}\} = 0$, that any functional $\mathcal{I} : \mathcal{U}^d \to \mathbf{R}$ who satisfies $\{\mathcal{I}, \mathcal{H}\} = 0$ is a first integral of (1.85) and that the Poisson bracket of any two first integrals is a first integral itself. Functional $\mathcal{C} : \mathcal{U}^d \to \mathbf{R}$ is called a Casimir if the Poisson bracket $\{\mathcal{C}, \mathcal{F}\} = 0$ for any functional $\mathcal{F}$. Note the resemblance in geometric structure of the Hamiltonian PDE (1.85) to the geometric structure of the Poisson system (1.80).

Recall the semi-linear sine-Gordon equations (1.29)–(1.32) from Section 1.1.1. These equations belong to the class of Hamiltonians PDEs and can be written in the form (1.85) with a canonical skew-symmetric system matrix

$$\mathcal{J} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{1.86}$$

and Hamiltonian functional (1.33). With system matrix (1.86) the associated Poisson bracket for any two functionals $\mathcal{F}, \mathcal{G} : \mathcal{C}^\infty([0,1])^2 \to \mathbf{R}$ of pair $(u, v)$ becomes

$$\{\mathcal{F}, \mathcal{G}\} = -\int_0^1 \left( \frac{\delta \mathcal{F}}{\delta v}\frac{\delta \mathcal{G}}{\delta u} - \frac{\delta \mathcal{F}}{\delta u}\frac{\delta \mathcal{G}}{\delta v} \right) \mathrm{d}x.$$

The first variation of the Hamiltonian functional (1.33) with respect to u and v are

$$\delta \mathcal{H} = \int_0^1 \left( v\delta v + \frac{\partial u}{\partial x}\frac{\partial \delta u}{\partial x} + \sin(u)\delta u \right) \mathrm{d}x = \int_0^1 \left( v\delta v - \frac{\partial^2 u}{\partial x^2}\delta u + \sin(u)\delta u \right) \mathrm{d}x,$$

where the boundary conditions (1.32) have been used to carry out the integration by parts. Writing down the equations in Hamiltonian PDE form (1.85) we recover the sine-Gordon equations (1.29)–(1.30), i.e.

$$\frac{\partial u}{\partial t} = \frac{\delta \mathcal{H}}{\delta v} = v, \tag{1.87}$$

$$\frac{\partial v}{\partial t} = -\frac{\delta \mathcal{H}}{\delta u} = \frac{\partial^2 u}{\partial x^2} - \sin(u). \tag{1.88}$$

In Chapter 2 we are concerned with linearized, 2D, inviscid, incompressible Euler-Boussinesq partial differential equations in the stream function formulation:

$$\frac{\partial q}{\partial t} = -\frac{\partial b}{\partial x}, \tag{1.89}$$

$$\frac{\partial b}{\partial t} = -N_f^2 \frac{\partial \psi}{\partial x}, \tag{1.90}$$

$$q = -\Delta \psi, \tag{1.91}$$

$$\psi = 0 \quad \text{on} \quad \partial \mathcal{D}, \tag{1.92}$$

where $\psi$ is the stream function, $q$ is the vorticity, $b$ is the buoyancy and $N_f$ is the stratification frequency. Equations (1.89)–(1.92) are defined in space on a simply connected bounded open set $\mathcal{D} \subset \mathbf{R}^2$ with boundary $\partial \mathcal{D}$, in vertical plane coordinates $(x, z) \in \mathcal{D}$. The Hamiltonian structure of the Euler equations for an ideal fluid is well-known [87]. As shown in [45], the nonlinear Euler-Boussinesq equations inherit the noncanonical Hamiltonian structure from the ideal fluid Poisson bracket. In Chapter 2 we verify and show that the linearized equations (1.89)–(1.92) preserve a linear Hamiltonian structure with system matrix $\mathcal{J}$ and Hamiltonian $\mathcal{H}$:

$$\mathcal{J} = -N_f^2 \begin{bmatrix} 0 & \frac{\partial}{\partial x} \\ \frac{\partial}{\partial x} & 0 \end{bmatrix}, \qquad \mathcal{H} = \frac{1}{2}\int_{\mathcal{D}} \left( \nabla\psi \cdot \nabla\psi + \frac{1}{N_f^2} b^2 \right) \mathrm{d}x\, \mathrm{d}z,$$

respectively. Since the matrix $\mathcal{J}$ is not invertible, system (1.89)–(1.92) possesses the Casimir invariant:

$$\mathcal{C} = \int_{\mathcal{D}} \frac{1}{N_f^2} b \, \mathrm{d}x\, \mathrm{d}z.$$

Its time derivative is equal to zero due to the zero Dirichlet boundary conditions (1.92) of the stream function.

In Chapter 4 we consider one-dimensional Korteweg-de Vries (KdV) model on the $2\pi$-periodic domain:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0. \tag{1.93}$$

The classical KdV equation is obtained with proper rescaling of time and space. Note that KdV equation (1.93) without dispersion term $\frac{\partial^3 u}{\partial x^3}$ reduces to the Burgers-Hopf equation. The structure differential operator $\mathcal{J} = -\frac{\partial}{\partial x}$ implies the corresponding Poisson bracket:

$$\{\mathcal{F}, \mathcal{G}\} := -\int_0^{2\pi} \frac{\delta \mathcal{F}}{\delta u}\frac{\partial}{\partial x}\frac{\delta \mathcal{G}}{\delta u}\,\mathrm{d}x.$$

Together with the Hamiltonian functional

$$\mathcal{H} = \int_0^{2\pi} \left(\frac{1}{6}u^3 - \frac{1}{2}\left(\frac{\partial u}{\partial x}\right)^2\right)\,\mathrm{d}x, \tag{1.94}$$

the KdV equation (1.93) can be written in the Hamiltonian PDE form (1.85). The first variation of the Hamiltonian (1.93) with respect to function u is

$$\delta\mathcal{H} = \int_0^{2\pi} \left(\frac{1}{2}u^2\delta u - \frac{\partial u}{\partial x}\frac{\partial \delta u}{\partial x}\right)\,\mathrm{d}x = \int_0^{2\pi} \left(\frac{1}{2}u^2\delta u + \frac{\partial^2 u}{\partial x^2}\delta u\right)\,\mathrm{d}x,$$

where result follows from the integration by parts and periodic boundary conditions. Then

$$\frac{\partial u}{\partial t} = -\frac{\partial}{\partial x}\frac{\delta \mathcal{H}}{\delta u} = -\frac{\partial}{\partial x}\left(\frac{1}{2}u^2 + \frac{\partial^2 u}{\partial x^2}\right) = -u\frac{\partial u}{\partial x} - \frac{\partial^3 u}{\partial x^3} \tag{1.95}$$

and we recover KdV equation (1.93). The differential operator $\mathcal{J} = -\frac{\partial}{\partial x}$ is not invertible and this gives rise to the Casimir functional

$$\mathcal{C} = \int_0^{2\pi} u\,\mathrm{d}x. \tag{1.96}$$

It is easy to check that $\{\mathcal{C}, \mathcal{F}\} = 0$ is true for any functional $\mathcal{F}$. Hamiltonian $\mathcal{H}$ and Casimir $\mathcal{C}$ are not the only conserved quantities of the KdV equation, in fact, there are infinitely many conserved quantities. We refer readers to Chapter 4 for discussion on conserved quantities of the KdV and Burgers-Hopf equations. For further reference we mention that the functional

$$\mathcal{E} = \int_0^{2\pi} \frac{1}{2}u^2\,\mathrm{d}x \tag{1.97}$$

is also a conserved quantity of the KdV equation (1.93).

### 1.1.7 Semi-discretized Hamiltonian PDEs

In this subsection we address questions related to the structure preserving numerical methods for Hamiltonian PDEs. Our objective when discretizing Hamiltonian PDEs in space is to derive semi-discretized Hamiltonian systems, i.e. Hamiltonian

ODEs (1.42), such that the approximated Hamiltonian functional and as many Casimirs and other conserved quantities of the original system as possible are conserved quantities of the semi-discretized Hamiltonian system, and the associated discrete bracket again satisfies all properties of a Poisson bracket. Thus preserving geometric properties as much as possible.

To ensure that the semi-discretized equations are at least a Hamiltonian (or Poisson) system, we separately approximate the Poisson bracket, i.e. structure operator $\mathcal{J}(u)$ such that all properties of the Poisson bracket are satisfied, and the Hamiltonian functional $\mathcal{H}$, (see McLachlan [81]). Unfortunately, this method does not automatically ensure that discrete approximations of other first integrals will be preserved in discrete sense.

We illustrate the approach described above for the KdV equation (1.93) on a $2\pi$-periodic domain. For comparison we consider two methods: a finite difference method and a Fourier-Galerkin method, the motivation will become evident in the following.

Consider $N$ equally spaced grid points $x_i$ on the 1-dimensional torus

$$\mathbf{T}_{2\pi} = \{x \bmod 2\pi \,|\, x \in \mathbf{R}\}$$

and define the grid size $\Delta x = 2\pi/N$ such that $x_i = i\Delta x$ for $i = 0, \ldots, N-1$. The discrete values of function $u$ at each grid point are defined by $u_i = u(i\Delta x)$ for each $i = 0, \ldots, N-1$. With $\boldsymbol{u} \in \mathbf{R}^N$ we define a vector of discrete function $u$ values $u_i$ where $i = 0, \ldots, N-1$ and with $\mathbf{U} = \mathbf{R}^N$ we denote the space of all periodic (with respect to subscript $i$) grid functions $\boldsymbol{u}$, i.e. periodicity condition implies that $u_k = u_{k \bmod N}$ for any $k \in \mathbf{Z}$.

We also define the discrete inner product on $\mathbf{U}$:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathbf{U}} = \sum_{i=0}^{N-1} u_i v_i \, \Delta x = \boldsymbol{u}^T \boldsymbol{v} \, \Delta x, \quad \forall \, \boldsymbol{u}, \boldsymbol{v} \in \mathbf{U},$$

where the last equality follows from the fact that we consider a uniform grid. With matrices $D_x, D_{2x} : \mathbf{U} \to \mathbf{U}$ of dimension $N \times N$ we define finite difference approximation matrices for the spatial derivatives:

$$(D_x \boldsymbol{u})_i = \frac{u_i - u_{i-1}}{\Delta x}, \quad (D_{2x} \boldsymbol{u})_i = \frac{u_{i+1} - u_{i-1}}{2\Delta x}, \quad i = 0, \ldots, N-1,$$

where $\boldsymbol{u} \in \mathbf{U}$, matrix $D_{2x}$ is skew-symmetric and matrix $D_{xx} = -D_x^T D_x \in \mathbf{R}^{N \times N}$ is symmetric with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathbf{U}}$. Matrix $D_{xx} : \mathbf{U} \to \mathbf{U}$ defines a finite difference approximation matrix for the second order spatial derivative. Then, in terms of the inner product on $\mathbf{U}$, the discrete approximation of the Hamiltonian functional (1.94) is defined by

$$H(\boldsymbol{u}) = \frac{1}{6} \langle \boldsymbol{u}, \boldsymbol{u} * \boldsymbol{u} \rangle_{\mathbf{U}} - \frac{1}{2} \langle D_x \boldsymbol{u}, D_x \boldsymbol{u} \rangle_{\mathbf{U}}, \tag{1.98}$$

where $*$ denotes pointwise vector multiplication. The variational derivative of $H(\boldsymbol{u})$ is defined in the discrete inner product by

$$\left\langle \frac{\delta H}{\delta \boldsymbol{u}}, \boldsymbol{v} \right\rangle_{\mathbf{U}} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left( H(\boldsymbol{u} + \epsilon \boldsymbol{v}) - H(\boldsymbol{u}) \right), \quad \forall \, \boldsymbol{v} \in \mathbf{U}.$$

We find that

$$\frac{\delta H}{\delta \boldsymbol{u}} = \frac{1}{2}(\boldsymbol{u} * \boldsymbol{u}) + D_{xx}\boldsymbol{u}.$$

By approximating the structure operator $\mathcal{J} = -\frac{\partial}{\partial x}$ with matrix $-D_{2x}$, the Hamiltonian semi-discretization of the KdV equation in the form of a Hamiltonian PDE (1.95) can be defined by

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = -D_{2x}\frac{\delta H}{\delta \boldsymbol{u}} = -D_{2x}\left(\frac{1}{2}(\boldsymbol{u} * \boldsymbol{u}) + D_{xx}\boldsymbol{u}\right), \tag{1.99}$$

which is a symmetric finite difference approximation on the uniform grid of the KdV equation (1.93) with $2\pi$-periodic boundary conditions.

Interestingly, when $N$ is odd, the matrix $D_{2x}$ is singular with rank $N-1$ and the finite difference approximation (1.99) preserves the discrete approximation to the Casimir functional (1.96):

$$C(\boldsymbol{u}) = \langle \mathbf{1}, \boldsymbol{u} \rangle_{\mathbf{U}}, \tag{1.100}$$

where $\mathbf{1} \in \mathbf{R}^N$ is a vector of ones. When the dimension of matrix $D_{2x}$ is even, its rank is $N-2$ and the finite difference approximation (1.99) preserves two Casimir invariants:

$$C_1(\boldsymbol{u}) = \sum_{i=0}^{N/2-1} u_{2i}\,\Delta x, \qquad C_2(\boldsymbol{u}) = \sum_{i=0}^{N/2-1} u_{2i+1}\,\Delta x, \tag{1.101}$$

which is an artefact of the numerical discretization. Clearly, their sum implies (1.100), i.e. $C(\boldsymbol{u}) = C_1(\boldsymbol{u}) + C_2(\boldsymbol{u})$.

The finite difference approximation (1.99) is a noncanonical Hamiltonian system (1.42) with constant system matrix $J = -D_{2x}$ and Hamiltonian (1.98). The Hamiltonian and the Casimir function (1.100) or Casimirs (1.101) are the only known conserved quantities of the system (1.99). Conservation of other first integrals of the KdV equation (1.93) was lost during the process of going from continuous to discrete equations.

In Chapter 2 we consider a similar approach for constructing structure preserving semi-discetized equations for the Euler-Boussinesq equations (1.89)–(1.92).

Now we can explain why the discrete approximation (1.41) of the energy functional (1.33) by the quadrature rule is a conserved quantity along the solution of the semi-discretized sine-Gordon equations (1.34)–(1.36). Recall that sine-Gordon equations are a Hamiltonian PDE (1.87)–(1.88) with canonical system matrix (1.86). With notation from Section 1.1.1 we define a space $\mathbf{V} \in \mathbf{R}^{N-1}$ of discrete grid functions $\boldsymbol{u}$, where $\boldsymbol{u}_i = u_i$ for $i = 1, \ldots N-1$, with zero Dirichlet boundary conditions (with respect to subscript $i$), i.e. $u_0 = u_N = 0$. Then by defining the discrete inner product on $\mathbf{V}$, i.e.

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathbf{V}} = \sum_{i=1}^{N-1} u_i v_i\,\Delta x = \boldsymbol{u}^T \boldsymbol{v}\,\Delta x, \quad \forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathbf{V},$$

the Hamiltonian functional (1.33) is approximated by

$$H(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2}\langle \boldsymbol{v}, \boldsymbol{v} \rangle_{\mathbf{V}} + \frac{1}{2}\langle D_x\boldsymbol{u}, D_x\boldsymbol{u} \rangle_{\mathbf{V}} - \langle \mathbf{1}, \cos(\boldsymbol{u}) \rangle_{\mathbf{V}}.$$

This is exactly the same approximation as (1.41). Since matrix (1.86) is constant, we do not need to approximate it. Hence with the proper definition of the variational derivative of $H(\boldsymbol{u}, \boldsymbol{v})$ in the discrete inner product on $\mathbf{V}$ we find that

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \frac{\delta H}{\delta \boldsymbol{v}} = \boldsymbol{v}, \\ \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} &= -\frac{\delta H}{\delta \boldsymbol{u}} = -D_x^T D_x \boldsymbol{u} - \sin(\boldsymbol{u}) \end{aligned}$$

and recover equations (1.34)–(1.35). This shows that in Section 1.1.1 considered semi-discretized sine-Gordon equations (1.34)–(1.36) are in fact a canonical Hamiltonian system and time reversible, since $H(\boldsymbol{u}, \boldsymbol{v}) = H(\boldsymbol{u}, -\boldsymbol{v})$. Conservation of the discrete energy function (1.41) follows from the construction of the Hamiltonian structure preserving finite difference method.

For comparison to the finite difference approximation (1.99) we consider the Fourier-Galerkin method, finite spectral truncation of the KdV equation (1.93). We follow the approach from Chapter 4.

Let $\mathcal{P}_N$ denotes the standard $N$-mode Fourier projection operator, i.e.

$$u_N = \mathcal{P}_N u(x) = \sum_{|k| \le N} \hat{u}_k e^{ikx}, \qquad \hat{u}_k = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-ikx} \mathrm{d}x,$$

where $\hat{u}_k$ is the $k^{th}$ Fourier coefficient of the smooth real $2\pi$-periodic function $u(x)$. Since $u(x)$ is real, we have

$$\hat{u}_{-k} = \hat{u}_k^*.$$

By $\mathcal{U}$ we denote the function space of the $N$-mode Fourier projection functions $u_N$, equipped with the $L^2$ inner product. The projection operator $\mathcal{P}_N$ is symmetric with respect to the inner product $(\cdot, \cdot)$ and commutes with the derivative operator $\frac{\partial}{\partial x}$. Consequently, the composite operator $\frac{\partial}{\partial x}\mathcal{P}_N$ is skew-symmetric with respect to the $L^2$ inner product. Hence the Poisson bracket (1.84) may be defined by

$$\{\mathcal{F}_N, \mathcal{G}_N\} = -\int_0^{2\pi} \frac{\delta \mathcal{F}_N}{\delta u_N} \frac{\partial}{\partial x} \mathcal{P}_N \frac{\delta \mathcal{G}_N}{\delta u_N} \, \mathrm{d}x,$$

where $\mathcal{F}_N$ and $\mathcal{G}_N$ are any two functionals restricted to the truncated function $u_N$ and its derivatives, i.e.

$$\mathcal{F}_N = \int_0^{2\pi} F\left(u_N, \frac{\partial u_N}{\partial x}, \frac{\partial^2 u_N}{\partial x^2}, \dots\right) \mathrm{d}x.$$

The Hamiltonian (1.94) and Casimir (1.96) restricted to the truncated function $u_N$ become:

$$\mathcal{H}_N = \int_0^{2\pi} \left(\frac{1}{6} u_N^3 - \frac{1}{2}\left(\frac{\partial u_N}{\partial x}\right)^2\right) \mathrm{d}x, \qquad \mathcal{C}_N = \int_0^{2\pi} u_N \, \mathrm{d}x,$$

respectively. It is easy to see that $\{\mathcal{C}_N, \mathcal{F}_N\} = 0$ for any functional $\mathcal{F}_N$.

Hence the finite truncation of the KdV Hamiltonian PDE (1.95) reads:

$$\frac{\partial u_N}{\partial t} = -\frac{\partial}{\partial x}\mathcal{P}_N \frac{\delta H}{\delta u_N}, \tag{1.102}$$

where from the definition of the variational derivative with respect to the $L^2$ inner product $(\cdot,\cdot)$ follows that

$$\frac{\delta H}{\delta u_N} = \frac{1}{2}u_N^2 + \frac{\partial^2 u_N}{\partial x^2}.$$

Note that the finite truncation (1.102) is still space dependent, i.e. $x$ dependent. By multiplying equation (1.102) from both sides with test functions $e^{-imx}$ where $m = 1, \ldots, N$ and then integrating from 0 to $2\pi$, the Galerkin approach, equations can be written in terms of the Fourier coefficients $\hat{u}_k$, see Chapter 4.

Opposite to the finite difference discretization (1.99), spectral truncation (1.102) of the KdV equation gives rise to the additional conserved quantity:

$$\mathcal{E}_N = \int_0^{2\pi} \frac{1}{2}u_N^2 \,\mathrm{d}x, \tag{1.103}$$

which is the functional (1.97) restricted to the truncated function $u_N$. Since

$$\begin{aligned}
\{\mathcal{E}_N, \mathcal{H}_N\} &= -\frac{1}{2}\int_0^{2\pi} u_N \frac{\partial}{\partial x}\mathcal{P}_N\left(u_N^2\right) \,\mathrm{d}x - \int_0^{2\pi} u_N \frac{\partial^3 u_N}{\partial x^3}\,\mathrm{d}x \\
&= \frac{1}{2}\int_0^{2\pi} u_N^2 \frac{\partial u_N}{\partial x}\,\mathrm{d}x + \int_0^{2\pi} \frac{\partial u_N}{\partial x}\frac{\partial^2 u_N}{\partial x^2}\,\mathrm{d}x \\
&= \frac{1}{6}\int_0^{2\pi} \frac{\partial u_N^3}{\partial x}\,\mathrm{d}x + \frac{1}{2}\int_0^{2\pi} \frac{\partial}{\partial x}\left(\frac{\partial u_N}{\partial x}\right)^2 \,\mathrm{d}x = 0,
\end{aligned}$$

where the result follows from the periodic boundary conditions, function $\mathcal{E}_N$ is a first integral of the truncated KdV equation (1.102). We take advantage of this in Chapter 4.

In passing we mention that equation system (1.102) can be efficiently solved using a standard pseudospectral approach where derivatives are computed in real space with discrete fast Fourier transform. This implies discrete representation of the truncated function $u_N$ at grid values $x_i$. Hence with the same number of grid points, solutions of the both numerical methods for the KdV equation can be compared in real space at these grid values. Alternatively, we can apply the discrete Fourier transform to the solution $\boldsymbol{u}$ of the finite difference method (1.99) and compare both methods in spectral representation.

While both numerical approximations of the KdV equation preserve the main structure of the Hamiltonian PDE, they have different number of conserved quantities which strongly constrain the dynamics and may affect statistical properties [20, 21]. The best way to illustrate this is by expressing functional (1.103) in spectral representation, i.e.

$$\mathcal{E}_N = 2\pi \sum_{|n|\leq N} \frac{1}{2}\hat{u}_n \hat{u}_n^* = \pi \hat{u}_0^2 + 2\pi \sum_{n=1}^{N} |\hat{u}_n|^2 = 2\pi \sum_{n=1}^{N} |\hat{u}_n|^2,$$

where we can assume $\hat{u}_0 = 0$ up to a Galilean change of coordinates. This implies that the dynamics of the Fourier coefficients are constrained to the $(2N-1)$-dimensional hypersphere, i.e. the phase space is compact. The finite difference method does not share this property of the spectral representation and the associated phase space is not compact, but $\mathbf{R}^{2N}$ with the initial conditions satisfying $\boldsymbol{C}(\boldsymbol{u}) \equiv 0$.

### 1.1.8 Geometric integrators

In this subsection we discuss time integration of Hamiltonian dynamics (1.42). We refer readers to [37] for time integration methods of Poisson systems (1.80). When constructing time integrators for Hamiltonian systems we are concerned with preservation of symplecticity, first integrals and time reversibility. Volume preservation follows from symplecticity.

Consider the canonical Hamiltonian system (1.55)–(1.56) with initial conditions $(q_0, p_0)$ and its approximation:

$$(q^{n+1}, p^{n+1}) = \Psi_H^\tau(q^n, p^n), \tag{1.104}$$

where the discrete flow map $\Psi_H^\tau$ maps the discrete solution $(q^n, p^n) \approx (q(t^n), p(t^n))$ at time $t^n$ to the discrete solution at time $t^{n+1}$ with time step $\tau = t^{n+1} - t^n$ for $n = 0, 1, \ldots$ At time $t = 0$ we have that $(q^0, p^0) \equiv (q_0, p_0)$ and we take $\tau = const$ such that $t^n = n\tau$. Equation (1.104) defines a one-step numerical method and we say that it is of order $m$ if the local error at each time $t^n$ between the exact and approximate solutions is $O(\tau^{m+1})$, i.e.

$$\Phi_H^\tau(q^n, p^n) - \Psi_H^\tau(q^n, p^n) = O(\tau^{m+1}).$$

We call the one-step method (1.104) symplectic if

$$\mathrm{d}q^{n+1} \wedge \mathrm{d}p^{n+1} = \mathrm{d}q^n \wedge \mathrm{d}p^n. \tag{1.105}$$

Recall the second order Störmer-Verlet method (StV) (1.22)–(1.24) from Section 1.1.1. We show that StV method applied to the canonical Hamiltonian system with separable Hamiltonian, i.e. $H(q,p) = H_1(p) + H_2(q)$, is symplectic. Equations read:

$$p^* = p^n - \frac{\tau}{2} \nabla H_2(q^n), \tag{1.106}$$

$$q^{n+1} = q^n + \tau \nabla H_1(p^*), \tag{1.107}$$

$$p^{n+1} = p^* - \frac{\tau}{2} \nabla H_2(q^{n+1}). \tag{1.108}$$

By differentiating equations (1.106)–(1.108) we arrive at the following system:

$$\mathrm{d}p^* = \mathrm{d}p^n + A\mathrm{d}q^n, \tag{1.109}$$

$$\mathrm{d}q^{n+1} = \mathrm{d}q^n + B\mathrm{d}p^*, \tag{1.110}$$

$$\mathrm{d}p^{n+1} = \mathrm{d}p^* + C\mathrm{d}q^{n+1}, \tag{1.111}$$

where $A$, $B$ and $C$ are symmetric Hessian matrices, i.e.

$$A = -\frac{\tau}{2}\nabla_{qq}H_2(q^n), \quad B = \tau\nabla_{pp}H_1(p^*), \quad C = -\frac{\tau}{2}\nabla_{qq}H_2(q^{n+1}).$$

Taking the wedge product of equation (1.111) with $\mathrm{d}q^{n+1}$ from the left and using the property of the wedge product that $\mathrm{d}X \wedge A\mathrm{d}X = 0$ for any symmetric matrix $A$ we find:

$$\mathrm{d}q^{n+1} \wedge \mathrm{d}p^{n+1} = \mathrm{d}q^{n+1} \wedge \mathrm{d}p^* = (\diamond). \tag{1.112}$$

Substituting $\mathrm{d}p^*$ from (1.109) in (1.112) and using $\mathrm{d}q^{n+1}$ from (1.110) we get:

$$\begin{aligned}(\diamond) &= \mathrm{d}q^{n+1} \wedge \mathrm{d}p^n + \mathrm{d}q^{n+1} \wedge A\mathrm{d}q^n \\ &= \mathrm{d}q^n \wedge \mathrm{d}p^n + B\mathrm{d}p^* \wedge \mathrm{d}p^n + \mathrm{d}q^n \wedge A\mathrm{d}q^n + B\mathrm{d}p^* \wedge A\mathrm{d}q^n \\ &= \mathrm{d}q^n \wedge \mathrm{d}p^n + B\mathrm{d}p^* \wedge (\mathrm{d}p^n + A\mathrm{d}q^n) \\ &= \mathrm{d}q^n \wedge \mathrm{d}p^n + B\mathrm{d}p^* \wedge p^* = \mathrm{d}q^n \wedge \mathrm{d}p^n.\end{aligned}$$

This completes the proof of the statement that the Störmer-Verlet method (1.22)–(1.24) applied to canonical separable Hamiltonian systems is symplectic. For example, in Chapter 2 we apply StV method (1.106)–(1.108) to Hamiltonian system of semi-discretized Euler-Boussinesq equations (1.89)–(1.92).

In fact, StV method can be extended to general Hamiltonian systems (1.55)–(1.56) and symplecticity of the method can be shown. The method reads:

$$q^{n+1/2} = q^n + \frac{\tau}{2}\nabla_p H(q^n, p^{n+1/2}), \tag{1.113}$$

$$p^{n+1/2} = p^n - \frac{\tau}{2}\nabla_q H(q^n, p^{n+1/2}), \tag{1.114}$$

$$q^{n+1} = q^{n+1/2} + \frac{\tau}{2}\nabla_p H(q^{n+1}, p^{n+1/2}), \tag{1.115}$$

$$p^{n+1} = p^n - \frac{\tau}{2}\nabla_q H(q^{n+1}, p^{n+1/2}). \tag{1.116}$$

Note that the method (1.113)–(1.116) for separable Hamiltonian functions reduces to (1.106)–(1.108) with $p^* = p^{n+1/2}$.

Symplecticity conservation (1.105) implies volume preservation in phase space by the numerical method. In Section 1.1.1 all considered example equations were canonical Hamiltonian systems. Hence all the numerical solutions of the StV method were symplectic and volume preserving. This confirms the statement in Section 1.1.1 that the area in phase space enclosed by the numerical solutions of the transformed Lotka-Volterra model (1.25)–(1.26) is preserved in time, see Figures 1.2(a) and 1.2(b).

The Störmer-Verlet method is not the only existing symplectic method for Hamiltonian systems. The method belongs to the general class of symplectic partitioned Runge-Kutta methods. On the another hand, the StV method (1.113)–(1.116) is also a composition method of two symplectic partitioned Runge-Kutta methods. The fact that the composition of symplectic methods is symplectic proves the statement that method (1.113)–(1.116) is symplectic. The conditions for general Runge-Kutta methods to be symplectic can be found in [37]. Additionally, we mention here the

implicit midpoint method (second order symplectic Runge-Kutta method), which preserves quadratic first integrals. We demonstrate this for the general Hamiltonian system (1.42). The method reads:

$$X^{n+1} = X^n + \tau J \nabla H \left( X^{n+1/2} \right), \qquad X^{n+1/2} = \frac{X^{n+1} + X^n}{2}. \tag{1.117}$$

The quadratic first integral of the system is defined by

$$I(X) = \frac{1}{2} X^T A X,$$

where $A \in \mathbf{R}^{n \times n}$ is a symmetric matrix. From (1.43) follows that

$$(AX)^T J \nabla H(X) = 0.$$

We multiply the first equation in (1.117) by $\left( AX^{n+1/2} \right)^T$ from the left. We get that

$$\begin{aligned} \left( AX^{n+1/2} \right)^T X^{n+1} &= \left( AX^{n+1/2} \right)^T X^n + \tau \left( AX^{n+1/2} \right)^T J \nabla H \left( X^{n+1/2} \right) \\ &= \left( AX^{n+1/2} \right)^T X^n. \end{aligned}$$

Expanding both sides yields:

$$\begin{aligned} \frac{1}{2} X^{n+1^T} A X^{n+1} + \frac{1}{2} X^{n^T} A X^{n+1} &= \frac{1}{2} X^{n^T} A X^n + \frac{1}{2} X^{n+1^T} A X^n, \\ \frac{1}{2} X^{n+1^T} A X^{n+1} &= \frac{1}{2} X^{n^T} A X^n, \end{aligned}$$

and proves the statement. In Chapter 4 we apply the implicit midpoint method to the semi-discretized KdV and Burgers-Hopf equations (1.102) to exactly preserve the quadratic invariant (1.103) in time. Conservation of linear invariants are shared by all Runge-Kutta methods and a large class of partitioned Runge-Kutta methods, including the StV method. Hence all these methods preserve linear Casimirs (1.51) of Hamiltonian dynamics (1.42).

Consider the discrete flow map $\Psi_H^\tau$ of the Hamiltonian system (1.42). The exact flow map $\Phi_H^t$ satisfies $\left[ \Phi_H^t \right]^{-1} = \Phi_H^{-t}$ and is time reversible with respect to the linear transformation (1.48)–(1.49) if

$$S \Phi_H^t(X) = \left[ \Phi_H^t \right]^{-1} (SX).$$

If the discrete flow map $\Psi_H^\tau$ is invertible and satisfies

$$S \Psi_H^\tau(X^n) = \left[ \Psi_H^\tau \right]^{-1} (SX^n),$$

then we call a flow map $\Psi_H^\tau$ and the method $X^{n+1} = \Psi_H^\tau(X^n)$ time reversible with respect to the transformation:

$$\begin{aligned} \tau &\to -\tau, \\ X^n &\to SX^n. \end{aligned}$$

Additionally, we call a flow map $\Psi_H^\tau$ symmetric if $[\Psi_H^\tau]^{-1} = \Psi_H^{-\tau}$. Hence the method $X^{n+1} = \Psi_H^\tau(X^n)$ is symmetric if exchanging $X^n \leftrightarrow X^{n+1}$ and $\tau \leftrightarrow -\tau$ leaves the method unaltered. For symmetric methods the time reversibility condition reduces to

$$S\Psi_H^\tau(X^n) = \Psi_H^{-\tau}(SX^n).$$

It is easy to see that the Störmer-Verlet method (1.113)–(1.116) and the implicit midpoint method (1.117) are both symmetric methods and time reversible if the continuous system is time reversible.

In Section 1.1.1 we saw that the StV method applied to the semi-discretized sine-Gordon equations (1.34)–(1.36) does not conserve the Hamiltonian function (1.41) exactly in time but up to the second order for long computational times. This is a result of a theorem for symplectic methods applied to canonical Hamiltonian systems. The theorem rests on regularity assumptions for the Hamiltonian function and flow maps, as well as on assumptions about the phase space, solutions of the numerical method and *backward error analysis*. In backward error analysis we are concerned with the derivation of the *modified differential equations* for which a numerical method with fixed time step $\tau$ is "exact", i.e.

$$\frac{\mathrm{d}X}{\mathrm{d}t} = J_{Id}\,\nabla H(X) + \tau f_1(X) + \tau^2 f_2(X) + \ldots, \qquad X(0) = X_0, \tag{1.118}$$

where $X \in \mathbf{R}^{2n}$. The correction terms to the Hamiltonian system in (1.118) form an asymptotic expansion with respect to time step $\tau$. Note that in general the expansion does not converge. We saw an exceptional case in Section 1.1.1 where we derived the modified Hamiltonian system (1.17)–(1.18) for the StV method applied to the linear Harmonic oscillator equations (1.1)–(1.2). There the result followed directly from the dispersion relation.

For a symplectic method $X^{n+1} = \Psi_H^\tau(X^n)$ applied to the Hamiltonian system with smooth Hamiltonian on a simply connected phase space $\Omega$, it can be shown that there exist smooth functions $H_i : \Omega \to \mathbf{R}$ for $i = 1, 2, \ldots$, such that $f_i(X) = J_{Id}\,\nabla H_i(X)$. This implies that after truncation the modified equation (1.118) is a Hamiltonian system itself with modified Hamiltonian

$$H_{[i]}(X) = H(x) + \tau H_1(X) + \tau^2 H_2(X) + \cdots + \tau^i H_i(X)$$

and flow map $\Phi^t_{H_{[i]}}$. In fact, if the method is of order $m > 1$ then $f_i(X) = 0$ for $1 \le i \le m-1$.

With an analytic Hamiltonian $H(X)$ and if the flow maps $\Psi_H^\tau$ and $\Phi^t_{H_{[i]}}$ are analytic and bounded on an open (complex) neighborhood of a compact subset $\mathcal{K} \subset \Omega$ of phase space, i.e. flow maps have convergent Taylor expansions in open set and their derivatives can be estimated, then estimates can be derived for all $X_0 \in \mathcal{K}$:

$$\left\| \Psi_H^\tau(X_0) - \Phi^\tau_{H_{[i]}}(X_0) \right\| \le c_1 \tau (c_2 (i+1)\tau)^{i+1},$$

where $c_1, c_2 > 0$ are independent of $i$ and $\tau$. The expression on the right hand side as a function of $i > 0$ for fixed value of $\tau$ shows that asymptotic expansion in (1.118) converges before it starts to diverge. By taking $i = i_*$ where $i_*$ is equal to the integer

part of $1/(\tau c_2 e) - 1$ and $\gamma = 1/(c_2 e)$, we can make the estimate exponentially small with respect to the time step $\tau$, i.e.

$$\left\| \Psi_H^\tau(X_0) - \Phi_{H_{[i_*]}}^\tau(X_0) \right\| \leq 3c_1 \tau e^{-\gamma\tau^{-1}}. \tag{1.119}$$

From the same regularity assumptions follow the existence of a global $\tau$-independent Lipschitz constant $\lambda > 0$ for the modified Hamiltonian $H_{[i_*]}$, and if $X^{n_*} \in \mathcal{K}$ for all $n_* = 1, \ldots, n$ where $X^{n_*} = \Psi_H^\tau(X^{n_*-1})$ then together with estimate (1.119) we can estimate its error over exponentially long time intervals $t^n = n\tau \leq e^{\frac{1}{2}\gamma\tau^{-1}}$:

$$\left\| H_{[i_*]}(X^n) - H_{[i_*]}(X_0) \right\| \leq 3\lambda n\tau c_1 e^{-\gamma\tau^{-1}} \leq 3\lambda c_1 e^{-\frac{1}{2}\gamma\tau^{-1}}.$$

For a method of order $m$ the modified Hamiltonian with $i = i_*$ is

$$H_{[i_*]}(X) = H(X) + \tau^m H_m(X) + \tau^{m+1} H_{m+1}(X) + \cdots + \tau^{i_*} H_{i_*}(X).$$

Since $H_m(X) + \tau H_{m+1}(X) + \cdots + \tau^{i_*-m} H_{i_*}(X)$ is uniformly bounded on $\mathcal{K}$ independently of $\tau$ and $i_*$, we find that

$$\begin{aligned} H(X^n) - H(X_0) + O(\tau^m) &= H_{[i_*]}(X^n) - H_{[i_*]}(X_0), \\ H(X^n) - H(X_0) + O(\tau^m) &= O\left(e^{-\frac{1}{2}\gamma\tau^{-1}}\right), \\ H(X^n) - H(X_0) &= O(\tau^m) \end{aligned}$$

over exponentially long time intervals $t^n \leq e^{\frac{1}{2}\gamma\tau^{-1}}$. This completes the discussion of the necessary ingredients for the proof of the following theorem. The numerical results supporting the theorem can be seen in Figures 1.3(a) and 1.3(b).

**Theorem 1.1.8.1.** *Consider a symplectic method $X^{n+1} = \Psi_H^\tau(X^n)$ with time step $\tau$ applied to the Hamiltonian system with analytic Hamiltonian function $H(X) : \Omega \to \mathbf{R}$ where $\Omega \subset \mathbf{R}^{2n}$ is simply connected open set. If the numerical solution stays in the compact set $\mathcal{K} \subset \Omega$ for $X_0 \in \mathcal{K}$, then there exists $\gamma > 0$ and $i_*$ such that*

$$\begin{aligned} H_{[i_*]}(X^n) - H_{[i_*]}(X_0) &= O\left(e^{-\frac{1}{2}\gamma\tau^{-1}}\right), \\ H(X^n) - H(X_0) &= O(\tau^m) \end{aligned}$$

*over exponentially long time intervals $n\tau \leq e^{\frac{1}{2}\gamma\tau^{-1}}$.*

In passing we describe the second order symplectic, symmetric and time reversible method used in Chapter 3 for Hamiltonian systems with holonomic constraints (1.69)–(1.71), the RATTLE algorithm:

$$\begin{aligned} q^{n+1} &= q^n + \tau M^{-1} p^{n+1/2}, \\ p^{n+1/2} &= p^n - \frac{\tau}{2} \nabla V(q^n) - \frac{\tau}{2} \nabla g(q^n)^T \lambda_1, \\ 0 &= g(q^{n+1}), \\ p^{n+1} &= p^{n+1/2} - \frac{\tau}{2} \nabla V(q^{n+1}) - \frac{\tau}{2} \nabla g(q^{n+1})^T \lambda_2, \\ 0 &= g(q^{n+1}) M^{-1} p^{n+1}. \end{aligned}$$

The RATTLE algorithm ensures that if $(q^n, p^n) \in \mathcal{TM}$ then also $(q^{n+1}, p^{n+1}) \in \mathcal{TM}$, where the tangent bundle $\mathcal{TM}$ is an associated phase space of (1.69)–(1.71). The Lagrange multiplier $\lambda_1$ enforces the constraint $0 = g(q^{n+1})$ such that $q^{n+1} \in \mathcal{M}$ and the Lagrange multiplier $\lambda_2$ enforces that $p^{n+1}$ belongs to the tangent plane of the constraint manifold $\mathcal{M}$ at position $q^{n+1}$.

This concludes the introductory section to geometric numerical integration.

## 1.2 Thermostated dynamics

Thermostats, applied to Hamiltonian systems, are in principle artificial modelling devices. Motivation for their use depends on the problem at hand: for example in constant temperature molecular dynamics enforcing the system to be in thermal equilibrium with their surroundings or modelling small scales in geophysical fluid dynamics [22]. Due to the limitations of computer power we are motivated to think about model reduction techniques with applications to molecular simulations, climate modelling and turbulence. We find that thermostat methods may serve this purpose well, especially, when we are concerned with statistical properties of the underlying model, such as an invariant probability density function and autocorrelation functions.

There are situations when statistical results on a dynamical system are the only meaningful statements we can make about the behaviour of solutions of the system, e.g. when dealing with chaotic (unpredictable, sensitive to initial conditions) dynamical systems or when considering microscopic scale modelling subject to stochastic (Brownian) motion. In these situations the single trajectory of the solution is meaningless and we have to rely on the statistical properties of the dynamical system.

In our terminology thermostat methods refer to stochastic-dynamical thermostats, since we stochastically perturb model equations (Hamiltonian dynamical systems), such that the resulting SDEs sample particular probability distribution, measure of the phase space. We have two objectives: sampling of the probability density function and computation of the autocorrelation functions as a measure of dynamical properties. To achieve these objectives we are concerned with the construction and use of gentle and efficient thermostat methods. Gentleness stands for the small errors in the autocorrelation functions and efficiency stands for optimal compromise between sampling rates and gentleness, which are in principle contradictory. Ultimately, we can only judge the efficiency of the thermostat method by comparing it to other thermostat methods. The concept of an efficient thermostat method is explored in Chapters 3 and 4.

In the following subsections we give a brief introduction to the concept of ergodicity and discuss its importance to statistical mechanics, thermostated dynamics and computational methods. In Sections 1.2.1–1.2.3 we describe the microcanonical and canonical ensembles considered in Chapters 3 and 4. Introduction to thermostats is presented in Section 1.2.4 with discussion of theoretical aspects of the methods. In Section 1.2.5 we describe the time integration algorithms for thermostated dynamics.

Most of the material presented in this section can be found in the following references: Bühler [5], Leimkuhler [59], Leimkuhler & Reich [63], Pavliotis & Stuart

[94], Khinchin [51], Penrose [95], Kloeden & Platen [53].

### 1.2.1 The ergodic hypothesis

In statistical mechanics the *ergodic hypothesis* refers to the attempt for providing a dynamical basis for statistical mechanics. It states that the time average value of an observable function of the dynamics is equivalent to an *ensemble average*, i.e. an average over the large number of different dynamical states (*microstates*) with identical thermodynamic properties. Indeed, this appears to be the case when the dynamical system is *ergodic*. Throughout the thesis we address ergodicity as a question of the existence and uniqueness of a unique invariant measure under the dynamics, for both ODEs and SDEs. While the presentations are different, the goal remains the same. In this subsection we discuss the mathematical theory of ergodicity for autonomous dynamical systems.

In statistical mechanics thermodynamic properties such as energy or temperature refer to the thermodynamic state (*macrostate*) of the system. In general, for a given phase space $\Omega$ and a probability measure $\mu$, one has defined an ensemble. The *ensemble* represents the configurations (microstates) of the system and the probability of realizing each configuration. In this thesis we are concerned with three ensembles: the microcanonical and canonical ensembles, as well as the mixed canonical ensemble.

The *microcanonical ensemble* describes a completely isolated system with a fixed number of particles, with a fixed volume and a fixed energy, and respects all the other conservation laws of the system, if such exist. In an isolated system, when at equilibrium, each of its accessible microstates is equally probable, the fundamental postulate in statistical mechanics.

The *canonical ensemble* describes a system with a fixed number of particles and a fixed volume in thermal equilibrium with its surrounding (energy reservoir, heat bath). The system may exchange energy with the energy reservoir only in the form of heat. The *mixed canonical ensemble* is the canonical ensemble extended to systems with additional conservation laws and constraints.

Correcting the measure, sampling the canonical ensemble and recreating scenario of the energy exchange with a reservoir is one of the main reasons for introducing thermostat methods. We discuss in detail the microcanonical and canonical statistical mechanics in Sections 1.2.2 and 1.2.3, respectively. The latter also contains a motivation for using thermostat methods.

Recall the objective of this subsection is to discuss the mathematics of the ergodic hypothesis and lay down the background for microcanonical statistical mechanics, which is the subject of the next subsection. To make this discussion more precise we recall some basic concepts and results from ergodic theory. Let a triple $(\Omega, \mathcal{A}, \mu)$ denote a *probability space* where $\mathcal{A}$ is a $\sigma$-algebra of $\Omega$ and $\mu : \mathcal{A} \to [0, 1]$ is a *probability measure*. The collection of subsets of a set $\Omega$ is called $\sigma$-algebra if it contains $\Omega$ and is closed under the operations of taking complements and countable unions of its elements. A function $\mu$ is a probability measure if $\mu(\Omega) = 1$, $\mu(\emptyset) = 0$ and if for a countable collection of pairwise disjoint sets $A_n \in \mathcal{A}$, i.e. $A_n \cap A_m = \emptyset$

for $n \neq m$, it holds that

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

We write $L^1(\Omega, \mathcal{A}, \mu)$ for the space of all functions $F : \Omega \to \mathbf{R}$ that are integrable with respect to the measure $\mu$, i.e. $F$ is $\mu$-measurable and $\int_\Omega |F| \, \mathrm{d}\mu < \infty$. We say that function $F$ is $\mu$-measurable if $F^{-1}(D) \in \mathcal{A}$ for every Borel subset $D \in \mathcal{B}(\mathbf{R})$. The Borel $\sigma$-algebra $\mathcal{B}(\mathbf{R})$ is a smallest $\sigma$-algebra containing all the open subsets of $\mathbf{R}$. The *space average* (ensemble average) of function $F$ is defined by

$$\langle F \rangle = \int_\Omega F \, \mathrm{d}\mu. \tag{1.120}$$

Let $\Phi : \Omega \to \Omega$ be a $\mu$-measurable transformation, i.e. $\Phi(A)^{-1} \in \mathcal{A}$ for all $A \in \mathcal{A}$. With a set $\{\Phi^n(X_0)\}_{n=0}^{\infty}$ we denote the *orbit* of $X_0 \in \Omega$. Then we can define the *discrete time average* of $F \in L^1(\Omega, \mathcal{A}, \mu)$ along the orbit of $X_0$ by

$$\bar{F}_D(X_0) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} F(\Phi^k(X_0)).$$

We say that $\Phi$ is *measure preserving* ($\mu$ is an invariant measure for $\Phi$) if

$$\mu(\Phi(A)) = \mu(A), \quad \forall\, A \in \mathcal{A}.$$

For measure preserving transformations $\Phi$ the following theorem is a basic result about the distribution of orbits, the *Poincaré recurrence theorem*:

**Theorem 1.2.1.1.** *Let $\Phi : \Omega \to \Omega$ be a measure preserving transformation of $(\Omega, \mathcal{A}, \mu)$ and let $A \in \mathcal{A}$ have $\mu(A) > 0$. Then for $\mu$-a.e. $X_0 \in A$, the orbit $\{\Phi^n(X_0)\}_{n=0}^{\infty}$ returns to $A$ infinitely often.*

A set $A \in \mathcal{A}$ is an *invariant* of transformation $\Phi$ if

$$\Phi(A) = A.$$

A transformation $\Phi$ is called *ergodic* if every invariant set $A \in \mathcal{A}$ of $\Phi$ is such that either $\mu(A) = 0$ or $\mu(A) = 1$. If there is a *unique* $\Phi$-invariant probability measure then we say that $\Phi$ is *uniquely ergodic*. The basic result in ergodic theory is the pointwise *Birkhoff's ergodic theorem*:

**Theorem 1.2.1.2.** *Let $\Phi : \Omega \to \Omega$ be a measure preserving transformation of $(\Omega, \mathcal{A}, \mu)$. Then for any $F \in L^1(\Omega, \mathcal{A}, \mu)$ the limit $\bar{F}_D := \bar{F}_D(X_0)$ exists for $\mu$-a.e. $X_0 \in \Omega$. The limit $\bar{F}_D \in L^1(\Omega, \mathcal{A}, \mu)$ is $\Phi$ invariant, i.e. $\bar{F}_D(\Phi(X_0)) = \bar{F}_D(X_0)$, and the space integrals are equal, $\langle F \rangle = \langle \bar{F}_D \rangle$. If the transformation $\Phi$ is ergodic, then $\bar{F}_D$ is constant and $\bar{F}_D = \langle F \rangle$ for $\mu$-a.e. $X_0 \in \Omega$.*

As a direct consequence of the ergodic Theorem 1.2.1.2 we can define the relative measure of any subset $A \in \mathcal{A}$ :

$$\mu_r(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} 1_A(\Phi^k(X_0)),$$

where $1_A$ is an indicator function of $A$, i.e.

$$1_A(X) = \begin{cases} 1, & X \in A, \\ 0, & X \notin A. \end{cases}$$

The relative measure $\mu_r$ measures the proportion of time that orbits of transformation $\Phi$ spend in a given subset of $\mathcal{A}$.

Let us illustrate how the theoretical considerations above translate to an autonomous dynamical system with phase space $\Omega \subset \mathbf{R}^n$:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = f(X), \qquad X(0) = X_0, \tag{1.121}$$

where $X \in \mathbf{R}^n$ and function $f(X) : \mathbf{R}^n \to \mathbf{R}^n$ is divergence-free, i.e. $\nabla \cdot f(X) = 0$. With $\mathcal{A}$ we will denote the $\sigma$-algebra of all open subsets of phase space $\Omega$. We will restrict our discussion to *finite* phase spaces $\Omega$, i.e. $\int_\Omega \mathrm{d}X < \infty$, such that all open subsets of $\Omega$ are finite. The flow map of system (1.121) is indicated by $\Phi^t$ where $t \in \mathbf{R}$. In Section 1.1.2 we noted that divergence free property of the right hand side vector field of the autonomous dynamical system plays an important role in statistical mechanics. Why it is important will become clear in the following discussions.

From the divergence free property of $f(X)$ follows that $\Phi^t$ is volume preserving, see Section 1.1.2. Hence it guaranties the conservation of the phase space volume element $\mathrm{d}X$ and induces measure on $\mathcal{A}$, i.e.

$$\mathrm{vol}\{A\} = \int_A \mathrm{d}X, \quad \forall\, A \in \mathcal{A}.$$

Since flow map $\Phi^t$ is volume preserving, it is measure preserving, i.e.

$$\mathrm{vol}\{\Phi^t(A)\} = \mathrm{vol}\{A\}, \quad \forall\, A \in \mathcal{A},\, t \in \mathbf{R}.$$

With the restriction to finite phase spaces $\Omega$, the measure is normalizable and we can apply the Poincaré recurrence Theorem 1.2.1.1. This implies that for almost every initial conditions $X_0 \in A$, where $A \in \mathcal{A}$ and $\mathrm{vol}\{A\} > 0$, the trajectory (orbit) defined by infinitely iterated map $\Phi^n$ where $n \in \mathbf{Z}$ will return to subset $A$ infinitely often. Note that the recurrence theorem holds only if the phase space $\Omega$ is finite, since the proof of Theorem 1.2.1.1 strictly relies on the fact that the measure is finite.

With $L^1(\Omega)$ we identify the space of integrable functions $F(X) : \Omega \to \mathbf{R}$ on $(\Omega, \mathcal{A})$ with respect to the Lebesgue measure $\mathrm{d}X$, that is, $F(X)$ is Lebesgue measurable and $\int_\Omega |F(X)|\, \mathrm{d}X < \infty$. The *continuous time average* of integrable function $F(X)$ is defined by

$$\bar{F} = \lim_{T\to\infty} \frac{1}{T} \int_0^T F(X(t))\, \mathrm{d}t, \qquad X(t) = \Phi^t(X_0). \tag{1.122}$$

Indeed, generalization of Birkhoff's ergodic theorem implies that the time average $\bar{F}$ exists for almost all initial conditions $X_0 \in \Omega$, but time averages may still be dependent on the initial condition.

To achieve the desirable result we consider invariant subsets of phase space $\Omega$. We call a subset $A \in \mathcal{A}$ *invariant* if all trajectories which start in $A$ never leave it, i.e. if $X_0 \in A$ then $\Phi^t(X_0) \in A$ for all times $t$. If subset $A$ is not *metrically decomposable*, i.e. set $A$ cannot be split into two disjoint invariant subsets $A_1$ and $A_2$ with nonzero volume, and if $X_0 \in A$, then system (1.121) is ergodic on invariant subset $A$ and generalization of Birkhoff's ergodic theorem implies that for almost all initial conditions $X_0 \in A$ time average (1.122) of function $F(X) \in L^1(\Omega)$ is equal to the space average (1.120) with respect to Lebesgue measure, i.e.

$$\bar{F} = \frac{1}{\mathrm{vol}\{A\}} \int_A F(X)\,\mathrm{d}X.$$

If the invariant subset $A$ is metrically decomposable, i.e. $A$ can be split into two invariant subsets $A_1$ and $A_2$ with nonzero volume, then the trajectory emanating from any point $X_0$ in $A_1$ will always stay in $A_1$ and will never reach subset $A_2$, and vice versa. Hence system (1.121) would be not ergodic on the whole subset $A$.

The general observation of the discussion above is that for the system (1.121) to be ergodic on the whole phase space $\Omega$, this phase space must be finite and metrically indecomposable. If the phase space $\Omega$ has metrically indecomposable invariant subsets of the flow map $\Phi^t$ then the system (1.121) can be ergodic only on these subsets, provided that the initial condition $X_0$ is in one of those sets. Unfortunately, there is no general method for finding such metrically indecomposable invariant subsets in practical problems.

The main essence of ergodic systems is that trajectories visit the whole of the invariant and metrically indecomposable phase space while spend the same proportion of time in any of its subsets. Then time averages (1.122) converge to a value that is independent of the initial conditions and equal to space averages (1.120). From the point of view of practical applications, approximation of space average integral by numerical quadrature is in general prohibitively expensive due to the large dimension of $X$. If the system is ergodic, then instead of space averages we can compute time averages in a much less expensive manner.

For example, recall the harmonic oscillator equations (1.1)–(1.2). We assume that the frequency $\omega = 1$ such that the total energy is $H(x, y) = \frac{1}{2}(y^2 + x^2)$ and defines circles as closed orbits in the phase space $\mathbf{R}^2$. We restrict our discussion to the finite phase space, in particular a finite open ball centered at the origin $(0, 0)$. The right hand side vector field of (1.1)–(1.2) is divergence free, hence the flow map is volume (measure) preserving. From the analytical solution (1.3) we identify invariant sets by considering any enclosed area between two energy values $E_1$ and $E_2$ where $E_1 < E_2$. Clearly, these invariant sets are metrically decomposable and the system is not ergodic on these invariant sets. This is because the harmonic oscillator has a conserved quantity, the energy, and dynamics is constrained on periodic orbits in the phase space.

This leads us to the general observation that for general Hamiltonian systems on phase space $\Omega \subset \mathbf{R}^n$ (we restate equation (1.42)):

$$\frac{\mathrm{d}X}{\mathrm{d}t} = J\,\nabla H(X), \qquad X(0) = X_0, \tag{1.123}$$

where $X \in \mathbf{R}^n$ and the Hamiltonian is a conserved quantity of dynamics, the surface of constant $H(X) = E$ separates the phase space into slices of invariant subspaces, which breaks the assumptions of Birkhoff's theorem. This suggests considering an *infinitesimal* phase space volume in the neighbourhood of $H(X) = E$, i.e. the subspace

$$D(E, \mathrm{d}E) = \{X \in \mathbf{R}^n \mid H(X) \in [E; E + \mathrm{d}E]\},$$

for which Birkhoff's theorem may apply. Then the density corresponding to the phase-space volume $D(E, \mathrm{d}E)$ is

$$\rho(X) = \begin{cases} 0, & H(X) \notin [E; E + \mathrm{d}E], \\ 1/\mathrm{vol}\{D\}, & H(X) \in [E; E + \mathrm{d}E], \end{cases}$$

where we assumed that the surface $H(X) = E$ is finite and connected. In the limit $\mathrm{d}E \to 0$ the density $\rho(X)$ becomes concentrated on the surface $H(X) = E$. By using Dirac delta function, the limiting density becomes a singular measure (microcanonical ensemble):

$$\rho_{mc}(X) = \frac{1}{\Sigma(E)}\delta(H(X) - E), \qquad \Sigma(E) = \int_\Omega \delta(H(X) - E)\,\mathrm{d}X. \tag{1.124}$$

The microcanonical measure $\rho_{mc}(X)$ depends on $X$ only through the Hamiltonian $H(X)$ which is a constant of the motion; therefore, $\rho_{mc}(X)$ is a stationary measure. Recall that the Hamiltonian system (1.123) may have other conserved quantities, the first integrals of the system, including the Casimirs. In particular, if (1.123) admits precisely $m+1$ independent first integrals $H(X), I_1(X), \ldots, I_m(X)$, then the subspace $D$ must be restricted to the infinitesimal shell around a surface determined by all first integrals of the system. Hence

$$\rho_{mc}(X) \propto \delta(H(X) - E)\delta(I_1(X) - I_1^0) \cdots \delta(I_m(X) - I_m^0).$$

For the harmonic oscillator considered above, the equations satisfy a form of ergodicity on the subspace

$$D(E, \mathrm{d}E) = \left\{ x, y \in \mathbf{R} \,\middle|\, \frac{1}{2}\left(y^2 + x^2\right) \in [E; E + \mathrm{d}E] \right\}, \qquad \mathrm{d}E \to 0,$$

which is an invariant set of the dynamics and metrically indecomposable. Consider the analytical solution (1.3) of the harmonic oscillator equations (1.1)–(1.2) with $\omega = 1$ and observable $F(x(t), y(t)) = \frac{1}{2}y(t)^2$, the kinetic energy. Since the Hamiltonian function $H(x, y) = \frac{1}{2}(y^2 + x^2) = E$, where $E$ is a positive constant, defines an equation for a circle in $\mathbf{R}^2$, we introduce a polar coordinate transformation:

$$x = \sqrt{2E}\cos(\phi),$$
$$y = \sqrt{2E}\sin(\phi),$$

where $\phi \in [0; 2\pi]$. Then

$$\Sigma(E) = \sqrt{2E}\int_0^{2\pi} \mathrm{d}\phi = 2\pi\sqrt{2E}$$

and straightforward computations show that the time average of $F$ is $\bar{F} = \frac{1}{2}E$ and the space average is

$$\langle F \rangle = \frac{1}{2\pi\sqrt{2E}} \int_\Omega \frac{1}{2} y^2 \delta(H(x,y) - E)\, \mathrm{d}x\, \mathrm{d}y = \frac{E}{2\pi} \int_0^{2\pi} \sin(\phi)^2\, \mathrm{d}\phi = \frac{1}{2}E.$$

Hence $\bar{F} = \langle F \rangle$. The fact that the time average of kinetic energy is half of the total energy is a consequence of the *Virial theorem*, which states that, on average, the kinetic and potential energy share the total energy equally.

### 1.2.2 Microcanonical statistical mechanics

In this subsection we discuss the concepts of the microcanonical entropy and the microcanonical statistical temperature. Consider Hamiltonian system (1.123) with its microcanonical ensemble (1.124). For our discussion we assume that the Hamiltonian $H(X)$ is the only conserved quantity of (1.123). We refer to the normalization constant $\Sigma(E)$ in (1.124) as the limiting measure of the surface $H(X) = E$. For example, in the previous subsection we saw that for the harmonic oscillator the limiting measure of the surface $\frac{1}{2}(y^2 + x^2) = E$ is the circumference of the circle, i.e. $\Sigma(E) = 2\pi\sqrt{2E}$.

The *microcanonical entropy* as a function of $E$ is defined by

$$S(E) = \ln \Sigma(E),$$

which is a monotonically increasing function with respect to $\Sigma(E)$. The importance of the logarithm follows from the consideration of two *independent* systems $A$ and $B$ with state variables $X_A \in \Omega_A$ and $X_B \in \Omega_B$, respectively. For each system we define the energy functions: $H_A(X_A) = E_A$ and $H_B(X_B) = E_B$, which are conserved quantities of each individual system. Hence both systems are equipped with the microcanonical ensemble and the associated limiting measures of surfaces are

$$\Sigma_A(E_A) = \int_{\Omega_A} \delta(H_A(X_A) - E_A)\, \mathrm{d}X_A, \quad \Sigma_B(E_B) = \int_{\Omega_B} \delta(H_B(X_B) - E_B)\, \mathrm{d}X_B.$$

Then the total measure for independent systems is given by

$$\begin{aligned} \Sigma(E_A, E_B) &= \Sigma_A(E_A)\Sigma_B(E_B) \\ &= \int_{\Omega_A} \int_{\Omega_B} \delta(H_A(X_A) - E_A)\delta(H_B(X_B) - E_B)\, \mathrm{d}X_A\, \mathrm{d}X_B \end{aligned}$$

and the corresponding entropy of $\Sigma(E_A, E_B)$ is

$$S(E_A, E_B) = \ln \Sigma(E_A, E_B) = \ln \Sigma_A(E_A) + \ln \Sigma_B(E_B) = S_A(E_A) + S_B(E_B).$$

This shows that the entropy is *additive* for the independent systems. In general, this is not true for *coupled* systems. In the following we consider the simplest form of

coupling which allows energy to be exchanged between two systems $A$ and $B$. The total energy of the coupled system $AB$ is defined by

$$H = H(X_A, X_B) = H_A(X_A) + H_B(X_B) = E.$$

The phase space $\Omega_{AB}$ of the coupled system $AB$ is the product space $\Omega_A \times \Omega_B$ and the associated microcanonical ensemble of the system $AB$ is

$$\rho_{mc}(X_A, X_B) = \frac{1}{\Sigma(E)}\delta(H(X_A, X_B) - E), \tag{1.125}$$
$$\Sigma(E) = \int_{\Omega_A}\int_{\Omega_B} \delta(H(X_A, X_B) - E)\,\mathrm{d}X_A\,\mathrm{d}X_B.$$

Note that

$$\Sigma(E) \neq \Sigma_A(E_A)\Sigma_B(E_B).$$

Hence for the coupled systems $A$ and $B$:

$$S(E) \neq S_A(E_A) + S_B(E_B).$$

Now we ask what is the probability of a particular state $X_A$ under the condition that the joint system $AB$ satisfies the microcanonical dynamics with the total energy $H = E$. The probability density function (1.125) in $X_B$ for fixed $X_A = \bar{X}_A$ is simply

$$\rho_{mc}(\bar{X}_A, X_B) = \frac{1}{\Sigma(E)}\delta(H(\bar{X}_A, X_B) - E) = \frac{1}{\Sigma(E)}\delta(H_A(\bar{X}_A) + H_B(X_B) - E).$$

To find the probability of a particular state $X_A$, we integrate the density function $\rho_{mc}(X_A, X_B)$ over all states of $X_B \in \Omega_B$, i.e.

$$\mathrm{P}\,\{X_A \mid H = E\} = \int_{\Omega_B} \rho_{mc}(X_A, X_B)\,\mathrm{d}X_B \tag{1.126}$$
$$= \frac{1}{\Sigma(E)}\int_{\Omega_B} \delta(H_A(X_A) + H_B(X_B) - E)\,\mathrm{d}X_B = \frac{\Sigma_B(E - H_A(X_A))}{\Sigma(E)}.$$

This probability will be of use in the next subsection.

Now we ask what is the probability that $H_A = E_A$ under the condition that $H = E$. The probability density function (1.125) reduces to the conditional probability density function

$$\rho_{mc}(X_A, X_B|H_A = E_A) = \frac{1}{\Sigma(E)}\delta(E_A + H_B(X_B) - E)\delta(H_A(X_A) - E_A).$$

To find the probability we integrate $\rho_{mc}(X_A, X_B|H_A = E_A)$ over all states of $X_A \in \Omega_A$ and $X_B \in \Omega_B$:

$$\mathrm{P}\,\{H_A = E_A \mid H = E\} = \int_{\Omega_A}\int_{\Omega_B} \rho_{mc}(X_A, X_B|H_A = E_A)\,\mathrm{d}X_A\,\mathrm{d}X_B$$
$$= \frac{1}{\Sigma(E)}\int_{\Omega_A} \delta(H_A(X_A) - E_A)\,\mathrm{d}X_A \int_{\Omega_B} \delta(E_A + H_B(X_B) - E)\,\mathrm{d}X_B$$
$$= \frac{\Sigma_A(E_A)\Sigma_B(E - E_A)}{\Sigma(E)}.$$

To find the most probable occurring macrostate $E_A$, which characterises the most likely energy split of $H = E$ into $H_A = E_A$ and $H_B = E - E_A$, we consider maximization problem:

$$\max_{E_A} \left[\Sigma_A(E_A)\Sigma_B(E - E_A)\right].$$

In terms of entropies the maximization problem reduces to

$$\max_{E_A} \left[S_A(E_A) + S_B(E - E_A)\right].$$

Thus the most probable macrostate $E_A$ occurs at the maximum of the total entropy. With the formal assumption that the function in the maximization problem is differentiable, the maximum then occurs at $E_A^*$ and we obtain:

$$\left.\frac{\partial S_A}{\partial E_A}\right|_{E_A = E_A^*} = \left.\frac{\partial S_B}{\partial E_B}\right|_{E_B = E - E_A^*}.$$

This condition has motivated the definition of the *microcanonical statistical temperature* $T$ by

$$\frac{\partial S(E)}{\partial E} = \frac{1}{\mathrm{T}},$$

such that $\mathrm{T}_A = \mathrm{T}_B$ at the most probable macrostate $E_A^*$. In practice we adopt the *inverse statistical temperature* $\beta = 1/\mathrm{T}$.

So far we have not specified anything about the relative sizes of the systems $A$ and $B$ in the coupled system $AB$. We address this question in the following subsection.

### 1.2.3 Canonical statistical mechanics

We proceed as in the previous subsection and consider a coupled system $AB$. In this subsection we will consider the special case when system $B$ in the coupled system $AB$ is very large relative to system $A$. The very large system $B$ is called an *energy reservoir* for system $A$. Then, if the entropy $S_B(E - E_A)$ is slowly varying over the relevant range of $E_A$, we can greatly simplify the computation of the probabilities for system $A$. We rewrite the probability (1.126) of $X_A$ in terms of the entropy of the system $B$, i.e.

$$\mathrm{P}\left\{X_A \mid H = E\right\} = \frac{\exp\left(S_B(E - E_A)\right)}{\Sigma(E)},$$

where $E_A = H_A(X_A)$. By Taylor expansion of the entropy function $S_B(E - E_A)$ around the value of the total energy $E$ and by truncating it after one term, we obtain:

$$\mathrm{P}\left\{X_A \mid H = E\right\} \propto \exp\left(S_B(E) - \beta_B E_A\right) \propto \exp\left(-\beta_B E_A\right),$$

where the inverse statistical temperature $\beta_B$ is evaluated at $E$ and the constant term $S_B(E)$ has been absorbed into the normalization constant. In practical applications

it is common to consider, as a good starting point, a perfect reservoir, which is characterized by constant inverse temperature $\beta$.

Consider the Hamiltonian system (1.123) in place of system $A$ and system $B$ being an energy reservoir to system $A$. Then $E_A = H(X)$, and the derivation above motivates the introduction of the *canonical* probability distribution (canonical ensemble) of a system (1.123) with an energy reservoir:

$$\rho_c(X) = \frac{1}{\Sigma(\beta)} \exp(-\beta H(X)), \qquad \Sigma(\beta) = \int_\Omega \exp(-\beta H(X)) \, \mathrm{d}X. \tag{1.127}$$

Clearly, a typical trajectory of (1.123) does not ergodically sample a distribution like (1.127). Due to preservation of the Hamiltonian $H(X)$ there is no energy exchange between systems $A$ and $B$, since $H_A = const$ for all times. In molecular dynamics it is desirable to sample distributions like (1.127) and a number of mechanisms have been introduced to model the thermal exchange with the reservoir. These *thermostats* perturb the Hamiltonian vector field so typical trajectories do ergodically sample the correct distribution. We describe various thermostat methods in the following subsection.

In passing we mention that if the initial condition $X_0$ of the general system (1.121) is a random variable with the probability density function $\rho_0(X) : \Omega \to \mathbf{R}$, so that $X(t)$ solving (1.121) is a random variable, then a probability density function $\rho(X,t) : \Omega \times \mathbf{R} \to \mathbf{R}$, satisfying

$$\int_\Omega \rho(X,t) \, \mathrm{d}X = 1, \qquad \rho(X,t) \geq 0 \quad \forall \, t \in \mathbf{R}, \tag{1.128}$$

is transported under the flow of (1.121) according to the *Liouville equation*:

$$\frac{\partial}{\partial t}\rho(X,t) = \mathcal{L}^* \rho(X,t) = -\nabla \cdot (f(X)\rho(X,t)), \qquad \rho(X,0) = \rho_0(X). \tag{1.129}$$

The Liouville operator $\mathcal{L}^*$ is the formal $L^2$-adjoint operator of the generator $\mathcal{L}$, where

$$\mathcal{L}F(X) = f(X) \cdot \nabla F(X),$$

for $F(X)$ some observable. If function $f(X)$ is divergence-free, then $\mathcal{L} = -\mathcal{L}^*$. Recall that if $X(t)$ solves (1.121) and $V(X) : \Omega \to \mathbf{R}$ is any continuously differentiable function, then

$$\frac{\mathrm{d}}{\mathrm{d}t} V(X(t)) = \mathcal{L}V(X(t)). \tag{1.130}$$

We say that the probability density function $\rho(X)$ is a stationary density function of the Liouville equation (1.129) if

$$\mathcal{L}^* \rho(X) = 0.$$

For the Hamiltonian system (1.123) the canonical probability distribution function (1.127) is a stationary density function, since

$$\mathcal{L}^* \rho_c(X) = -\rho_c(X) \, \nabla \cdot (J \, \nabla H(X)) + \beta \rho_c(X) \, \nabla H(X)^T J \, \nabla H(X) = 0. \tag{1.131}$$

In fact, any density function dependent only on the Hamiltonian $H(X)$ and/or any other first integral of the system (1.123) is a stationary density function of the Liouville equation (1.129).

### 1.2.4 Stochastic-dynamical thermostats

In this subsection we introduce stochastic-dynamical thermostat methods applied to the Hamiltonian system (1.123) for sampling the canonical distribution function (1.127). We begin our introduction with the definition of Wiener process.

A one-dimensional *Wiener process* (also called *Brownian motion*) is a stochastic process $\{w(t)\}_{t\geq 0}$ with the following properties: $w(0) = 0$, the function $w(t)$ is continuous in $t$ with probability 1, the process has *stationary*, *independent increments* and the increment $w_t - w_s \sim \mathcal{N}(0, t-s)$ for all $0 \leq s < t$, i.e. normally mean zero distributed with variance $t - s$. The term stationary increments means that the distribution of $w_t - w_s$ is independent of $s$, and so identical to the distribution of $w_t$. The term independent increments means that for nonoverlapping time intervals the respective increments are jointly independent.

Consider a stochastic differential equation system defined on $\Omega \subset \mathbf{R}^n$:

$$\mathrm{d}X = f(X)\,\mathrm{d}t + \Sigma(X)\,\mathrm{d}W, \qquad X(0) = X_0, \tag{1.132}$$

where $X \in \mathbf{R}^n$, $f(X) : \Omega \to \mathbf{R}^n$ and $\Sigma(X) : \Omega \to \mathbf{R}^{n\times m}$ are smooth functions of $X$, and $W(t)$ is a vector of $m \leq n$ independent Wiener processes. We assume that the distribution of $X$ has a probability density function $\rho(X,t) : \Omega \times \mathbf{R} \to \mathbf{R}$ satisfying conditions (1.128) and $X_0$ is a random variable with density $\rho_0(X) : \Omega \to \mathbf{R}$. Then $\rho(X,t)$ satisfies the *Fokker-Planck equation*, also known as forward Kolmogorov equation:

$$\begin{aligned}\frac{\partial}{\partial t}\rho(X,t) = \mathcal{L}^*\rho(X,t) &= -\nabla\cdot(f(X)\rho(X,t)) \\ &\quad + \frac{1}{2}\nabla\cdot\nabla\cdot(\Sigma(X)\Sigma(X)^T\rho(X,t)), \qquad \rho(X,0) = \rho_0(X),\end{aligned} \tag{1.133}$$

where $\nabla \cdot A(X)$ denotes the divergence over the columns of matrix $A(X)$ for each row.

The Fokker-Planck operator $\mathcal{L}^*$ is the formal $L^2$-adjoint operator of the generator $\mathcal{L}$ defined by

$$\mathcal{L}F(X) = f(X)\cdot\nabla F(X) + \frac{1}{2}\Sigma(X)\Sigma(X)^T : \nabla\nabla F(X),$$

where $\nabla\nabla F$ denotes the Hessian matrix of $F$ and $A : B = \mathrm{trace}(AB^T)$, that is, the sum over all components of the element-wise product of matrices $A$ and $B$. For any twice continuously differentiable function $V(X) : \Omega \to \mathbf{R}$, evaluated at the solution of (1.132), the generator $\mathcal{L}$ yields *Itô's formula*:

$$\mathrm{d}V(X) = \mathcal{L}V(X)\,\mathrm{d}t + \langle \nabla V(X), \Sigma(X)\,\mathrm{d}W\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes Euclidean inner product in $\mathbf{R}^n$. Note that if $\Sigma \equiv 0$ we recover equation (1.130) and the Fokker-Planck equation (1.133) reduces to the Liouville equation (1.129).

The density function $\rho(X)$ is an *equilibrium* (stationary) density function of the system (1.132) if it is a stationary solution of the Fokker-Planck equation (1.133), i.e.

$$\mathcal{L}^*\rho(X) = 0.$$

For example, consider the scalar Ornstein-Uhlenbeck (OU) process on $\mathbf{R}$:

$$\mathrm{d}\xi = -\gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w, \qquad \xi(0) = \xi_0, \tag{1.134}$$

where $\xi \in \mathbf{R}$, $\gamma, \sigma > 0$ and $w(t)$ is a scalar Wiener process. With $\alpha = 2\gamma/\sigma^2$, the normal distribution density function with mean zero and variance $\alpha^{-1}$, i.e.

$$\vartheta(\xi) = \sqrt{\frac{\alpha}{2\pi}} \exp\left(\frac{-\alpha}{2}\xi^2\right), \tag{1.135}$$

satisfies the stationary Fokker-Planck equation

$$\gamma\frac{\partial}{\partial\xi}(\xi\vartheta(\xi)) + \frac{1}{2}\sigma^2\frac{\partial^2}{\partial\xi^2}\vartheta(\xi) = 0. \tag{1.136}$$

It is well known that the density (1.135) is the unique steady state solution of the Fokker-Planck equation associated to the Ornstein-Uhlenbeck process (1.134). Hence, solutions of (1.134) ergodically sample (1.135). Recall that the OU process has an analytical solution:

$$\xi(t) = e^{-\gamma t}\xi_0 + \sigma\sqrt{\frac{1-e^{-2\gamma t}}{2\gamma}}\Delta w, \tag{1.137}$$

where $\Delta w \sim \mathcal{N}(0,1)$, that is, normally mean zero distributed random number with unit variance. Hence the time dependent expectation values of $\xi$ and its variance are

$$\mathbf{E}\{\xi(t)\} = e^{-\gamma t}\xi_0, \qquad \mathbf{E}\{(\xi(t) - \mathbf{E}\{\xi(t)\})^2\} = \frac{1}{\alpha}\left(1 - e^{-2\gamma t}\right),$$

i.e. $\mathbf{E}$ denotes at given time $t$ the mean value over all realizations of $\Delta w$. In time, when $t \to \infty$, expectation values converge to the ensemble averages in the measure (1.135):

$$\langle\xi\rangle = \sqrt{\frac{\alpha}{2\pi}}\int_{\mathbf{R}} \xi\exp\left(\frac{-\alpha}{2}\xi^2\right)\mathrm{d}\xi = 0, \quad \langle\xi^2\rangle = \sqrt{\frac{\alpha}{2\pi}}\int_{\mathbf{R}} \xi^2\exp\left(\frac{-\alpha}{2}\xi^2\right)\mathrm{d}\xi = \frac{1}{\alpha},$$

respectively.

The Fokker-Planck equation (1.133) is a second order partial differential equation and the associated operator $\mathcal{L}^*$ is a second order differential operator. Any second order differential operator $\mathcal{P}$ with $C^\infty$ coefficients is called *hypoelliptic* on an open set $U$ if all distributional solutions $\rho$ of the differential equation $\mathcal{P}\rho = 0$ are $C^\infty$. A sufficient criterion for hypoellipticity is provided by *Hörmander's condition.* Let $U \subset \mathbf{R}^n$ be an open set, let $V_0(X) : U \to \mathbf{R}^n$ be a vector filed and let $\mathcal{I}(V_0, V_1, \ldots, V_m)$ denote the ideal of the vector fields $V_k(X) : U \to \mathbf{R}^n$ with $k = 1, \ldots, m$ within the Lie algebra generated by all of the $\{V_0(X), \ldots, V_m(X)\}$:

$$\mathcal{I}(V_0, V_1, \ldots, V_m) = \{V_{k_0}, [V_{k_0}, V_{k_1}], [[V_{k_0}, V_{k_1}], V_{k_2}], \ldots\},$$

where $[\cdot,\cdot]$ denotes the commutator of vector fields, $k_0$ takes values in the set $\{1, \ldots, m\}$, and $k_1$, $k_2$, etc. take values in $\{0, \ldots, m\}$. Then the vector fields $V_0(X), \ldots, V_m(X)$ satisfy Hörmander's condition at $X \in U$ if

$$\mathbf{R}^n \subset \operatorname{span}\mathcal{I}(V_0, V_1, \ldots, V_m).$$

The main application of Hörmander's condition is the following theorem, *Hörmander's theorem*:

**Theorem 1.2.4.1.** *Let $U \subset \mathbf{R}^n$ be open and let $\mathrm{V}(X) = [V_1(X)\, V_2(X) \ldots V_m(X)]$ be a matrix of vector fields $V_1(X), \ldots, V_m(X)$. If Hörmander's condition $\mathbf{R}^n \subset span\mathcal{I}(V_0, V_1, \ldots, V_m)$ is satisfied at every $X \in U$, then the operator $\mathcal{P}$ which is defined by*

$$\mathcal{P}\rho(X) = -\nabla(V_0(X)\rho(X)) + \frac{1}{2}\nabla \cdot \nabla \cdot (\mathrm{V}(X)\mathrm{V}(X)^T \rho(X)),$$

*where $\rho(X) : U \to \mathbf{R}$, is hypoelliptic.*

Clearly, hypoellipticity provides smoothness for stationary solutions of Fokker-Planck equation (1.133), if the associated ideal of vector fields $V_0(X) = f(X)$ and $V_k(X) = \Sigma(X)_k$, where $\Sigma(X)_k$ is the $k^{th}$ column of matrix $\Sigma(X)$, span $\mathbf{R}^n$ at every $X \in \Omega$. For example, the trivial case is when $\Sigma(X) \equiv I_n$ where $I_n \in \mathbf{R}^{n\times n}$ is an identity matrix. Hörmander's condition is automatically satisfied, since the column vectors of $I_n$ form a basis of $\mathbf{R}^n$.

In the following we assume that the solutions of (1.132) exist for all times. If there exists everywhere a positive stationary density function $\rho(X)$ of the Fokker-Planck equation (1.133) on an open, connected set $\Omega \subset \mathbf{R}^n$, that is invariant under the flow (1.132), and Hörmander's condition is satisfied at every $X \in \Omega$, then $\rho(X)$ is the unique invariant measure on $\Omega$, and hence ergodic. Thermostated dynamic equations, presented in the following, are 'special', in the sense that the invariant measure is known by construction. Hence our main concern is to verify Hörmander's condition. Often, in practical applications, we can only rely only on the numerical verification of ergodicity.

As the first thermostat method, applied to the Hamiltonian system (1.123), we mention the generalized Langevin dynamics on phase space $\Omega \subset \mathbf{R}^n$:

$$\mathrm{d}X = J\,\nabla H(X)\,\mathrm{d}t - \frac{\beta}{2}\Sigma\Sigma^T\,\nabla H(X)\,\mathrm{d}t + \Sigma\,\mathrm{d}W, \tag{1.138}$$

where $X \in \mathbf{R}^n$, $W(t)$ is a vector of $m$ independent Wiener processes and $0 < m \leq n$. A constant matrix $\Sigma \in \mathbf{R}^{n\times m}$ has rank $m$ and the matrix product $\Sigma\Sigma^T$ is positive definite. The Langevin dynamics (1.138) is constructed by adding stochastic noise with balanced dissipation to the Hamiltonian equations (1.123), such that the canonical probability distribution function (1.127) is an equilibrium density function of the associated Fokker-Planck equation (1.133). Here and in the following we assume that the Hamiltonian function $H(X)$ is such that the measure (1.127) can be normalized, i.e. $\Sigma(\beta) < \infty$.

One limitation of the Langevin approach is that it destroys all invariants of the original system. To retain some of these it would be necessary to introduce constraint projections which may create significant difficulties in discretization. It is recognized that additive noise is much easier to treat accurately in discretization than multiplicative noise.

Note that if Hamiltonian $H(X)$ is a quadratic function and the constant matrix $\Sigma$ is a rectangular diagonal matrix with positive entries, then the balanced noise

and dissipation in the Langevin dynamics (1.138) consists of independent scalar OU processes (1.134).

We show that the probability density function (1.127) is a stationary density function of the Fokker-Planck equation (1.133). Note the result (1.131) and that

$$\nabla \cdot (\Sigma\Sigma^T \rho_c(X)) = \Sigma\Sigma^T \nabla\rho_c(X) = -\beta\Sigma\Sigma^T \nabla H(X)\rho_c(X).$$

Then

$$\begin{aligned}\mathcal{L}^*\rho_c(X) &= \frac{\beta}{2}\nabla \cdot (\rho_c(X)\Sigma\Sigma^T \nabla H(X)) + \frac{1}{2}\nabla\cdot\nabla\cdot(\Sigma\Sigma^T\rho_c(X)) \\ &= \frac{\beta}{2}\nabla \cdot (\rho_c(X)\Sigma\Sigma^T \nabla H(X) - \Sigma\Sigma^T \nabla H(X)\rho_c(X)) = 0.\end{aligned}$$

This proves the statement.

In the molecular dynamics, Langevin dynamics applied to the canonical Hamiltonian (Newtonian) dynamics (1.58)–(1.59) for $q, p \in \mathbf{R}^n$ reads:

$$\mathrm{d}q = M^{-1}p\,\mathrm{d}t, \tag{1.139}$$

$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t - \gamma M^{-1}p\,\mathrm{d}t + \sigma\mathrm{d}W, \tag{1.140}$$

where we consider constant $\sigma > 0$, $\gamma = \beta\sigma^2/2$ and $W(t)$ is a vector of $n$ independent Wiener processes. The canonical density function

$$\rho_c(q,p) \propto \exp\left(-\beta\left(\frac{1}{2}p^T M^{-1}p + V(q)\right)\right)$$

is a stationary density function of the associated Fokker-Planck equation (1.133). In Chapter 3 we will discuss the extension of Langevin dynamics (1.139)–(1.140) to Hamiltonian systems with holonomic constraints (1.69)–(1.71).

Another approach, originally proposed by Nosé [88, 89] and modified by Hoover [46], involves the introduction of an auxiliary variable $\xi$, embedding the Hamiltonian flow in a higher dimensional phase space. This thermostat method is deterministic and the equations are constructed such that the extended probability density function

$$\pi(X,\xi) = \rho(X)\vartheta(\xi) \tag{1.141}$$

is an invariant of the Liouville equations (1.129). The deterministic approach is often non-ergodic, however, motivating the addition of Langevin forcing to the auxiliary variable, which leads to the so called Nosé-Hoover-Langevin thermostat (NHL) in molecular dynamics [61]. Deterministic methods can be extended by including multiple auxiliary variables and by introducing more general coupling than originally considered. A broadened framework was proposed in [58] and termed Generalized Bulgac-Kusnezov (GBK) thermostatting. Here we follow the formulation and derivations from Chapter 5. In the simplest form of a GBK thermostat, we augment the Hamiltonian system (1.123) with a small number of additional variables $\xi_k$, $k = 1, \ldots, d_T$, and perturbation vector fields, which for our purposes may be assumed to be linear in the $\xi_k$. Let $g_k(X) : \Omega \to \mathbf{R}^n$, $k = 1, \ldots, d_T$, be smooth

vector fields. The complete system is then a set of coupled ordinary and stochastic differential equations of the form:

$$\mathrm{d}X = J\,\nabla H(X)\,\mathrm{d}t + \sum_{k=1}^{d_T} \xi_k g_k(X)\,\mathrm{d}t, \tag{1.142}$$

$$\mathrm{d}\xi_k = h_k(X)\,\mathrm{d}t - \gamma\xi_k\,\mathrm{d}t + \sigma\,\mathrm{d}w_k, \quad k = 1,\ldots,d_T, \tag{1.143}$$

where $\gamma = \alpha\sigma^2/2$ and the $w_k(t)$ are independent scalar Wiener processes. In general we consider the number of thermostat variables $d_T$ to be small, say $d_T = 1$ or $d_T = 2$, so that the computational cost of simulating the thermostat variables is negligible to that of simulating the physical model.

For a given distribution $\rho(X)$, we seek functions $h_k(X) : \Omega \to \mathbf{R}$, $k = 1,\ldots,d_T$, such that the extended probability density function (1.141) is a stationary solution of the Fokker-Planck equation (1.133) associated with (1.142)–(1.143), i.e.

$$\begin{aligned}\nabla\cdot\pi(X,\xi)\,&\left(J\,\nabla H(X) + \sum_k \xi_k g_k(X)\right) \\ &+ \sum_k \left[\frac{\partial}{\partial\xi_k}\left(\pi(X,\xi)(h_k(X) - \gamma\xi_k)\right) - \frac{\sigma^2}{2}\frac{\partial^2}{\partial\xi_k^2}\pi(X,\xi)\right] = 0.\end{aligned} \tag{1.144}$$

For simplicity we consider the canonical distribution (1.127) in place of $\rho(X)$ in (1.141). The expression (1.144) simplifies under the conditions $\nabla\cdot(J\,\nabla H(X)) = 0$, $\nabla H(X)^T J\,\nabla H(X) = 0$. Using the fact that the terms of the OU process (1.134) satisfy the stationary Fokker-Planck equation (1.136), the relation (1.144) reduces to

$$\begin{aligned}0 &= \sum_k \xi_k\,\nabla\cdot\pi(X,\xi)g_k(X) + h_k(X)\frac{\partial}{\partial\xi_k}\pi(X,\xi) \\ &= \sum_k \xi_k\pi(X,\xi)\,\nabla\cdot g_k(X) - \beta\xi_k\pi(X,\xi)\,(\nabla H(X)\cdot g_k(X)) - \alpha\xi_k\pi(X,\xi)h_k(X) \\ &= \sum_k \xi_k\,(\nabla\cdot g_k(X) - \beta\,\nabla H(X)\cdot g_k(X) - \alpha h_k(X))\,.\end{aligned}$$

Hence it is sufficient to take

$$h_k(X) = \frac{1}{\alpha}\,(\nabla\cdot g_k(X) - \beta\,\nabla H(X)\cdot g_k(X))$$

for given vector fields $g_k(X)$. We have yet to specify these vector fields. In doing so, choices may be motivated from the application point of view, since the vector fields $g_k$ determine the direction of perturbation, or through analytical considerations. For more discussion, see Chapter 4.

When the Hamiltonian system (1.123) is (1.58)–(1.59), $(q,p)^T \in \mathbf{R}^{2n}$, $d_T = 1$ and $g_1 = (0,p)^T$, we recover the NHL thermostat method considered in molecular

dynamics, i.e.

$$\mathrm{d}q = M^{-1}p\,\mathrm{d}t, \tag{1.145}$$
$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t + \xi p\,\mathrm{d}t, \tag{1.146}$$
$$\mathrm{d}\xi = \frac{1}{\alpha}\left(n - \beta p^T M^{-1} p\right)\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w, \tag{1.147}$$

with extended stationary canonical ensemble:

$$\pi(q,p,\xi) \propto \exp\left(-\beta\left(\frac{1}{2}p^T M^{-1}p + V(q)\right)\right)\exp\left(\frac{-\alpha}{2}\xi^2\right).$$

As an example we consider Langevin dynamics (1.139)–(1.140) and the NHL method (1.145)–(1.147) applied to a Hamiltonian system with a triple well potential

$$V(q) = \frac{1}{2}q^2 - \frac{0.6}{4}q^4 + \frac{0.07}{6}q^6, \tag{1.148}$$

where $q \in \mathbf{R}$, and the mass matrix $M = 1$. In both methods we fix $\beta = 1$. All numerical methods of this subsection are described in the next subsection in the context of time integration for thermostated dynamics. In Figures 1.4(a) and 1.4(b) we plot a single stochastic trajectory of both methods for relative comparison. Additionally in these two figures we plot a few closed orbits associated to the original Hamiltonian system. The trajectory of the NHL method (1.145)–(1.147) looks relatively more regular than the trajectory of the Langevin dynamics (1.139)–(1.140). This can be explained from the fact that noise is once integrated before perturbing the Hamiltonian system through the auxiliary variable $\xi$ and the perturbations are always in the direction of $p$. In contrast, in Langevin dynamics noise enters directly into the equation of $p$ and acts in all directions, and this is true for any positive value of $\sigma$. The parameter values, i.e. $\gamma$ and $\alpha$, were chosen for presentation purposes only. The noted differences between Langevin and NHL methods have direct effects on autocorrelation functions, see [60] and Chapter 3.

We show that the Langevin dynamics (1.139)–(1.140) with potential (1.148) satisfy Hörmander's condition. In this case we have only two vector fields:

$$V_0(q,p) = (p, -q + 0.6q^3 - 0.07q^5 - \gamma p)^T, \qquad V_1(q,p) = (0,\sigma)^T.$$

The commutator of these vector fields reads:

$$\begin{aligned}[V_0, V_1] &= (\nabla V_0)V_1 - (\nabla V_1)V_0 \\ &= \begin{bmatrix} 0 & 1 \\ -1 + 1.8q^2 - 0.35q^4 & -\gamma \end{bmatrix}\begin{pmatrix} 0 \\ \sigma \end{pmatrix} = \begin{pmatrix} \sigma \\ -\gamma\sigma \end{pmatrix}.\end{aligned}$$

Hence Hörmander's condition is satisfied at each point $(q,p)^T \in \mathbf{R}^2$, since the vectors $(0,\sigma)^T$ and $(\sigma, -\gamma\sigma)^T$ span the whole $\mathbf{R}^2$. In the same way Hörmander's condition can be verified for the NHL method (1.145)–(1.147) with potential (1.148) everywhere except on the line $p = 0$. Note that the function $f(q) = -\nabla V(q) = -q + 0.6q^3 - 0.07q^5$ has five real roots. Hence NHL dynamics has five stationary

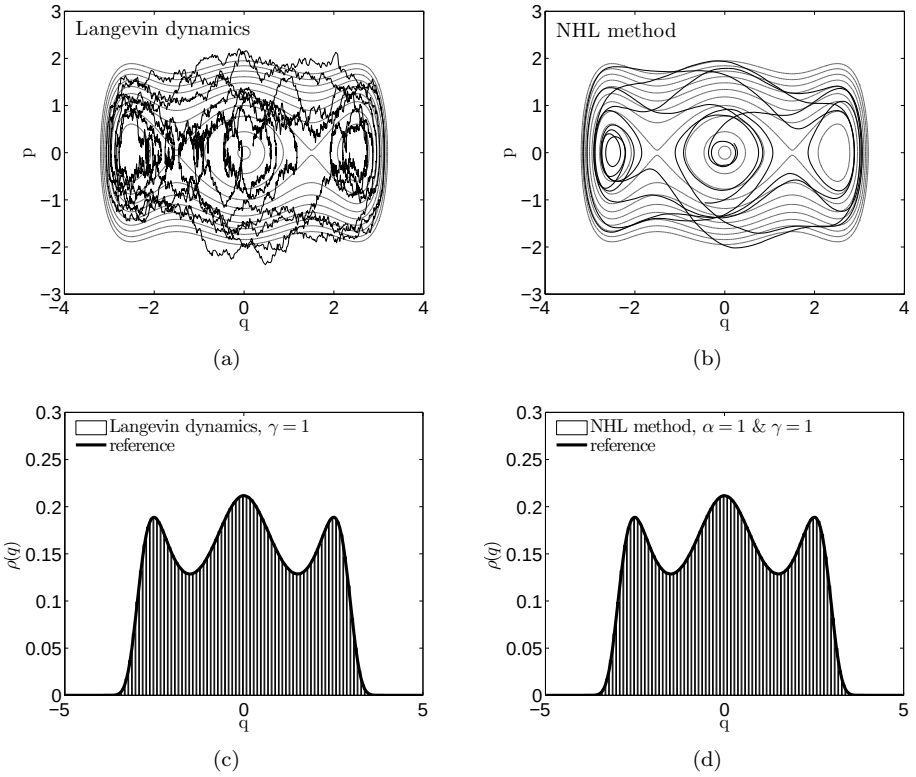



Figure 1.4: Langevin dynamics (1.139)–(1.140) and NHL method (1.145)–(1.147) with triple well potential (1.148). Top: one trajectory of stochastic dynamics. Bottom: probability density function of $q$. (a) Langevin dynamics $\gamma = 0.1$. (b) NHL method with $\alpha = 5$ and $\gamma = 0.1$. (c) Langevin dynamics compared to reference. (d) NHL method compared to reference.

points with $p = 0$. This implies that if an initial condition is one of those stationary points then the solution with respect to variables $q$ and $p$ will stay there for all times.

We perform a numerical test of ergodicity. In Figures 1.4(c) and 1.4(d) we compare the numerically computed probability density function (pdf) of $q$ to Monte Carlo simulations using the Metropolis-Hastings algorithm. The histograms were computed from $10^9$ data points collected during the long time simulation of Langevin and NHL equations. Results show that a single numerically computed trajectory produces what is essentially a perfect pdf of the variable $q$.

Hörmander's condition can be tailored neatly to the GBK thermostat (1.142)–(1.143), as demonstrated next. Denoting by $\partial_{\xi_k}$ the unit vector in $\mathbf{R}^{n+d_T}$ corresponding to the variable $\xi_k$, Hörmander's condition for this system is

$$\mathbf{R}^{n+d_T} \subset \operatorname{span} \mathcal{I}(F, \partial_{\xi_1}, \ldots, \partial_{\xi_{d_T}}),$$

where

$$F = \begin{pmatrix} J\nabla H(X) + \sum_k \xi_k g_k(X) \\ h_1(X) - \gamma\xi_1 \\ \vdots \\ h_{d_T}(X) - \gamma\xi_{d_T} \end{pmatrix}$$

denotes the deterministic vector field of (1.142)–(1.143). Defining

$$G_k(X) = [F, \partial_{\xi_k}] = \begin{pmatrix} g_k(X) \\ -\gamma\partial_{\xi_k} \end{pmatrix}, \quad k = 1, \ldots, d_T,$$

we find that

$$[F, G_k] = \begin{pmatrix} [J\nabla H, g_k] \\ 0 \end{pmatrix} + c_1 G_k(X) + c_2(X)\partial_{\xi_k}. \tag{1.149}$$

Since the unit vectors $\partial_{\xi_k}$ form a globally defined basis for the auxiliary space of the thermostat variables $\xi_k$, it remains to construct a basis for the original space $\mathbf{R}^n$. Eliminating the $\xi_k$ and the $G_k(X)$ from (1.149), shows that the following reduced Hörmander condition holds (this Lemma appears in Chapter 4):

**Lemma 1.2.4.1.** *The GBK method (1.142)–(1.143) satisfies Hörmander's condition at a point* $(X, \xi_1, \ldots, \xi_{d_T}) \in \mathbf{R}^{n+d_T}$ *if the related Hörmander condition on* $\mathbf{R}^n$ *holds at* $X$*:*

$$\mathbf{R}^n \subset \operatorname{span} \mathcal{I}(J\nabla H, g_1, g_2, \ldots, g_{d_T}).$$

When choosing appropriate vector fields $g_k(X)$, it is important to ensure that the vector fields $J\nabla H(X)$ and the $g_k(X)$ do not all share an invariant manifold of co-dimension one, since sets of co-dimension one can divide the phase space into left-right or inner-outer regions that cannot be reached and Hörmander's condition will fail there.

To illustrate the above considerations, as a counterexample we consider the thermostated harmonic oscillator equations (1.1)–(1.2) with GBK method:

$$\mathrm{d}q = p\,\mathrm{d}p, \tag{1.150}$$

$$\mathrm{d}p = -q\,\mathrm{d}t + \xi p(1 - q^2 - p^2)\,\mathrm{d}t, \tag{1.151}$$

$$\mathrm{d}\xi = \frac{1}{\alpha}\left(1 - q^2 - 3p^2 - \beta p^2(1 - q^2 - p^2)\right)\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w, \tag{1.152}$$

where $w(t)$ is a scalar Wiener process and $\alpha = 2\gamma/\sigma^2$. Equations (1.150)–(1.152) were constructed such that the extended canonical ensemble

$$\pi(q, p, \xi) \propto \exp\left(-\beta\frac{1}{2}\left(p^2 + q^2\right)\right)\exp\left(\frac{-\alpha}{2}\xi^2\right)$$

is a stationary density of the associated Fokker-Planck equation (1.133). Method (1.150)–(1.152) is the GBK method (1.142)–(1.143) with $d_T = 1$ and $g_1(q, p) = (0, p(1 - q^2 - p^2))^T$. Note that equations (1.150)–(1.151) have a stationary point $(0, 0)$, and the unit circle $q^2 + p^2 = 1$ is an invariant manifold of co-dimension one,

which divides the phase space into inner-outer open invariant sets, such that, if the initial condition of $q$ and $p$ is inside the circle, the solutions will stay inside the circle for all times and vice versa.

We consider a numerical test with two different initial conditions: $(q_0, p_0, \xi_0) = (0.25, 0.25, 0)$ and $(q_0, p_0, \xi_0) = (1.25, 1.25, 0)$, see Figures 1.5(a) and 1.5(b), respectively. From these figures it is evident that a single stochastic trajectory stays in one of the regions determined by the initial condition. Additionally, in Figures 1.5(c) and 1.5(d) we compare the numerically computed histogram of $R = q^2 + p^2$ to Monte Carlo simulations using the Metropolis-Hastings algorithm. The histograms were computed from $10^9$ data points collected during the long time simulation of the GBK method (1.150)–(1.152). The computations were performed for two different initial conditions. In Figure 1.5(c) the system was initialized with the initial condition for $q$ and $p$ inside the circle and in Figure 1.5(d) the initial condition for $q$ and $p$ was chosen outside the circle. Both Figures 1.5(c) and 1.5(d) suggest that the method is ergodic in each of the regions of phase space separated by the unit circle. Hence the numerical test confirms the statement that the circle $q^2 + p^2 = 1$ divides the phase space of $(q, p)$ into two invariant regions.

Compare this to the previous example, where Hörmander's condition was not satisfied for $p = 0$. Since the line $p = 0$ is not invariant (besides at the stationary points, which have co-dimension two), it does not separate the phase space into invariant sets, and therefore does not impede ergodicity.

In conclusion we state that while there exist general analytical tools, such as Hörmander's condition, to verify certain aspects necessary for proving ergodicity; in practical applications it may be hard or even impossible to apply them. Hence we are often left with numerical verification of ergodicity and therefore we need good time integration methods, which are the subject of the following subsection.

### 1.2.5 Time integration and sampling

In this subsection we discuss time integration methods for thermostated dynamics. For thermostated dynamics, all methods in this thesis are based on the splitting methods. Consider dynamical system:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y), \qquad y(0) = y_0, \tag{1.153}$$

where $y \in \mathbf{R}^n$ and vector field $f(y) : \mathbf{R}^n \to \mathbf{R}^n$ can be split into the sum of multiple vector fields. For our presentation we consider the splitting into two vector fields, i.e.

$$f(y) = f_1(y) + f_2(y).$$

Then we split the dynamical system (1.153) into two dynamical systems:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f_1(y), \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = f_2(y), \tag{1.154}$$

with associated flow maps $\Phi_1^t$ and $\Phi_2^t$, respectively. Note that in general flow maps do not commute, that is $\Phi_1^t \circ \Phi_2^t \neq \Phi_2^t \circ \Phi_1^t$, and $\Phi^t \neq \Phi_1^t \circ \Phi_2^t$, where $\Phi^t$ is the exact

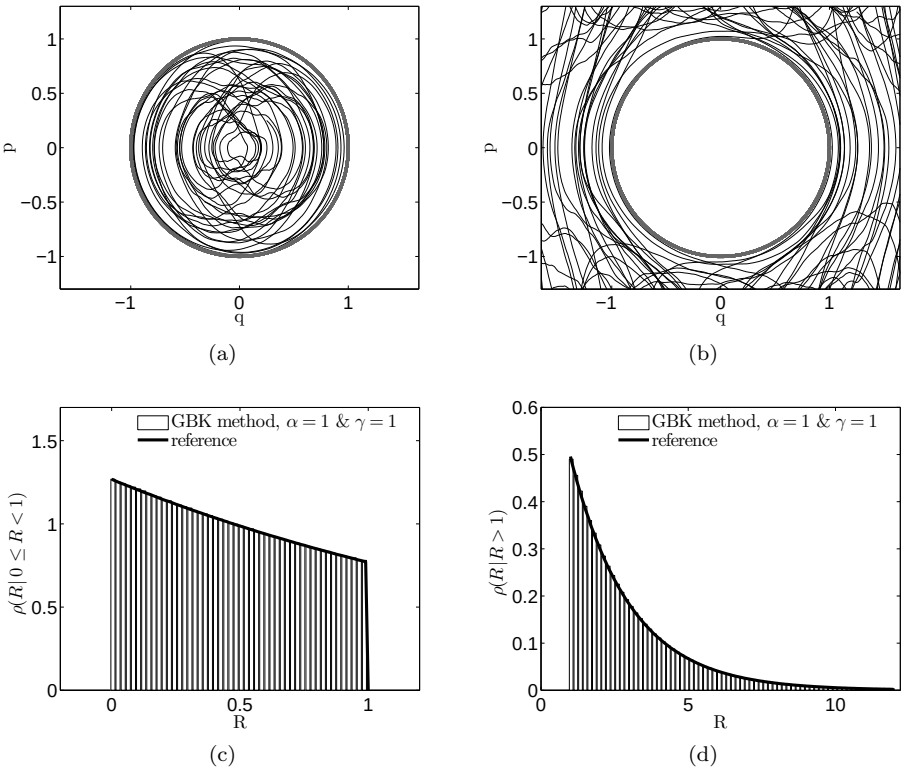



Figure 1.5: Stochastic dynamics of GBK method (1.150)–(1.152), $\beta = 1$, $\alpha = 1$ and $\gamma = 1$. Top: one trajectory of stochastic dynamics. Bottom: GBK method compared to reference for the probability density function of $R = q^2 + p^2$. (a) numerical solution with initial condition $(q_0, p_0, \xi_0) = (0.25, 0.25, 0)$. (b) numerical solution with initial condition $(q_0, p_0, \xi_0) = (1.25, 1.25, 0)$. (c) pdf of $R$ inside the circle, i.e. $0 \leq R < 1$. (d) pdf of $R$ outside the circle, i.e. $R > 1$.

flow map associated with (1.153). Now we apply numerical approximation methods to each of the systems in (1.154):

$$y_1^{n+1} = \Psi_1^\tau(y_1^n), \qquad y_2^{n+1} = \Psi_2^\tau(y_2^n), \tag{1.155}$$

where $\tau$ is a time step, $y_1^n$ and $y_2^n$ are the numerical solutions of two dynamical systems (1.154) at time $t^n = n\tau$ where $n = 0, 1, \ldots$ $\Psi_1^t$ and $\Psi_2^t$ are the associated discrete flow maps of the numerical methods, respectively. Note that if any of the dynamical systems in (1.154) are exactly integrable, then we can consider $\Psi_1^\tau = \Phi_1^\tau$ or $\Psi_2^\tau = \Phi_2^\tau$. This already indicates the advantage of splitting methods.

The idea behind the splitting methods is that the numerical solution of (1.153) can be constructed as the composition method $\Psi^\tau$ of the numerical methods in (1.155). The simplest first order method, *Trotter splitting method*, consists of direct

composition of $\Psi_1^\tau$ and $\Psi_2^\tau$, i.e.

$$\Psi^\tau = \Psi_1^\tau \circ \Psi_2^\tau.$$

As a method it reads:

$$y^* = \Psi_1^\tau(y^n), \qquad y^{n+1} = \Psi_2^\tau(y^*).$$

Another idea would be to use a symmetric version, a so called *Strang splitting*:

$$\Psi^\tau = \Psi_1^{\frac{\tau}{2}} \circ \Psi_2^\tau \circ \Psi_1^{\frac{\tau}{2}}. \tag{1.156}$$

As a method it takes the following form:

$$y^* = \Psi_1^{\frac{\tau}{2}}(y^n), \qquad y^{**} = \Psi_2^\tau(y^*), \qquad y^{n+1} = \Psi_1^{\frac{\tau}{2}}(y^{**}).$$

If both numerical methods $\Psi_1^\tau$ and $\Psi_2^\tau$ are symmetric methods, see Section 1.1.8, or exact analytical flow maps, then the Strang splitting (1.156) yields symmetric method $\Psi^\tau$ of order two. Indeed,

$$[\Psi^\tau]^{-1} = \left[\Psi_1^{\frac{\tau}{2}}\right]^{-1} \circ [\Psi_2^\tau]^{-1} \circ \left[\Psi_1^{\frac{\tau}{2}}\right]^{-1} = \Psi_1^{-\frac{\tau}{2}} \circ \Psi_2^{-\tau} \circ \Psi_1^{-\frac{\tau}{2}} = \Psi^{-\tau}.$$

For example we consider canonical Hamiltonian system (1.55)–(1.56) with separable Hamiltonian, i.e. $H(q,p) = H_1(p) + H_2(q)$. Equations read:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \nabla H_1(p), \tag{1.157}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla H_2(q). \tag{1.158}$$

We can split Hamiltonian system (1.157)–(1.158) into two exactly integrable systems:

$$\begin{array}{rcl} q & = & const, \\ \dfrac{\mathrm{d}p}{\mathrm{d}t} & = & -\nabla H_2(q), \end{array} \qquad \begin{array}{rcl} \dfrac{\mathrm{d}q}{\mathrm{d}t} & = & \nabla H_1(p), \\ p & = & const. \end{array}$$

Application of Strang splitting method (1.156) yields the numerical method:

$$\begin{aligned} p^* &= p^n - \frac{\tau}{2}\nabla H_2(q^n), \\ q^{n+1} &= q^n + \tau\,\nabla H_1(p^*), \\ p^{n+1} &= p^* - \frac{\tau}{2}\nabla H_2(q^{n+1}), \end{aligned}$$

that is exactly the symmetric and symplectic Störmer-Verlet method (1.106)–(1.108) from Section 1.1.8, derived as the Strang splitting for the Hamiltonian system (1.157)–(1.158).

For the thermostat methods presented in the previous subsection we adopt the Strang splitting approach by splitting thermostated equations in deterministic and stochastic parts. Let us illustrate this for the Langevin dynamics (1.139)–(1.140)

and NHL method (1.145)–(1.147). Without loss of generality we take a mass matrix $M$ to be an identity matrix.

We split the Langevin dynamics (1.139)–(1.140) into a deterministic Hamiltonian part:

$$\mathrm{d}q = p\,\mathrm{d}t, \tag{1.159}$$

$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t, \tag{1.160}$$

and stochastic part:

$$\mathrm{d}p = -\frac{\beta\sigma^2}{2}p\,\mathrm{d}t + \sigma\,\mathrm{d}W. \tag{1.161}$$

For solving Hamiltonian part (1.159)–(1.160) we choose any desirable symmetric and symplectic method, e.g. the Störmer-Verlet method, and denote its discrete flow map by $\Psi_H^\tau$. Note that the stochastic part (1.161) decouples into scalar OU processes (1.134) which have analytical solutions given by (1.137). Denote the exact flow map of (1.161) by $\Phi_{OU}^t$. Then the Strang splitting method (1.156) for the Langevin dynamics (1.139)–(1.140) reads:

$$\Psi_{LD}^\tau = \Phi_{OU}^{\frac{\tau}{2}} \circ \Psi_H^\tau \circ \Phi_{OU}^{\frac{\tau}{2}}. \tag{1.162}$$

We split the NHL dynamics (1.145)–(1.147) into three systems, the Hamiltonian dynamics (1.159)–(1.160), a linear equation for $p$:

$$\mathrm{d}p = \xi p\,\mathrm{d}t, \tag{1.163}$$

and an equation for $\xi$, that is equation (1.147). We solve the Hamiltonian system (1.159)–(1.160) with a symmetric and symplectic numerical method $\Psi_H^\tau$. For a fixed value of $\xi$ the equation (1.163) for $p$ can be solved exactly and we denote its flow map by $\Phi_P^t$. Equation (1.147) is a scalar OU process with nonzero mean. For fixed value of $p$, equation has analytical solution:

$$\xi(t) = e^{-\gamma t}\xi_0 + \frac{1}{\alpha\gamma}\left(n - \beta p^T p\right)\left(1 - e^{-\gamma t}\right) + \sigma\sqrt{\frac{1 - e^{-2\gamma t}}{2\gamma}}\Delta w,$$

where $\Delta w \sim \mathcal{N}(0,1)$. We denote its exact flow map by $\Phi_\xi^t$. Then the numerical method reads:

$$\Psi_{NHL}^\tau = \Phi_\xi^{\frac{\tau}{2}} \circ \Phi_P^{\frac{\tau}{2}} \circ \Psi_H^\tau \circ \Phi_P^{\frac{\tau}{2}} \circ \Phi_\xi^{\frac{\tau}{2}}. \tag{1.164}$$

Note that the method is symmetric, since $\left[\Psi_{NHL}^\tau\right]^{-1} = \Psi_{NHL}^{-\tau}$.

These two examples above demonstrate the convenience of using splitting methods for time integration of thermostated dynamics, as well as, for Hamiltonian dynamics in general. There is a freedom for choosing different splitting tactics based on the problem at hand and there is a freedom for choosing the time integration method $\Psi_H^\tau$. In Chapter 3, for example, we consider symmetric splitting methods for thermostated Hamiltonian systems with holonomic constraints. For general theory and introduction to numerical simulations of stochastic differential equations we refer readers to [44, 53].

Numerical results in Figure 1.4 were obtained by numerical methods (1.162) and (1.164) with time step $\tau = 0.01$. We used the Störmer-Verlet method (1.106)–(1.108) for the time integration of Hamiltonian part, $\Psi_H^\tau$. For numerical integration of system (1.150)–(1.152) we adopted a different splitting tactic by splitting the system in two parts, that is, in equations (1.150)–(1.151) and (1.152). For given value of $\xi$ we solved system of equations (1.150)–(1.151) with the implicit midpoint method (1.117) from Section 1.1.8. We denote its flow map by $\Psi_{IM}^\tau$. The nonlinear relations were solved using fixed point iteration to a tolerance of $10^{-14}$. Equation (1.152) for given value of $p$ was solved analytically with flow map $\Phi_\xi^t$. Overall the splitting method reads:

$$\Psi^\tau = \Phi_\xi^{\frac{\tau}{2}} \circ \Psi_{IM}^\tau \circ \Phi_\xi^{\frac{\tau}{2}}.$$

All computations were performed with time step $\tau = 0.005$.

In Chapters 3 and 4 we are concerned with sampling the probability density functions, computation of autocorrelation functions and showing convergence rates for observables. To numerically compute reference pdfs, i.e. histograms, of observables in a particular ensemble, we use a Monte-Carlo method. The Monte-Carlo method is an iteration strategy that combines a randomly generated step with a Metropolis-Hastings accept/reject condition in order to guarantee that the points generated have the desired distribution. To compute a histogram for an observable we restrict a segment of the size relative to the expected values of the observable. Then we divide a segment into $N$ subsegments of the same lengths. The Monte-Carlo algorithm generates points and we count how many points have appeared in the particular subsegment. After that we normalize the data set and plot the numerically computed pdf. We use a Monte-Carlo method for computing reference pdfs and sets of initial conditions sampling a particular distribution, i.e. an ensemble of initial conditions. On the contrary, pdfs of the thermostated methods are computed from the long time simulation using the same counting mechanism. The data points are collected after each $n^{th}$ time step $\tau$ for constant $\tau$. Unless specified otherwise, $n = 1$.

The autocorrelation function $c(s)$ of observable $F(X)$ is defined in the ensemble average by

$$c(s) = \frac{1}{c_0}\langle F(\Phi^s(X))F(X)\rangle, \qquad c_0 = \langle F(X)^2\rangle, \tag{1.165}$$

where $\Phi^s$ is an associated flow map of the Hamiltonian system or thermostated dynamics. If the flow is ergodic, the autocorrelation can be computed from the time average according to

$$\begin{aligned} c(s) &= \frac{1}{c_0}\lim_{T\to\infty}\frac{1}{T}\int_0^T F(X(t))F(X(t+s))\,\mathrm{d}t, \\ c_0 &= \lim_{T\to\infty}\frac{1}{T}\int_0^T F(X(t))^2\,\mathrm{d}t, \end{aligned} \tag{1.166}$$

where $X(t) = \Phi^t(X_0)$, $X(t+s) = \Phi^{t+s}(X_0)$ and $X_0$ is an initial condition.

In the numerical computations, integrals are approximated by the quadrature rules and the function values are computed at the discrete values obtained from the

numerical time integration. The autocorrelation function (1.165) is approximated by

$$\tilde{c}(k\tau) = \frac{1}{\tilde{c}_0} \sum_{m=1}^{M} F(\Psi^{k\tau}(X_0^m))F(X_0^m), \quad \tilde{c}_0 = \sum_{m=1}^{M} F(X_0^m)^2, \quad k = 0, 1, \ldots, K,$$

where $\Psi^\tau$ is the discrete flow map of $\Phi^s$ and $X_0^m$ belongs to the ensemble of $M$ initial conditions, and the autocorrelation function (1.166) is approximated by

$$\tilde{c}(k\tau) = \frac{1}{\tilde{c}_0} \sum_{n=0}^{N} F(\Psi^{n\tau+k\tau}(X_0))F(\Psi^{n\tau}(X_0)),$$
$$\tilde{c}_0 = \sum_{n=0}^{N} F(\Psi^{n\tau}(X_0))^2, \quad k = 0, 1, \ldots, K,$$

where $X_0$ is the initial condition and $N$ denotes the number of time steps. The value $K\tau$ defines the time window for the computed autocorrelation function, i.e. $k\tau$ takes values in segment $[0, K\tau]$.

All reference autocorrelation functions are computed by averaging over constant Hamiltonian simulations, that is, $\Phi^s = \Phi_H^s$ and $\Psi^\tau = \Psi_H^\tau$. The ensembles of initial conditions are provided by a Monte-Carlo method. On the contrary, the autocorrelation functions of thermostated dynamics are computed from the long time simulations which serves as a good test for ergodicity.

In numerical study of the convergence of an observable $F(X)$ in time to its ensemble average $\langle F \rangle$ we compute a time dependent discrete expected value:

$$\mathbf{E}\{F(X(n\tau))\} = \frac{1}{M} \sum_{m=1}^{M} F(\Psi^{n\tau}(X_0^m)), \quad n = 0, 1, \ldots, N,$$

for an ensemble of $M$ initial conditions. To estimate the rate of convergence we compute the log error function given by

$$\log \text{error}(n\tau) = \log \left| \mathbf{E}\{F(X(n\tau))\} - \langle F \rangle \right|, \quad n = 0, 1, \ldots, N.$$

This concludes the introductory section to thermostated dynamics.

# Chapter 2

# Emergence of Internal Wave Attractors

## 2.1 Introduction

Internal gravity waves in uniformly stratified fluids retain their frequency and consequently also their angle with respect to gravity upon reflection from an inclined boundary. Waves do change their wavelength and become focused or defocused when reflecting from plane, inclined surfaces. Laboratory experiments confirm that when a container filled with a uniformly stratified fluid is excited vertically or horizontally, internal gravity waves appear that become focused when reflecting from a sloping wall and converge towards a limit cycle, a so called wave attractor (Maas & Lam [73]; Maas et al. [72]; Hazewinkel et al. [43]). Energy propagates along the straight lines of the attractor, which are normal to the direction of phase propagation. Understanding the behavior of internal waves in bounded domains may be important for explaining the mixing processes in ocean basins and lakes and has relevance to astrophysics and fluid dynamics in general (Bühler & Holmes-Cerfon [6]).

The ideal setting, considered above and used in typical laboratory and theoretical settings (including ours), assumes the fluid's stratification to be uniform, the domain's boundaries to be smooth and the setting to be 2D. Non-uniform stratification, rough topography and three-dimensionality may, however, all lead to scattering of the internal wave field. Moreover dissipation and nonlinear wave interaction limit the amplification of internal waves and might thus prohibit the ultimate localization of internal waves onto wave attractors.

Nevertheless, laboratory and numerical experiments have shown that wave attractors may be resilient to some of these perturbations. In the laboratory, attractors were shown to persist despite basins having non-uniform stratification, small-scale boundary corrugations (Hazewinkel et al. [42]) or being forced non-centrally in a 3D (paraboloidal) domain (Hazewinkel et al. [41]). Numerically, attractors were obtained using multi-purpose numerical codes in idealized 2D trapezoidal domains

(Grisouard et al. [36]), in 3D parabolic channel domains (Drijfhout & Maas [19]) or in geometries mimicking realistically the Luzon Strait in the South China Sea (Tang & Peacock [108]; Echeverri et al. [26]). Because of the interest in the dynamics of the Earth's liquid outer core and of stellar interiors, special attention has been devoted to wave attractors in spherical shells, where they are relevant to tidal dissipation and where they are resolved using spectral codes (e.g. Dintrans et al. [18]; Tilgner [109]; Rieutord et al. [98]).

But the actual relevance of internal wave attractors to real lakes, seas, oceans, atmospheres, the Earth's outer core, or planets and stars is unclear at present. Many factors may after all 'dilute' the ideal setting, and the evidence from direct observations is inconclusive or contradictory. Field observations in the small, 1 km wide stratified lake Mystic, show that the horizontal velocity reaches its maxima at the sloping sides of the lake. This suggests that internal waves are steered towards a wave attractor instead of taking the shape of a seiche, a sloshing mode which would have its velocity maximum near the center (Fricker & Nepf [31]). Earlier lake observations revealed the dominance of high-wavenumber vertical modes, indicative of the presence of the small-scales associated with an attractor (LaZerte [57]). The nonuniform stratification and presence of sheared background currents, all affecting internal wave ray paths, have been held responsible for the apparent absence of an attractor in the much larger Faroe-Shetland Channel (Gerkema & van Haren [32]). The absence of an attractor may, however, also be due to a mismatch between aspect ratio and the ratio of wave and stratification frequencies. Recent satellite observations of internal solitary waves suggest that wave attractors might actually have served as the amplification mechanism required to explain the enigmatic appearance of internal solitary waves from weak surface tides over a particular 80 km stretch of the Red Sea (da Silva et al. [14]). This seems to emphasize that higher spatial resolution of periodic internal wave fields is needed in in situ measurements.

Here we concentrate on an unsolved 'academic aspect', addressing the response of a uniformly stratified 2D fluid to an initial perturbation in a basin whose shape breaks the reflection symmetry of internal gravity waves. The ansatz of a time-periodic, single frequency (monochromatic) solution to the linearized internal gravity wave equations yields a wave equation in space with Dirichlet boundary conditions. This makes the problem quite unusual, as it is ill-posed due to nonuniqueness. The problem allows for weak solutions that can be solved using the method of characteristics or through a regularization technique (Swart et al. [106]). Via the method of characteristics one can study the limit behavior of reflecting rays in bounded domains. The most generic asymptotic solution is an attractor, which is a finite closed orbit of rays within the domain. The particular structure of internal gravity wave attractors in a tilted square domain depends on: the rotation angle of the square $\theta$, the wave frequency $\omega$ and the stratification frequency $N_f$. A family of wave attractors is characterized by the number of reflections of a member-attractor from the boundary. By symmetry considerations, an attractor must reflect an equal number $n$ times with the top and bottom domain boundaries, and an equal number $m$ times with the left and right boundaries. Such an attractor is called an $(n, m)$-attractor. Figure 2.1 shows a discrete sample of the attractor geometries from the infinite classes of (1,1)- and (1,3)-attractors in a tilted square domain (see Section

2.2).

Due to the ill-posedness of the monochromatic wave problem, we are motivated to study the initial value problem for internal gravity waves in a confined region. Alternatively, one could introduce viscosity, which regularizes the monochromatic wave problem, allowing for its approximate analytical solution (Ogilvie [90]). Lighthill [68] considered the initial value problem for the evolution of a localized disturbance in an unbounded domain, deriving the dispersion relation and noting that vortical structures remain stationary after internal gravity waves have propagated away horizontally. In this chapter we study internal waves in a stratified fluid filling a domain with solid walls, so that wave motion is trapped inside. We consider the simplest case that admits wave attractors: perturbations to a linearly stratified inviscid fluid, either freely evolving or parametrically excited. To guarantee that viscous effects play no role—not even implicitly via "numerical diffusion"—we construct a numerical discretization that conserves total energy and symmetry in the absence of forcing and study two idealized theoretical configurations: freely evolving (i.e. unforced) flow, and parametrically excited flow. We proceed with a normal mode analysis of the discrete model. For the freely evolving case, we analyze the unforced initial boundary value problem, to show how linear dynamics is partitioned into normal modes for different classes of initial conditions. Figure 2.2 illustrates the free evolution from Fourier modes with wave numbers $(1, 1)$ and $(1, 3)$, respectively. Evident in the plots at later times, we observe structures reminiscent of the full class of $(1, 1)$- and $(1, 3)$-attractors, suggesting a relationship between the Fourier modes and attractor geometries, for which we give some motivation. For the parametrically excited case, the normal mode analysis reveals that the flow may be decomposed into independent Mathieu equations, and that those modes whose associated frequencies lie within the resonance zones (Arnold tongues) will be amplified, forming a wave attractor.

It is important to note that the existence of a complete normal mode decomposition for the discretized model contrasts sharply with the continuum model, for which the eigenspectrum is continuous and no such decomposition exists (Maas [70]). The continuous spectrum for the continuum model actually implies the existence of an uncountable infinity of time-periodic solutions, corresponding to the arbitrary definition of the boundary condition on the fundamental intervals, which we discuss. For the discretized system, the finite basis of normal modes are precisely the time-periodic solutions. The complete normal mode decomposition for the discrete model is also non-robust with respect to viscous perturbation of the system. For the forced system with viscosity, the normal mode basis becomes time dependent, meaning the solution cannot be decomposed into scalar problems.

The chapter is organized as follows: In Section 2.2 we recall the 2D linear hydrostatic inviscid Euler-Boussinesq equations which govern internal gravity waves in stratified fluids, discuss monochromatic solutions in a tilted square domain, and review the Hamiltonian structure. In Section 2.3 we describe a structure-preserving finite difference discretization on the tilted square and present the normal mode analysis of the discretized model in the unforced and forced cases. Using the symmetries of the discrete differential operators we show that in both cases the dynamics may be projected onto an invariant basis of normal modes, such that they entirely

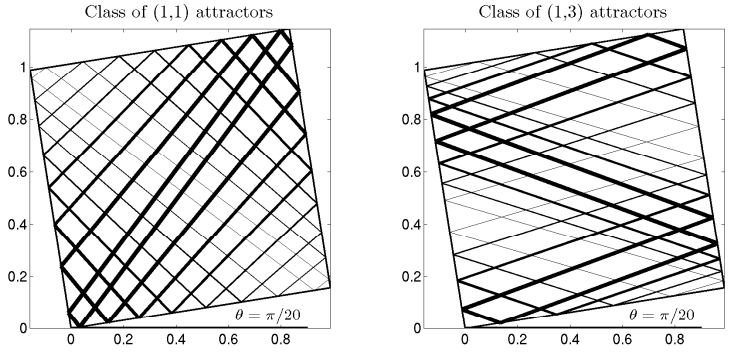



Figure 2.1: Limit cycle wave attractors corresponding to a discrete set of frequencies from the respective continuum ranges. Different line thicknesses correspond to distinct wave attractors. Left: class of $(1, 1)$ attractors. Right: class of $(1, 3)$ attractors.

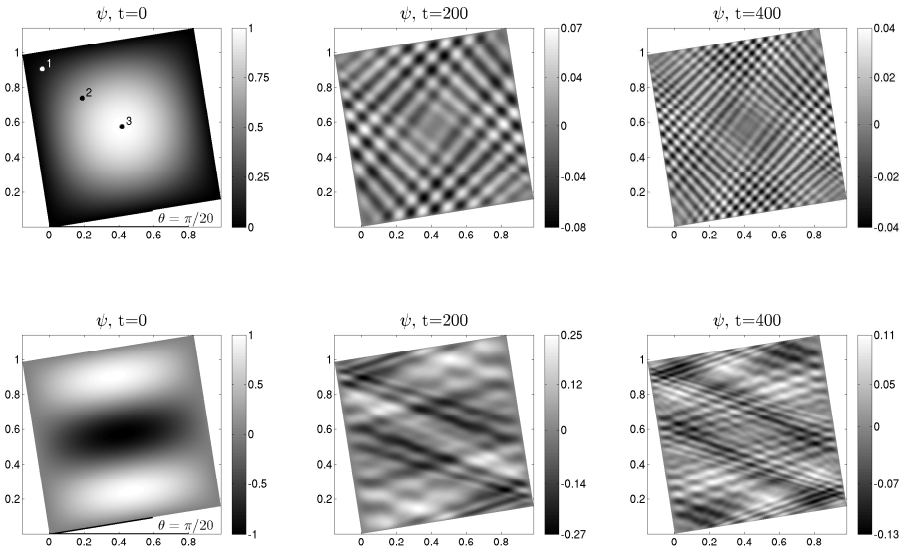



Figure 2.2: Evolution of the stream function in time from two distinct Fourier mode initial conditions.

decompose into independent scalar problems: harmonic oscillators in the unforced case or Mathieu equations in the forced case. In Section 2.4 we present numerical experiments of the unforced and forced models. We observe that an $(n, m)$

Fourier mode initial condition projects mostly onto the range of the associated $(n, m)$-attractor, explaining the similarities of Figures 2.1 and 2.2. For the forced model we observe that if the initial condition has a nontrivial projection onto normal modes with amplified Mathieu dynamics, a wave attractor will emerge. Conclusions are summarized in Section 2.5.

## 2.2 Euler-Boussinesq equations

### 2.2.1 Internal gravity wave equations

We consider a vertical slice domain $\mathcal{D} \subset \mathbf{R}^2$ with boundary $\partial\mathcal{D}$ and Cartesian coordinates $\boldsymbol{x} = (x, z)$, where z is directed antiparallel to the direction of gravity, $\mathbf{g}$. We decompose the fluid density field and the pressure field as follows:

$$\rho(x, z, t) = \rho_0 + \bar{\rho}(z) + \rho'(x, z, t), \qquad p(x, z, t) = \bar{p}(z) + p'(x, z, t),$$

where $\rho_0$ is an average constant mean density and $\bar{\rho}(z)$ is a mean static density stratification, i.e. a monotonically decreasing function of $z$. The sum $\rho_0 + \bar{\rho}(z)$ defines a stable background density field in hydrostatic balance with the pressure field $\bar{p}(z)$:

$$\frac{\partial \bar{p}}{\partial z} = -g(\rho_0 + \bar{\rho}(z)),$$

where $g$ is the gravitational acceleration. The quantities $\rho'(x, z, t)$ and $p'(x, z, t)$ are small amplitude perturbations about the (steady state) background density and pressure fields.

In geophysical and astrophysical fluid dynamics it is common to treat the density field distinctly, defining both an 'inertial mass' and a 'gravitational mass'. The Boussinesq approximation consists of assuming a constant density value $\rho_0$ for the inertial mass in the momentum equation (from which the density may be consequently removed), while maintaining the full density $\rho$ for the gravitational mass. We enforce the inequality $|\rho'| \ll |\bar{\rho}(z)| \ll \rho_0$ to justify the Boussinesq approximation. Such flows are termed 'buoyancy-driven'. The background stratification defines a stratification frequency, $N_f$, (Brunt-Väisälä frequency), where $N_f^2 = -g\rho_0^{-1}\mathrm{d}\bar{\rho}/\mathrm{d}z$. In the following we assume that $N_f$ is a constant, i.e. the fluid is linearly stratified in the background density.

Wave focusing occurs when a boundary of the domain is inclined with respect to gravity. For this reason we assume that the coordinate system is rotated through an angle $0 \leq \theta \leq \pi/4$. With the above considerations in mind, the inviscid linear Euler-Boussinesq equations describing the propagation of perturbations in this rotated frame read:

$$\frac{\partial \boldsymbol{u}}{\partial t} = -\nabla \hat{p} + b\hat{\boldsymbol{k}}(\theta), \tag{2.1}$$

$$\frac{\partial b}{\partial t} = -N_f^2 \boldsymbol{u} \cdot \hat{\boldsymbol{k}}(\theta), \tag{2.2}$$

$$\nabla \cdot \boldsymbol{u} = 0, \tag{2.3}$$

$$\boldsymbol{u} \cdot \hat{\boldsymbol{n}} = 0 \quad \text{on} \quad \partial\mathcal{D}, \tag{2.4}$$

where $\boldsymbol{u} = (u, w)$ is a velocity field in the $x$ and $z$ direction respectively (now *tilted* relative to the original direction), $\hat{p} = \rho_0^{-1} p'$ is scaled pressure with respect to the mean constant density, $b = -g\rho'\rho_0^{-1}$ is the buoyancy, $\hat{\boldsymbol{k}}(\theta) = (\sin\theta, \cos\theta)$ is the unit vector in the direction opposite to gravity and $\hat{\boldsymbol{n}}$ is the unit outward normal to the boundary $\partial\mathcal{D}$.

In two-dimensions it is convenient to consider the stream function formulation of the Euler-Boussinesq equations (2.1)–(2.4). The divergence-free condition (2.3) allows us to define a stream function $\psi$ on $\mathcal{D}$ such that

$$u = -\frac{\partial\psi}{\partial z}, \qquad w = \frac{\partial\psi}{\partial x}.$$

By taking the curl of the momentum equations (2.1) we eliminate the pressure from (2.1), obtaining the 2D linear inviscid Euler-Boussinesq equations in stream function formulation:

$$\frac{\partial q}{\partial t} = -\frac{\partial b}{\partial x}\cos\theta + \frac{\partial b}{\partial z}\sin\theta, \tag{2.5}$$

$$\frac{\partial b}{\partial t} = -N_f^2\left(\frac{\partial\psi}{\partial x}\cos\theta - \frac{\partial\psi}{\partial z}\sin\theta\right), \tag{2.6}$$

$$q = -\Delta\psi, \tag{2.7}$$

$$\psi = 0 \quad \text{on} \quad \partial\mathcal{D}, \tag{2.8}$$

where $q = \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}$ is vorticity.

The model (2.5)–(2.8) is a system of partial differential equations that conserves total energy:

$$\mathcal{H} = \frac{1}{2}\int_{\mathcal{D}}\left(\nabla\psi\cdot\nabla\psi + \frac{1}{N_f^2}b^2\right)\,\mathrm{d}\boldsymbol{x}, \tag{2.9}$$

equal to the sum of kinetic and potential energies.

### 2.2.2 Forcing

Wave attractors are generated by periodically forcing a stratified fluid in a domain with inclined boundaries. In the ocean, the forcing is primarily tidal forcing. In laboratory experiments (Maas et al. [72]; Lam & Maas [56]), wave attractors were generated by vertically oscillating a container with a sloping wall. To incorporate such *parametric excitation* (McEwan & Robinson [80]) equation (2.5) is modified by multiplication with a time dependent function $\alpha(t)$ to obtain:

$$\frac{\partial q}{\partial t} = \alpha(t)\left(-\frac{\partial b}{\partial x}\cos\theta + \frac{\partial b}{\partial z}\sin\theta\right).$$

An alternative approach is *external excitation*, e.g. a horizontal oscillation of the container, for which time dependent terms may be added to (2.5) and (2.6) (Ogilvie [90]), or by means of boundary forcing (Grisouard et al. [36]).

Vertical oscillation of the container can be viewed as time-dependent modulation of the gravitational parameter $g$, which originally enters the momentum equation,

and should thus be present only in the vorticity equation (2.5). Hence, we can realize this kind of forcing as parametric excitation with

$$\alpha(t) = 1 - \epsilon \cos(2\omega t),$$

where $\epsilon$ is a positive constant smaller than one and $2\omega$ is the forcing frequency.

### 2.2.3 Dispersion properties of internal gravity waves

Consider a time periodic solution

$$\psi(x, z, t) = \Psi(x, z)e^{-i\omega t}, \qquad b(x, z, t) = B(x, z)e^{-i\omega t}.$$

Substituting the above ansatz into (2.5)–(2.7), eliminating $B$ and taking $\theta = 0$ without loss of generality yields

$$\frac{\partial^2 \Psi}{\partial z^2} - \frac{(N_f^2 - \omega^2)}{\omega^2} \frac{\partial^2 \Psi}{\partial x^2} = 0, \tag{2.10}$$

which is recognized as a wave equation when $\omega^2 < N_f^2$ for the scalar state variable $\Psi$. In other words, internal gravity waves are spatially governed by the wave equation. Substituting the plane wave

$$\Psi(x, z) = a \exp(i(\kappa_x x + \kappa_z z))$$

into (2.10), where $a$ is the amplitude and $\kappa_x$ and $\kappa_z$ are wave numbers, yields the dispersion relation

$$\omega^2 = N_f^2 \frac{\kappa_x^2}{\kappa_x^2 + \kappa_z^2} = N_f^2 \cos^2 \phi, \tag{2.11}$$

the last equality of which follows from the polar coordinate description of the wave number vector $\boldsymbol{\kappa} = |\boldsymbol{\kappa}|(\cos\phi, \sin\phi)$ where $|\boldsymbol{\kappa}|$ is the wave number magnitude and $\phi$ its direction. Hence, $\omega^2 \leq {N_f}^2$ and the frequencies of internal gravity waves are bounded by the stratification frequency $N_f$. It is also apparent that the wave frequency is independent of the wave number magnitude and depends only on its angle $\phi$. Consequently an incident wave retains its propagation direction upon reflection from a plane surface independent of the slope of the surface, leading to monoclinic (single-angled) waves. A wave does, in general, change its wavelength and can become focused or defocused upon reflection from an inclined boundary. It is well known that the wave phase travels in the phase velocity direction $\boldsymbol{c_p} = \omega\boldsymbol{\kappa}/|\boldsymbol{\kappa}|^2$ and wave packet energy is transported by the group velocity $\boldsymbol{c_g} = \nabla_{\boldsymbol{\kappa}}\omega$, Whitham [113]. The internal wave group velocity vector $\boldsymbol{c_g}$ and phase velocity vector $\boldsymbol{c_p}$ are mutually perpendicular, i.e. $\boldsymbol{c_g} \cdot \boldsymbol{c_p} = 0$. Hence internal waves propagate energy parallel to the wave crests and troughs (i.e. along these).

### 2.2.4 Monochromatic wave solutions in a tilted square

The wave equation (2.10) with Dirichlet boundary conditions (2.8) is formally an ill-posed problem (Swart et al. [106]). One not only finds a trivial solution $\psi \equiv 0$,

but there exist infinitely many solutions. For example, the hyperbolic wave equation (2.10) can be solved on a non-inclined ($\theta = 0$) rectangular domain $(x, z) \in [0, 1] \times [0, \ell]$ by separation of variables. The function

$$\Psi = A_{n,m} \sin(n\pi x) \sin(m\pi z/\ell)$$

satisfies the hyperbolic equation (2.10) and boundary condition (2.8) provided that

$$\ell = \sqrt{\frac{\omega^2}{N_f^2 - \omega^2}} \frac{m}{n}. \tag{2.12}$$

Replacing integer $(n, m)$ in (2.12) by $(jn, jm)$ leaves $\ell$ unchanged, and for integer $j$, $\Psi$ still vanishes at the boundaries. In this noninclined case there is a denumerable infinite set of solutions to the wave equation (2.10); in the inclined case this set is not denumerable, resulting in the ill-posedness.

The general solution of the wave equation (2.10) is given by

$$\Psi(x, z) = f(x - \gamma z) - g(x + \gamma z), \qquad \gamma = \sqrt{\frac{N_f^2 - \omega^2}{\omega^2}},$$

for arbitrary functions $f$ and $g$. Hence the function $g$ is constant along a characteristic line $x + \gamma z = const.$, and likewise $f$ is constant along lines $x - \gamma z = const.$ Furthermore, the Dirichlet boundary condition, $\Psi = 0$, implies that $f \equiv g$ on the boundary. Therefore, from any point $p$ in the domain, one can define an orbit, consisting of a characteristic passing through $p$ and the infinite sequence of successive reflections of that characteristic in both forward and backward orientation upon which $f$ and $g$ are alternately constant. Such a sequence of characteristics will be referred to as a *characteristic orbit*. Two characteristic orbits intersect at each point $p$ in the interior of $\mathcal{D}$, and the difference $f - g$ determines the stream function at $p$. One can follow characteristic orbits that intersect at $p$ until they reach a boundary segment upon which the function $f = g$. The problem of determining a well-posed monochromatic solution is reduced to that of identifying a minimal set of distinct intervals, the so called fundamental intervals, on the boundary where the functions $f$ and $g$ may be prescribed (see Maas & Lam [73]).

In this chapter we will study internal waves in a tilted square domain. In the tilted unit square the topology of a complete characteristic orbit passing through a point depends on the angle of tilt $\theta$ and the ratio of wave frequency to stratification frequency $\omega/N_f$. In the *subcritical case* all characteristic orbits asymptotically approach diagonally opposite corners of the square. This occurs when the characteristic slopes $\pm\gamma$ are both either larger or smaller than the inclination of both horizontal and vertical boundaries. In the *supercritical case* one can distinguish an additional three types of limit behavior: periodic, ergodic and limit cycle orbits (John [49]; Kopecz [54]). In the periodic case all characteristic orbits reflect from the boundary at a finite number of points, the fundamental intervals collapse onto one another, and the characteristic orbit through every point is periodic. In the ergodic case, the characteristic orbit through any point passes arbitrarily close to every other point in the domain, the fundamental interval shrinks to a single point,

and the stream function then necessarily vanishes, implying no flow. However, the most generic case of limit behavior of the characteristic orbits is an attractor or limit cycle, i.e. one or more distinct periodic orbits that attracts a neighborhood of itself. Such attractors are characterized by the number of boundary reflections from the horizontal and vertical boundaries. Considering the symmetry of the top and bottom boundary and of the two side boundaries, we denote by $(n, m)$ an attractor having $n$ reflections from the boundary on the upper side of the square and $m$ reflections from the left side of the square. The overall number of reflections with the boundary $(2n + 2m)$ is called the attractor's period. In the unit square domain all attractors are globally attracting.

The choice of the fundamental intervals on the boundary and the functions prescribed on them is not unique. In the subcritical case it is sufficient to prescribe only one interval between two successive characteristic reflections from the boundary. In the ergodic case the solution may be prescribed at only one point on the boundary yielding the trivial solution $\psi \equiv 0$ of the wave equation (2.10) due to the zero Dirichlet boundary conditions (2.8). For the periodic and attractor cases one must prescribe one or two intervals on one of the square's boundaries, respectively. For a complete discussion see Maas & Lam [73].

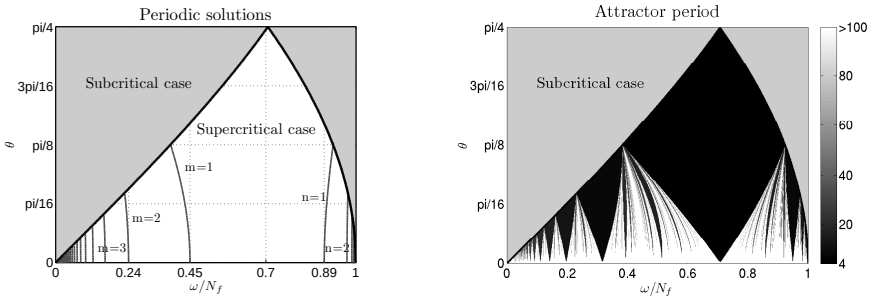
Let us take a closer look at periodic solutions and limit cycles. The experimental variables are the wave frequency $\omega$, stratification frequency $N_f$ and rotation angle of the square $\theta$. In the periodic solution regime, all orbits correspond to odd-even pairs $(2n, 2m+1)$ or $(2n+1, 2m)$. But the periodic regime is non-robust with respect to perturbations in domain geometry. In the tilted square domain these solutions occur only for a discrete set of frequencies. In contrast the limit cycle attractors persist over a continuous range of frequencies, hence are robust with respect to frequency perturbations. In the simplest periodic case the characteristic orbit emanating from, say, the lower left corner of the square will precisely intersect the lower right corner after making $n$ successive reflections from the top of the square, or will intersect the upper left corner after $m$ successive reflections from the right side of the square. In both such situations we have analytic expressions relating the wave frequency $\omega$, stratification frequency $N_f$ and rotation angle of the square $\theta$:

$$\cot\left(\theta + \tan^{-1}\sqrt{\frac{\omega^2}{N_f^2 - \omega^2}}\right) - \cot\left(\theta - \tan^{-1}\sqrt{\frac{\omega^2}{N_f^2 - \omega^2}}\right) = \frac{1}{n},$$

$$\tan\left(\theta + \tan^{-1}\sqrt{\frac{\omega^2}{N_f^2 - \omega^2}}\right) - \tan\left(\theta - \tan^{-1}\sqrt{\frac{\omega^2}{N_f^2 - \omega^2}}\right) = \frac{1}{m},$$

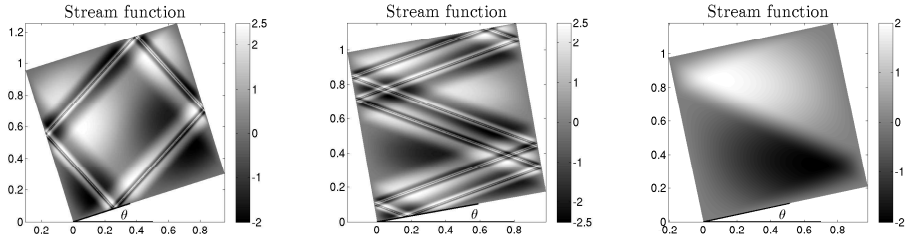
respectively. Hence these periodic solutions are indicated as $(2n, 1)$ and $(1, 2m)$ with periods $2(2n + 1)$ and $2(2m + 1)$, respectively. Similar periodic solutions can be computed when the characteristic orbits have multiple reflections from both the left and top boundaries, and geometries $(2n, 2m + 1)$ or $(2n + 1, 2m)$.

Figure 2.3 illustrates the parameter space $\omega/N_f$ versus $\theta$. The bold line separates subcritical and supercritical regimes. Within the supercritical region of Figure 2.3(a), we indicate the loci of parameter values corresponding to periodic solutions of the classes $(2n, 1)$ and $(1, 2m)$. Note that for a given rotation angle $\theta$, the periodic solutions correspond to discrete values of $\omega/N_f$. Limit cycle solutions are




(a) (b)

Figure 2.3: Parameter space for monochromatic solutions. Left: loci in parameter space corresponding to periodic solutions $(2n, 1)$ and $(1, 2m)$. Right: limit cycle attractor period, indicated by color.




(a) (b) (c)

Figure 2.4: Monochromatic stream function solutions. Left: within the (1,1) attractor frequency range ($\omega/N_f = 0.74$, $\theta = 7\pi/72$). Middle: within the (1,3) attractor frequency range ($\omega/N_f = 0.34$, $\theta = \pi/18$). Right: the unique $(1, 2)$ periodic solution ($\omega/N_f = 0.43$, $\theta = \pi/15$).

indicated in Figure 2.3(b), where the color denotes the period of the attractor. Periodic solutions (Figure 2.3(a)) are found where the attractor period (Figure 2.3(b)) approaches infinity.

Figure 2.4 shows solutions of the monochromatic wave equation (2.10) for the $(1, 1)$ and $(1, 3)$ attractor cases and for the $(1, 2)$ periodic case, for specific values of $\theta$ and $\omega/N_f$. In Figures 2.4(a) and 2.4(b) we show two typical members from the respective continuum ranges of limit cycle solutions. In both cases one can observe a self-similar structure approaching the attractor. The solutions were constructed using the method of characteristics; on the fundamental intervals we prescribe two cosines with an offset at the chosen intervals. For a square-shaped attractor in a trapezoidal geometry, a free wave solution possessing a logarithmic self similar Fourier spectrum was computed analytically (Maas [71]).

## 2.3 Numerical discretization and linear analysis

In this section we describe our discrete model equations and show that in the special case of linear inviscid flow, the dynamics decouples into scalar oscillators.

### 2.3.1 Fourier analysis of the continuum model, non-tilted

For a non-tilted square domain ($\theta = 0$), the initial boundary value problem for the linear Euler-Boussinesq equations (2.5)–(2.8) with initial conditions $\psi_0(x,z)$ and $b_0(x,z)$ and zero Dirichlet boundary conditions (2.8) can be solved analytically using separation of variables. The solution is

$$\psi(x,z,t) = \sum_{n,m=1}^{\infty} \psi_{n,m}(x,z)\frac{\mathrm{d}}{\mathrm{d}t}T_{n,m}(t), \tag{2.13}$$

$$b(x,z,t) = -N_f^2 \sum_{n,m=1}^{\infty} \frac{\partial}{\partial x}\psi_{n,m}(x,z)T_{n,m}(t), \tag{2.14}$$

where $\psi_{n,m}(x,z) = \sin(n\pi x)\sin(m\pi z)$ are Fourier modes on the unit square, i.e. the eigenfunctions of the operators $\frac{\partial^2}{\partial x^2}$ and $\frac{\partial^2}{\partial z^2}$ under the given boundary conditions, and $T_{n,m}$ is a solution to the simple harmonic oscillator equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}T_{n,m} = -\omega_{n,m}^2 T_{n,m}, \qquad \omega_{n,m}^2 = N_f^2\frac{n^2}{n^2+m^2}, \tag{2.15}$$

with the frequencies given by the dispersion relation (2.11).

The total energy functional (2.9) of the general solution in the form (2.13)–(2.14) is

$$\mathcal{H} = \frac{\pi}{8}\sum_{n,m=1}^{\infty}\left[(n^2+m^2)\left(\frac{\mathrm{d}}{\mathrm{d}t}T_{n,m}\right)^2 + N_f^2 n^2 T_{n,m}^2\right] = \sum_{n,m=1}^{\infty}\mathcal{H}_{n,m},$$

where for each $(n,m)$, the term in square brackets, $\mathcal{H}_{nm}$, is the independently conserved Hamiltonian of (2.15). Note that there is no coupling between wave numbers. The initial conditions may be projected onto the Fourier modes, but each mode evolves independently, and there is no energy exchange between modes.

The situation for $\theta \neq 0$ is very different. The initial boundary value problem (2.5)–(2.8) cannot be solved analytically by the method of separation of variables as it was done above. The eigenfunctions in the tilted case correspond to the ill-posed solutions of (2.10), and have no simple representation. However, as we show in the next section, the numerical discretization does admit a normal mode analysis.

### 2.3.2 Energy conserving numerical discretization and analysis

Making use of the Hamiltonian structure of (2.5)–(2.8), we construct in Appendix 2.A an energy preserving numerical discretization. Discretizing in space while leav-

ing time continuous yields the following system of linear ordinary differential equations (cf. (2.A.9)–(2.A.11)):

$$-L\frac{\mathrm{d}\boldsymbol{\psi}}{\mathrm{d}t} = \alpha(t)\left(D_x^T M_z \boldsymbol{b}\cos\theta - D_z^T M_x \boldsymbol{b}\sin\theta\right), \tag{2.16}$$

$$\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}t} = -N_f^2\left(M_z^T D_x \boldsymbol{\psi}\cos\theta - M_x^T D_z \boldsymbol{\psi}\sin\theta\right), \tag{2.17}$$

where $\boldsymbol{\psi} \in \mathbf{R}^M$ and $\boldsymbol{b} \in \mathbf{R}^N$, $M < N$, are vectors containing the values of $\psi$ and $b$ at (staggered) grid positions. The finite difference matrices $M_x$, $M_z$, $D_x$, $D_z$ and $L$, defined in Appendix 2.A.1, represent discretized mean ($M_*$), difference ($D_*$) and Laplacian ($L$) operators and superscript T denotes the transpose. Here we introduced the factor $\alpha(t)$, that allows us to add forcing by means of parametric excitation. Introducing the matrix $K = D_x^T M_z \cos\theta - D_z^T M_x \sin\theta$, this system can be written in matrix form

$$\begin{bmatrix} -L & 0 \\ 0 & I \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix} = \begin{bmatrix} 0 & \alpha(t)K \\ -N_f^2 K^T & 0 \end{bmatrix}\begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix}. \tag{2.18}$$

By construction, when forcing is absent ($\alpha \equiv 1$) the discretization possesses a first integral, the discrete Hamiltonian $H$ (2.A.8), which approximates the total energy (2.9), i.e.

$$H = \frac{1}{2}\left(-\boldsymbol{\psi}^T L\boldsymbol{\psi} + \frac{1}{N_f^2}\boldsymbol{b}^T\boldsymbol{b}\right)\Delta x\,\Delta z. \tag{2.19}$$

In Appendix 2.B we derive the normal mode bases $X = (X_1, \ldots, X_M)$ and $Y = (Y_1, \ldots, Y_N)$, in which $\boldsymbol{\psi}$ and $\boldsymbol{b}$ are expressed as, cf. (2.B.22),

$$\boldsymbol{\psi} = X\tilde{\boldsymbol{\psi}}, \qquad \boldsymbol{b} = Y\tilde{\boldsymbol{b}}.$$

In the new basis, the system (2.18) decouples into $M$ second order problems:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{\psi}_i = -\alpha(t)\omega_i^2\tilde{\psi}_i + \dot{\alpha}(t)\omega_i\tilde{b}_i, \tag{2.20}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{b}_i = -\alpha(t)\omega_i^2\tilde{b}_i, \tag{2.21}$$

for $i = 1, \ldots, M$, plus the trivial dynamics $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{b}_i = 0$, $i = M+1, \ldots, N$.

When forcing is absent, $\alpha(t) \equiv 1$, the dynamics further decouples into $2M$ independent harmonic oscillators

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{\psi}_i = -\omega_i^2\tilde{\psi}_i, \qquad \frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{b}_i = -\omega_i^2\tilde{b}_i, \qquad i = 1, \ldots, M.$$

In particular the total energy can be expressed as the sum of the harmonic oscillator energies

$$H = \sum_{i=1}^{M} H_i^{\psi} + H_i^b, \quad H_i^{\psi} = \frac{1}{2}\left[\left(\frac{\mathrm{d}\tilde{\psi}_i}{\mathrm{d}t}\right)^2 + \omega_i^2\tilde{\psi}_i^2\right], \quad H_i^b = \frac{1}{2}\left[\left(\frac{\mathrm{d}\tilde{b}_i}{\mathrm{d}t}\right)^2 + \omega_i^2\tilde{b}_i^2\right],$$

each of which is a conserved quantity.

**Remark.** In Section 2.2 we saw that there are infinitely many monochromatic wave solutions to the linearized Euler-Boussinesq equations, corresponding to an arbitrary specification of the solution on a fundamental interval. For the discretized equations, of course, there can be only a finite number of periodic solutions, each corresponding to a normal mode of the discretization matrix. This situation is analogous to the case of the advection equation $\rho_t + u\rho_x = 0$ on a periodic domain, for which any initial condition $\rho(x,0) = f(x)$ is periodic in time. Upon numerical discretization of this equation, the dispersion relation is altered, an arbitrary initial condition may be expanded in normal modes, and each of these evolves with a different phase speed, causing artificial dispersion. Only the (finite denumerable) normal modes themselves are periodic.

When parametric forcing is present in (2.21), i.e. $\alpha(t) = 1 - \epsilon\cos 2\omega t$, the buoyancy modes evolve independently according to the *Mathieu equation*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{b}_i = -(1 - \epsilon\cos(2\omega t))\omega_i^2\tilde{b}_i. \tag{2.22}$$

The Mathieu equation supports resonance zones in parameter space for which the solution grows unbounded in magnitude, as well as stable (non-resonant) zones for which the solution remains bounded for all time. The first and most important instability region originates at the subharmonic frequency $\omega$ of the driving frequency $2\omega$, (see Arnold [3]).

### 2.3.3 Dynamics of the Mathieu equation

Rescaling time with respect to the stratification frequency $N_f$, i.e. $t' = N_f t$, in equation (2.22) yields, dropping primes,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{b}_i = -\left(1 - \epsilon\cos\left(2\frac{\omega}{N_f}t\right)\right)\frac{\omega_i^2}{N_f^2}\tilde{b}_i, \tag{2.23}$$

where $\omega_i^2/N_f^2 \leq 1$ from the dispersion relation. For given value of the (normalized) first subharmonic forcing frequency $|\omega/N_f| \leq 1$ we are interested in knowing for which normal mode frequencies $\omega_i/N_f$ and forcing amplitude $\epsilon$ equation (2.23) and equation (2.22) support resonances.

Introducing a second time transformation, $t' = \omega N_f^{-1} t$, we write the scalar Mathieu equations (2.23) in the general form

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\beta + (a - 2q\cos(2t))\beta = 0, \tag{2.24}$$

where $\beta = \tilde{b}_i$, $a = \omega_i^2/\omega^2 \leq N_f^2/\omega^2$ and $q = \frac{\epsilon}{2}\omega_i^2/\omega^2 \leq \frac{\epsilon}{2}N_f^2/\omega^2$ for a given normal mode $i$. According to the Floquet multiplier theorem, the Mathieu equation for fixed $a$ and $q$ admits a complex valued general solution of the form

$$\beta(t) = c_1 e^{\mu t}P(a,q,t) + c_2 e^{-\mu t}P(a,q,-t),$$

where $\mu \neq 0$ is a complex Floquet exponent and $P(a, q, t)$ is a complex valued, $\pi$-periodic, special function, i.e. $P(a, q, t + \pi) = P(a, q, t)$. If $\mathrm{Re}\,\mu = 0$, the solution $\beta(t)$ is bounded for all time. If $\mathrm{Re}\,\mu \neq 0$, the amplitude of the oscillations grows exponentially. For the degenerate case $\mu = 0$, the solutions are linearly dependent and the amplitude grows linearly in time.

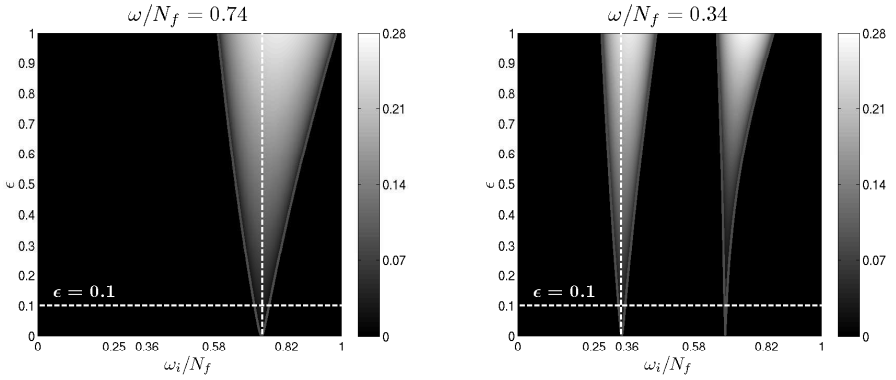
To determine the Floquet exponent $\mu$ we note that taking initial conditions $\beta(0) = 1$, $\frac{\mathrm{d}\beta}{\mathrm{d}t}\big|_{t=0} = 0$, one finds $c_1 = c_2 = (2P(a, q, 0))^{-1}$, hence the solution at time $t = \pi$ is

$$\beta(\pi) = \cosh \mu\pi.$$

Therefore $\mu$ can be estimated by solving (2.24) numerically on the interval $[0, \pi]$. For a given forcing $\omega/N_f$, we solve for $\mu$ numerically using the Störmer-Verlet method (Hairer et al. [37]) over a discrete set of values $\epsilon \in [0, 1]$ and $\omega_i/N_f \in [0, 1]$.

Our goal is to investigate the emergence of the two internal wave attractors presented in Section 2.2 by use of the parametric excitation mechanism described above. We expect that after an initial transient phase, the solution will be dominated by those normal modes having positive Floquet exponents. We fix $\epsilon = 0.1$ and choose forcing frequencies $2\omega/N_f$ whose subharmonics excite the patterns in Figure 2.4, i.e. we choose $\omega/N_f = 0.74$ or $\omega/N_f = 0.34$, respectively. In Figure 2.5 we plot the real part of the Floquet exponent $\mu$ as a function of normal mode frequency $\omega_i/N_f \in [0, 1]$ (regarding $\omega_i/N_f$ as a continuous variable). For these two cases we obtain the instability tongues shown in Figures 2.5(a) and 2.5(b), respectively. Figure 2.5(a) shows the real part of the Floquet exponent $\mu$ for subharmonic forcing frequency $\omega/N_f = 0.74$. The resonant instability tongue originates at $\omega_i/N_f = 0.74$, and superharmonic resonances ($n\omega/N_f$, $n = 2, 3, \dots$) are absent because they fall outside the admissible range of normal mode frequencies. Figure 2.5(b) shows $\mathrm{Re}\,\mu$ for subharmonic forcing frequency $\omega/N_f = 0.34$. The first resonant instability tongue then originates at $\omega_i/N_f = 0.34$, and also the first superharmonic resonance at $\omega_i/N_f = 2\omega/N_f = 0.68$ falls within the admissible range of normal mode frequencies. For a given value of subharmonic forcing frequency $\omega/N_f$, the rotation angle $\theta \in [0, \pi/4]$ determines the type of limit behavior observed, e.g. an attractor or a periodic solution, see Figure 2.3.

Since the forced internal wave equations (2.16)–(2.17) can be decomposed into the Mathieu type equations (2.20)–(2.21), the theory of Mathieu equations suggests that depending on the values of the Floquet exponent there will be resonant normal modes which will grow exponentially in time and there will be other modes which will stay bounded. The presence of resonant normal modes is dependent on the initial conditions. If a particular initial condition is such that its projection onto normal modes has no components within resonant zones of the Mathieu equation, then the solution of the forced linear internal wave equations (2.16)–(2.17) will stay bounded for all times. Hence the choice of initial conditions for computations is not arbitrary. The analysis in Section 2.4.1 of the system's response to different initial conditions in the unforced, undamped linear case suggests that the natural choice for finding $(1, 1)$ and $(1, 3)$ attractors would be initial conditions $\psi_{1,1}$ and $\psi_{1,3}$, respectively. This implies that there will be resonant normal modes.




(a) (b)

Figure 2.5: Instability tongues of the Mathieu equation, color denotes the magnitude of Floquet exponent $\mathrm{Re}\,\mu$, as a function of normal mode frequency $\omega_i/N_f$ for different forcing amplitudes $\epsilon$. Left: subharmonic forcing frequency $\omega/N_f = 0.74$, one instability tongue in the computation of the $(1,1)$ attractor. Right: subharmonic forcing frequency $\omega/N_f = 0.34$, two instability tongues in the computation of the $(1,3)$ attractor. The vertical and horizontal dashed lines indicate forcing frequencies and amplitudes respectively.

## 2.4 Numerical experiments

### 2.4.1 Freely evolving flow

Armed with the theory of internal gravity wave attractors in a tilted square from Section 2.2 and the structure preserving discretization of the Euler-Boussinesq equations in the stream function formulation from Section 2.3 we study the initial boundary value problem. Since we consider the inviscid equations, the system does not depend on spatial scales and time can be rescaled with respect to stratification frequency $N_f$ to cast the system in dimensionless form. As we will see in the following, the response of the system will depend on tilt angle $\theta$ and on the choice of the initial conditions.

We study the response of the system with the Fourier mode initial conditions:

$$\psi_0(x,z) = \psi_{n,m}(x,z), \qquad b_0(x,z) \equiv 0, \qquad (n,m) = (1,1),(1,2),(1,3). \tag{2.25}$$

These initial conditions correspond to low wavenumber smooth functions. When $\theta = 0$ the Fourier modes are eigenfunctions, as described in Section 2.3.1, and all three initial conditions result in single frequency standing wave solutions whose frequency is determined by the dispersion relation (2.11). When $\theta \neq 0$, i.e. the domain is tilted by the angle $\theta$ or the direction of gravity is changed, the Fourier

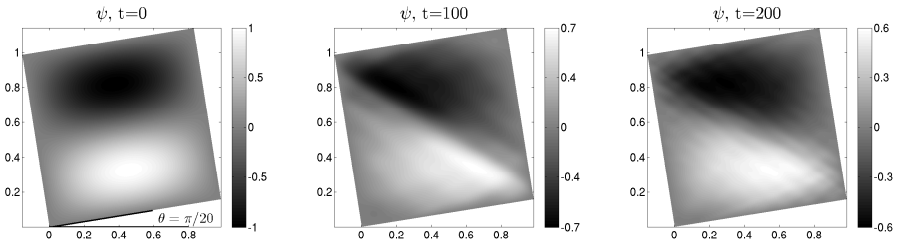



Figure 2.6: Evolution of the stream function in time from the initial condition $\psi_{1,2}$.

modes are no longer eigenfunctions, and we observe a different response from the system for initial conditions (2.25).

In all three numerical examples we use the same numerical parameters and parameter values. We compute to final time $T_{\text{end}} = 400$ with time step $\tau = 0.05$. The spatial mesh sizes in both space dimensions are equal, $\Delta x = \Delta z = 2 \times 10^{-3}$. We fix the stratification frequency $N_f = 1$ and choose $\theta = \pi/20$ for the rotation angle of the square. The Störmer-Verlet method (2.A.12)–(2.A.15) conserves energy in time up to fluctuations of amplitude $\mathcal{O}(\tau^2)$. For this choice of $\tau$ the relative error of the Hamiltonian function (2.19) remained smaller than $10^{-3}$ in all three numerical experiments. Computational results with initial conditions $\psi_{1,1}$ and $\psi_{1,3}$ are shown in Figure 2.2. Results with the initial condition $\psi_{1,2}$ are shown in Figure 2.6. In all three examples we plot the evolution of the stream function at three distinct times.

Complementary to the state variables we also look at the energy density function, i.e. the distribution of the energy in space. Hence we define the discrete energy density function at the cell centers, making use of the discrete velocities defined by (2.A.7),

$$E_{i+1/2,j+1/2} = \frac{1}{2}u^2_{i+1/2,j+1/2} + \frac{1}{2}w^2_{i+1/2,j+1/2} + \frac{1}{2N_f^2}b^2_{i+1/2,j+1/2}. \tag{2.26}$$

In the numerical example with initial condition $\psi_{1,1}$ we observe that energy that is initially concentrated at the low wavenumber is transported to large wave numbers. Evidently, in Figure 2.2 the whole family of $(1, 1)$ attractors is observable. The evolution from initial condition $\psi_{1,3}$ is similar, but in this case the family of $(1, 3)$ wave attractors is obtained, see Figure 2.2. On the other hand, with initial condition $\psi_{1,2}$ the solution appears to consist mainly of a strong periodic component, plus small scale fluctuations.

Despite the fact that the energy functional (2.9) is conserved along the solution of the continuous system (2.5)–(2.8) and the discrete energy function (2.19) is conserved up to second order in time[1] along the solution of the discrete system

[1]Backward error analysis of symplectic numerical integrators (Hairer et al. [37]; Leimkuhler & Reich [63]) shows the existence of a perturbed Hamiltonian of the form $H + \mathcal{O}(\tau^2)$ which is

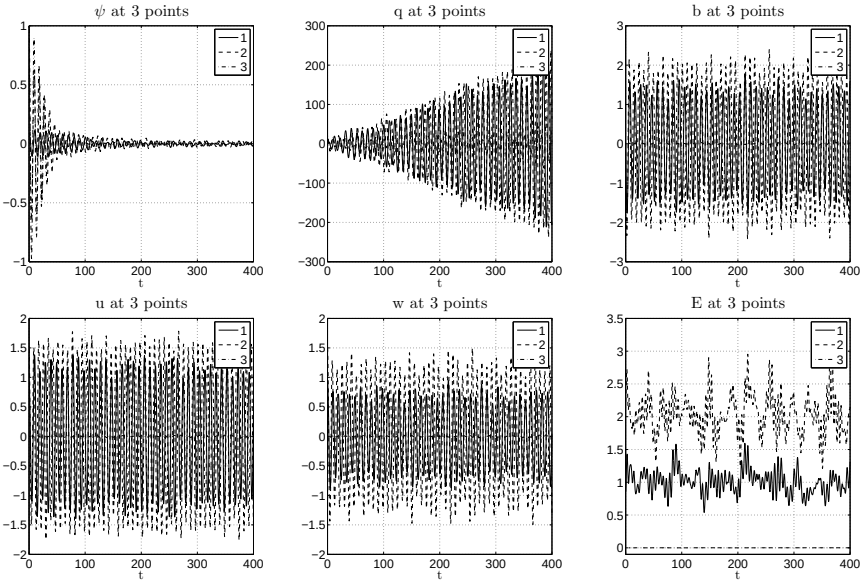
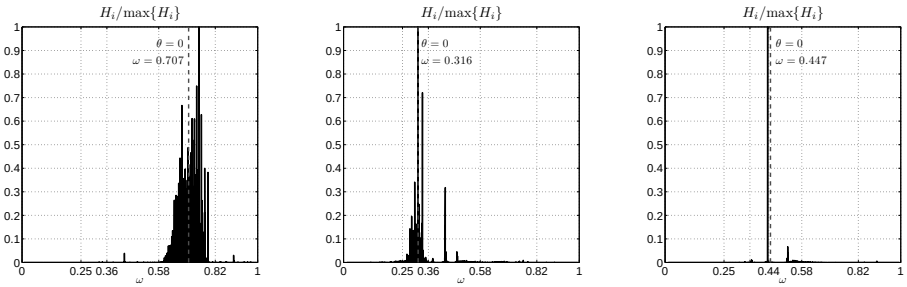



Figure 2.7: Time series of the stream function, vorticity, buoyancy, velocity $u$, velocity $w$ and energy density function $E$ at 3 points in space from computations with initial condition $\psi_{1,1}$ and $b = 0$.

(2.A.12)–(2.A.15), the amplitude of the stream function decays. That can be seen by comparing the color bars in Figures 2.2 and 2.6. For total energy to remain constant, there should either be a net exchange of kinetic into potential energy, or the amplitude of vorticity should grow commensurate to the loss in stream function. To confirm this we study the time series of the state variables: stream function, vorticity, buoyancy, velocities (2.A.7) and the energy density function (2.26), at three arbitrarily chosen points in space. These three points are shown in the top left plot of Figure 2.2. In Figure 2.7 we plot numerical time series data at these three points for the initial condition $\psi_{1,1}$. From Figure 2.7 we see that for energy to stay bounded when the amplitude of the stream function decays the amplitude of the vorticity grows and buoyancy, energy density function and the components of the velocity field stay bounded. This is reminiscent of the familiar cascade of vorticity to large wave numbers in 2D fluids, but note that the nonlinear advection terms are neglected in this model, so the observed effect is really due to dispersion among the normal modes.

The presence of only a single family of wave attractors in the time evolution of the initial conditions $\psi_{1,1}$ and $\psi_{1,3}$ suggests the excitation of only those frequencies associated to the respective class of $(1, 1)$ and $(1, 3)$ wave attractors, respectively. Similarly, the nearly periodic evolution from the $\psi_{1,2}$ Fourier mode suggests the dominance of the periodic $(1, 2)$ solution.

exactly conserved. For our problem, this implies the total energy will be conserved up to bounded fluctuations with amplitude $\mathcal{O}(\tau^2)$.




(a) (b) (c)

Figure 2.8: Energy projections upon normal modes of the semi-discrete system (2.16)–(2.17) for initial conditions $\psi_{1,1}$ (left), $\psi_{1,3}$ (middle), and $\psi_{1,2}$ (right).

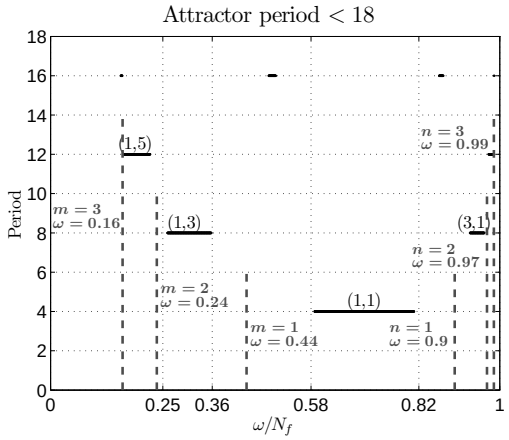
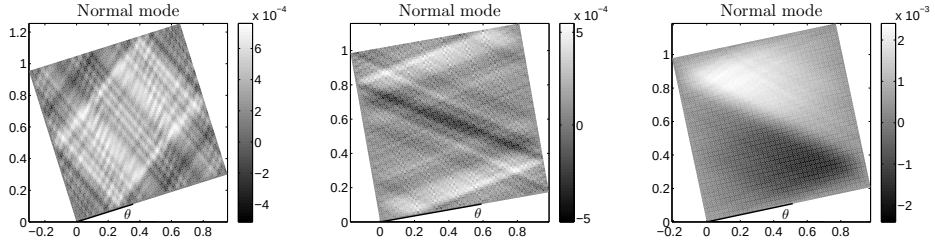



Figure 2.9: Attractor period as a function of subharmonic forcing frequency $\omega/N_f$ for the fixed angle $\theta = \pi/20$. Horizontal bars indicate families of limit cycle attractors, dashed lines indicate the discrete periodic cases.

To understand this, we project the Fourier modes onto the normal modes of the tilted system. We expand the initial conditions (2.25) in the normal modes of the semi-discretization (2.16)–(2.17) for $\theta = \pi/20$ and $N_f = 1$ and plot the scaled discrete energy values $H_i/\max\{H_i\}$ with respect to the frequencies of the discrete system in Figures 2.8(a), 2.8(b) and 2.8(c). In each of these Figures we plot a dashed line to indicate the standing wave solution frequency for $\theta = 0$. The data for Figure 2.9 was taken from the cross-section of Figure 2.3(b) corresponding to tilt angle $\theta = \pi/20$, and were computed by following characteristics. The figure indicates the attractor periods of the limit cycles observed as a function of (subharmonic) forcing frequency, for attractors having period less than eighteen. The horizontal bars reflect the fact that there is a continuous range of forcing frequencies that lead to limit cycle attractors of a given geometry, e.g. the class of $(1,1)$-attractors having period 4. For




(a) (b) (c)

Figure 2.10: Normal modes of the stream function ($N_f = 1$). Left: within the $(1,1)$ attractor frequency range ($\omega = 0.74$, $\theta = 7\pi/72$). Middle: within the $(1,3)$ attractor frequency range ($\omega = 0.34$, $\theta = \pi/18$). Right: the $(1,2)$ periodic solution, ($\omega = 0.43$, $\theta = \pi/15$).

$\theta = \pi/20$ there exist precisely six *periodic* solutions of type $(2n,1)$ and $(1,2m)$ whose discrete frequencies are indicated by the vertical dashed lines. Comparing Figures 2.9 and 2.8(a) we see that the $(1,1)$-Fourier mode projects almost entirely onto the range of $(1,1)$-attractors. Since there is no energy transfer between normal modes the solution of the semi discrete system with initial conditions $\psi_{1,1}$ at any time is a linear combination of the normal modes with frequencies in the range of the $(1,1)$ attractors. Similarly, most of the energy in the $(1,3)$-Fourier mode projects into the range of $(1,3)$ attractors, Figure 2.8(b). In contrast, Figure 2.8(c) illustrates that the $(1,2)$-Fourier mode is concentrated at one discrete frequency, which is very near that of the $(1,2)$ periodic solution, explaining the nearly periodic behavior of this solution.

For future reference, Figure 2.10 shows normal modes with frequencies within the $(1,1)$ and $(1,3)$ attractor ranges, as well as the distinct normal mode with $(1,2)$ periodic solution frequency. The normal modes shown in Figure 2.10 are those whose frequencies are closest to the forcing frequencies of the monochromatic solutions in Figure 2.4. The same frequencies were used to generate the Floquet exponents plotted in Figures 2.5(a) and 2.5(b), and to force the solutions shown in Figure 2.11. The normal modes displayed in Figures 2.10(a) and 2.10(b) are irregular, with high frequency oscillations near the grid scale, but a low frequency plateau structure is also evident. We have inspected a number of the normal modes having frequencies in the (1,1) and (1,3) attractor regimes. A subset of these possess a large scale structure in which attractor geometry is discernible, as with Figures 2.10(a) and 2.10(b). On the other hand, many of the normal modes have no apparent relation to the attractor structure. Furthermore, we were unable to see any functional relation between the normal mode structure and either frequency or resolution. This is perhaps unsurprising, when one considers that these solutions form an orthogonal basis (in an appropriate inner product) for the discrete stream function space.

In summary, for the untilted case the response to an initial perturbation corre-

sponds to an $(n, m)$ normal mode that simply 'sloshes' sinusoidally in time at the single frequency associated with that mode. In this case there are no other frequencies excited. When the same initial spatial perturbation is given in the tilted square domain, most of its energy is projected onto the whole ensemble of $(n, m)$ attractor modes, each associated with a *different* frequency residing in the $(n, m)$ frequency window.

### 2.4.2 Computation of wave attractors

In Section 2.2 we described how to compute monochromatic wave solutions in a tilted square. We illustrated this with two examples of internal gravity wave attractors, see Figures 2.4(a) and 2.4(b). In this section we compute internal wave attractors as an initial value problem with parametric excitation, so-called parametric resonance solutions.

We solve (2.16)–(2.17) with the Störmer-Verlet method. Since we generate instability in the system by parametric excitation, the amplitude of the solution grows in time, and energy is no longer conserved. We choose forcing frequency $2\omega = \pi$ such that the wave period is $T = 4$ and choose the normalized subharmonic frequency $\omega/N_f$ and tilt angle $\theta$ on the basis of the type of limit behavior we want to simulate. We compute a $(1, 1)$ attractor with parameter values $\omega/N_f = 0.74$ and $\theta = 7\pi/72$, and a $(1, 3)$ attractor with parameter values $\omega/N_f = 0.34$ and $\theta = \pi/18$.

Numerical parameters are fixed for both experiments: the forcing amplitude $\epsilon = 0.1$, time step $\tau = 0.05$ and grid step sizes $\Delta x = \Delta z = 2 \times 10^{-3}$. Initial conditions are chosen to be the Fourier modes $\psi_{1,1}$ and $\psi_{1,3}$ in the computation of the $(1, 1)$ and $(1, 3)$ attractors, respectively. We force the system for 50 wave periods and plot the stream function, buoyancy and the discrete energy density function (2.26) at the final time in Figure 2.11.

Figure 2.11 (top) displays the results for the $(1, 1)$ limit cycle attractor. The energy is focused on the attractor, which reflects from each side of the square once. We observe a standing wave solution with growing amplitude and a 'plateau' type of attractor with piecewise constant stream function. After about 10 wave periods, i.e. at time $t = 40$, the wave motion becomes localized along the straight lines of the attractor. The same 'plateau' type of attractors where observed in laboratory experiments (Hazewinkel et al. [43]). Since all sides of the tilted square are inclined, in the case of a simple $(1, 1)$ attractor, internal waves become focused at all boundaries, because the energy is transported in a counter-clockwise orientation around the attractor, as is indicated in the plots of the energy density function[2], see the right top plot of Figure 2.11.

In Figure 2.11 (bottom) we consider an example of a $(1, 3)$ attractor. It has one reflection point with the upper and lower boundaries of the square, and three reflection points each on the left and right sides of the square. Similarly to the case of the $(1, 1)$ attractor, we observe a standing wave solution that grows in amplitude, and the wave energy is localized along the straight lines of an attractor. The form of the attractor is again of 'plateau' type. Internal waves become highly focused

[2]Due to focusing, the energy density increases *after* reflection. Hence, the anticlockwise direction of energy propagation on the attractor can be deduced from the energy density plots.

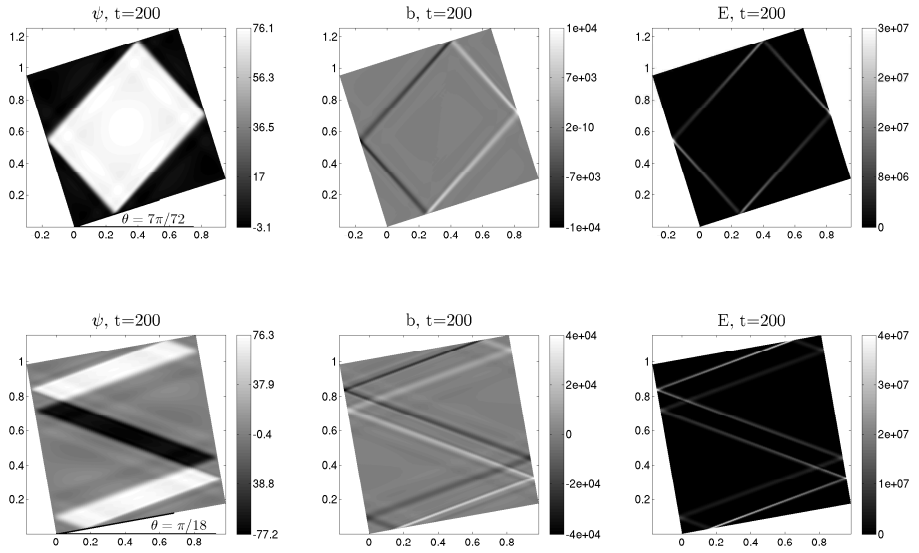



Figure 2.11: Wave attractors after 50 forcing periods (i.e. $t = 200$). Top: $(1, 1)$ attractor regime. Bottom: $(1, 3)$ attractor regime. Shown are the stream function (left), buoyancy (middle) and energy density (right). Initial conditions are the same as in Figure 2.2 at time $t = 0$.

upon reflection from the upper and lower boundaries of the square and gradually defocus in the rest of the domain, see the right bottom plot of Figure 2.11.

Following the discussion of Section 2.4.1, the choice of the initial conditions $\psi_{1,1}$ and $\psi_{1,3}$ ensures that there will be significant energy in the normal modes corresponding to $(1, 1)$ and $(1, 3)$ attractors, a subset of which will grow in amplitude due to resonance of the underlying Mathieu equations. Those modes with frequencies outside the instability tongue of the Mathieu equations remain bounded for all times and eventually become negligible compared to the unstable modes. Since we do not have external damping (like in the experiment discussed in Maas et al. [72]; Lam & Maas [56]), these modes also do not dissipate. Evolution of the stable modes is primarily significant only during the early part of the simulation, before the wave attractor dominates.

Experiments with smaller values of $\epsilon$ result in increased focusing in the neighborhood of the attractor. Figure 2.5 suggests that early on in the computation all the normal modes with frequencies in the resonant zone contribute to the dynamics. But since those modes for which the real part of the Floquet exponent is greater grow much faster in time, these become more prominently visible than others. Because of this energy becomes more and more focused near the attractor as time progresses.

Since there is no exchange of energy between normal modes, the precise structure observed at large times will depend both on the associated Floquet multipliers, and on the initial distribution of energy among the resonant frequencies. In other words, the initial condition is relevant to what is observed in Figure 2.11. On an intermediate time scale (here, 50 forcing periods), those normal modes whose frequencies are associated with the largest Floquet multipliers dominate the solution, and the observed steadily focusing attractor structure is a linear combination of these modes. If integration is carried out for much, much longer times (e.g. thousands of forcing periods for the current resolution), eventually only the distinct normal mode of largest Floquet multiplier will be observable. This can largely be considered a numerical artifact, in many cases having no recognizable attractor pattern, nor corresponding to any physical solution. In the presence of viscosity, the various normal modes do not evolve independently (cf. equation (2.B.24) in Appendix 2.B), and the asymptotic solution is independent of the initial condition (Ogilvie [90]).

Typical normal modes are nonsmooth, for example, as shown in Figure 2.10. The solutions observed in Figure 2.11 are primarily of plateau type. These solutions are composite, consisting of a linear combination of the most resonant modes. Close inspection of the solutions in Figure 2.11 reveals that the plateaus are not perfectly flat, but that there are secondary oscillations of smaller amplitude present. To better observe these, we subtract the plateau solution using the following formula:

$$\delta\psi_{i,j} = \text{trunc}\left(\frac{\psi_{i,j} - \min\{\boldsymbol{\psi}\}}{\max\{\boldsymbol{\psi} - \min\{\boldsymbol{\psi}\}\}}k\right), \qquad \text{trunc}(f) = f - \lfloor f \rfloor,$$

where $\lfloor f \rfloor$ indicates the largest integer less than $f$. The idea of the formula is to rescale the stream function, such that the oscillations about the plateau solution have an amplitude that is less than unity, and then subtract the integer part of the solution everywhere. This is achieved for the empirically chosen value $k = 12$. We plot the secondary wave solution in Figure 2.12 for the stream function at final time $t = 200$. Note the symmetry of the solution and a passing resemblance to Figure 2.4, for which half cosine waves were prescribed on the fundamental intervals. The secondary solutions are also robust with respect to spatial resolution and time step $\tau$. The shape of the secondary solution and its robustness with respect to numerical parameters and perturbation amplitude $\varepsilon$ suggests that the attractor shape is not truly piecewise constant, but has higher order secondary waveforms.

## 2.5 Conclusions

In this chapter we have considered the simplest time dependent configuration in which internal wave attractors can be generated in stratified fluids: linearized, inviscid flow with parametric forcing. We constructed a symmetric, energy conserving finite difference method. For the case of a tilted square geometry we simulated both the free evolution (unforced) wave evolution from Fourier mode initial conditions, and the parametrically forced evolution towards a wave attractor. This simple configuration, as well as the symmetries of the discretization, permit a complete normal mode analysis of the initial value problem in the discrete case. Based on this analysis we can conclude that the finite dimensional approximation has a complete basis

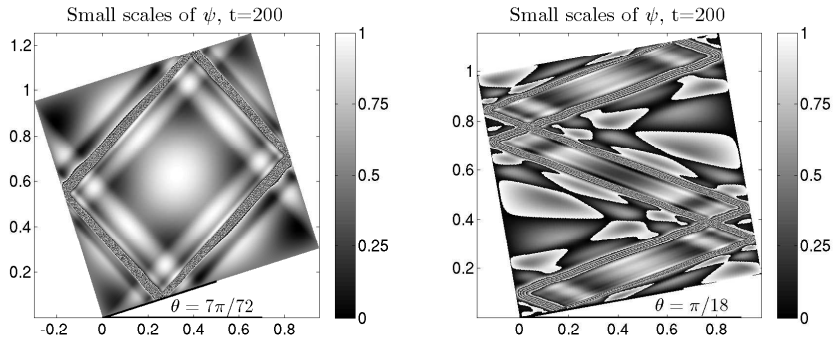



Figure 2.12: Deviation from a piecewise constant solution, after 50 wave periods. Left: $(1,1)$ attractor. Right: $(1,3)$ attractor.

of normal modes that is invariant in time, meaning the initial value problem can be fully decoupled into scalar harmonic oscillators, each of which preserves its initial energy. Therefore, the numerical solution is quasiperiodic, although the Poincaré recurrence time (the time over which a discrete, energy conserving system recovers its initial state) may be quite large. The same analysis can be carried out for the parametrically forced case, showing that the forced system of ODEs can be completely decoupled into Mathieu equations. For a generic initial condition, and depending on the frequency and magnitude of forcing, a range of normal mode frequencies will lie in an Arnold tongue of instability, and the corresponding modes will grow in time, eventually dominating the solution and forming a wave attractor. The shape of the stream function is to first order a plateau, or piecewise constant function, but there are secondary solutions that are robust with respect to discretization and forcing parameters.

We remark that for a given forcing, it is possible to choose judiciously an initial condition whose projection onto the amplified frequencies of the Mathieu equation is zero. In this case, a wave attractor will never be generated. However, this no longer holds if nonlinear advection is taken into account, due to nonlinear coupling. In fact, even for the linearized model, if viscosity is included there is no global decomposition into scalar dynamics, since the normal mode decomposition becomes time dependent.

## 2.A Hamiltonian numerical discretization

The Euler equations for an ideal fluid have a well-known Hamiltonian structure (Arnold [3]; Morrison [87]) that strongly constrains the dynamics. When constructing approximate models such as the Euler-Boussinesq equations (2.1)–(2.4), it is

usually advised to preserve such structure (Salmon [102]). As shown in Holm et al. [45], the nonlinear Euler-Boussinesq equations inherit the noncanonical Hamiltonian structure from the ideal fluid Poisson bracket. Here we verify that the linearization leading to (2.5)–(2.8) also preserves a linear Hamiltonian structure. A system of PDEs on a function space $\boldsymbol{F}^d$ equipped with an inner product $(\cdot,\cdot) : \boldsymbol{F}^d \times \boldsymbol{F}^d \to \mathbf{R}$ is said to constitute a Hamiltonian system (Olver [91]) in the variables $\boldsymbol{f}(x,t) = (f_1(x,t),\ldots,f_d(x,t))^T \in \boldsymbol{F}^d$ if there exists a functional $\mathcal{H}(\boldsymbol{f}) : \boldsymbol{F}^d \to \mathbf{R}$ and a constant, $d \times d$ matrix differential operator (structure matrix) $\mathcal{J} : \boldsymbol{F}^d \to \boldsymbol{F}^d$, that is skew-symmetric with respect to $(\cdot,\cdot)$, such that the PDE can be expressed as

$$\frac{\partial \boldsymbol{f}}{\partial t} = \mathcal{J}\frac{\delta \mathcal{H}}{\delta \boldsymbol{f}}, \tag{2.A.1}$$

where the variational derivative $\delta\mathcal{H}/\delta\boldsymbol{f}$ is defined by

$$\left(\frac{\delta \mathcal{H}}{\delta \boldsymbol{f}}, \boldsymbol{g}\right) = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[\mathcal{H}(\boldsymbol{f}+\varepsilon\boldsymbol{g}) - \mathcal{H}(\boldsymbol{f})\right], \quad \forall\, \boldsymbol{g} \in \boldsymbol{F}^d.$$

One consequence of Hamiltonian structure is the conservation of the Hamiltonian along solutions of (2.A.1), which follows from:

$$\frac{\mathrm{d}\mathcal{H}}{\mathrm{d}t} = \left(\frac{\delta \mathcal{H}}{\delta \boldsymbol{f}}, \frac{\partial \boldsymbol{f}}{\partial t}\right) = \left(\frac{\delta \mathcal{H}}{\delta \boldsymbol{f}}, \mathcal{J}\frac{\delta \mathcal{H}}{\delta \boldsymbol{f}}\right) = 0,$$

by the skew-symmetry condition on $\mathcal{J}$.

We show:

**Proposition 2.A.0.1.** *For any value of $\theta$ the linearized Euler-Boussinesq equations in the stream function formulation (2.5)–(2.8) can be written as a noncanonical Hamiltonian system (2.A.1) in the $L^2$ inner product with $\boldsymbol{f} = (q, b)$, structure matrix*

$$\mathcal{J} = -N_f^2\cos\theta\begin{bmatrix} 0 & \frac{\partial}{\partial x} \\ \frac{\partial}{\partial x} & 0 \end{bmatrix} + N_f^2\sin\theta\begin{bmatrix} 0 & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial z} & 0 \end{bmatrix} \tag{2.A.2}$$

*and Hamiltonian*

$$\mathcal{H} = \frac{1}{2}\int_{\mathcal{D}}\left(\nabla\psi\cdot\nabla\psi + \frac{1}{N_f^2}b^2\right)\mathrm{d}\boldsymbol{x}. \tag{2.A.3}$$

*Proof.* The first variations of the Hamiltonian functional (2.A.3) with respect to $q$ and $b$ are

$$\delta\mathcal{H} = \int_{\mathcal{D}}\left(\nabla\psi\cdot\nabla\delta\psi + \frac{1}{N_f^2}b\delta b\right)\mathrm{d}\boldsymbol{x} = \int_{\mathcal{D}}\left(-\psi\Delta\delta\psi + \frac{1}{N_f^2}b\delta b\right)\mathrm{d}\boldsymbol{x} = \int_{\mathcal{D}}\left(\psi\delta q + \frac{1}{N_f^2}b\delta b\right)\mathrm{d}\boldsymbol{x},$$

where the boundary condition (2.8) has been used to carry out the integration by parts. It follows that the variational derivatives of the Hamiltonian (2.A.3) with respect to the vorticity $q$ and the buoyancy $b$ are

$$\frac{\delta \mathcal{H}}{\delta q} = \psi, \qquad \frac{\delta \mathcal{H}}{\delta b} = \frac{1}{N_f^2} b. \tag{2.A.4}$$

Substituting (2.A.4) and (2.A.2) into (2.A.1) we get that

$$\mathcal{J}\frac{\delta \mathcal{H}}{\delta \boldsymbol{f}} = \mathcal{J}\begin{pmatrix} \frac{\delta \mathcal{H}}{\delta q} \\ \frac{\delta \mathcal{H}}{\delta b} \end{pmatrix} = \begin{pmatrix} -\frac{\partial b}{\partial x}\cos\theta + \frac{\partial b}{\partial z}\sin\theta \\ -N_f^2(\frac{\partial \psi}{\partial x}\cos\theta - \frac{\partial \psi}{\partial z}\sin\theta) \end{pmatrix} = \begin{pmatrix} \frac{\partial q}{\partial t} \\ \frac{\partial b}{\partial t} \end{pmatrix} = \frac{\partial \boldsymbol{f}}{\partial t}$$

which agree with (2.5)–(2.8) □

It follows that the Hamiltonian functional (2.A.3) is conserved along the solution of the equation system (2.5)–(2.8).

### 2.A.1 Finite difference matrices

In this section we describe a numerical discretization for the Euler-Boussinesq equations that preserves a discrete analogue of the Hamiltonian structure in the inviscid, unforced limit. In particular the spatially discrete system of ODEs has a first integral approximating the energy. The scheme also preserves the symmetries of the continuous differential operators. Our approach is to discretize the Hamiltonian and structure operator $\mathcal{J}$ separately, while enforcing the skew-symmetry of $\mathcal{J}$, (see McLachlan [81]). Although this approach leads to a rather standard staggered central difference scheme here, it can be used to construct a Hamiltonian discretization on more general domains and nonuniform grids, which will be important for studying internal waves in ocean basins.

Consider the unit square domain $\mathcal{D} = [0,1]^2$ divided into $N_x \times N_z$ uniform rectangular cells. Subscripted indices shall indicate grid nodes $\boldsymbol{x}_{i,j} = (i\Delta x, j\Delta z)$, where $\Delta x = 1/N_x$ and $\Delta z = 1/N_z$ are the grid sizes in $x$ and $z$ direction, respectively. We shall construct a Hamiltonian structure-preserving staggered finite difference scheme. To this end let us denote by $\mathbf{U} = \mathbf{R}^{N_x \times N_z}$ the space of cell-centered grid functions and by $\mathbf{V} = \mathbf{R}^{(N_x-1)\times(N_z-1)}$ the space of grid functions defined at cell vertices, where in the latter case, we only include inner vertices, since the boundary vertices are either known, or not needed in the discretization.

The discrete stream function $\psi_{i,j}$ and vorticity $q_{i,j}$ are defined at cell vertices and the buoyancy $b_{i+1/2,j+1/2}$ at cell centers. The discrete analogue of the boundary condition on the stream function (2.8) is

$$\psi_{0,j} = \psi_{N_x,j} = 0, \quad \forall j, \qquad \psi_{i,0} = \psi_{i,N_z} = 0, \quad \forall i. \tag{2.A.5}$$

We define column vectors $\boldsymbol{q}, \boldsymbol{\psi} \in \mathbf{V}$ consisting only of the interior grid point values of $q_{i,j}$ and $\psi_{i,j}$. The buoyancy column vector $\boldsymbol{b} \in \mathbf{U}$ consists of all the values of $b_{i+1/2,j+1/2}$ defined at cell centers.

We also define discrete inner products on $\mathbf{U}$ and $\mathbf{V}$:

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{\mathbf{U}} = \sum_{i,j=0}^{N_x-1,N_z-1} a_{i+1/2,j+1/2} b_{i+1/2,j+1/2} \, \Delta x \, \Delta z, \quad \boldsymbol{a}, \boldsymbol{b} \in \mathbf{U},$$
$$\langle \boldsymbol{q}, \boldsymbol{r} \rangle_{\mathbf{V}} = \sum_{i,j=1}^{N_x-1,N_z-1} q_{i,j} r_{i,j} \, \Delta x \, \Delta z, \quad \boldsymbol{q}, \boldsymbol{r} \in \mathbf{V}.$$

For the inner product on $\mathbf{V}$ we assume zero boundary data for at least one of its arguments.

Taking into account the discrete boundary conditions (2.A.5), the following matrices implement the central finite difference approximations to the first derivatives on cell edges:

$$(D_x \boldsymbol{\psi})_{i+1/2,j} = \frac{\psi_{i+1,j} - \psi_{i,j}}{\Delta x}, \qquad (D_z \boldsymbol{\psi})_{i,j+1/2} = \frac{\psi_{i,j+1} - \psi_{i,j}}{\Delta z},$$

where $D_x \in \mathbf{R}^{N_x(N_z-1)\times(N_x-1)(N_z-1)}$ and $D_z \in \mathbf{R}^{N_z(N_x-1)\times(N_x-1)(N_z-1)}$. The dual operators $-D_x^T$ and $-D_z^T$ represent central finite difference approximations to the first derivatives on cell vertices from cell edges.

Additionally we define the averaged operator matrices from cell centers to cell edges:

$$(M_x \boldsymbol{b})_{i,j+1/2} = \frac{b_{i+1/2,j+1/2} + b_{i-1/2,j+1/2}}{2},$$
$$(M_z \boldsymbol{b})_{i+1/2,j} = \frac{b_{i+1/2,j+1/2} + b_{i+1/2,j-1/2}}{2},$$

where $M_x \in \mathbf{R}^{N_z(N_x-1)\times N_x N_z}$, $M_z \in \mathbf{R}^{N_x(N_z-1)\times N_x N_z}$ and their transposes are averaged operator matrices from the cell edges to the cell centers.

The matrices above can be composed in various ways to construct approximate derivative operators from $\mathbf{V}$ to $\mathbf{U}$ and vice versa.

$$M_z^T D_x : \mathbf{V} \to \mathbf{U}, \quad M_x^T D_z : \mathbf{V} \to \mathbf{U}, \quad -D_x^T M_z : \mathbf{U} \to \mathbf{V}, \quad -D_z^T M_x : \mathbf{U} \to \mathbf{V}.$$

The discrete Laplacian operator $L : \mathbf{V} \to \mathbf{V}$, defined as

$$L = -(D_x^T D_x + D_z^T D_z) \in \mathbf{R}^{(N_x-1)(N_z-1)\times(N_x-1)(N_z-1)}, \tag{2.A.6}$$

is the standard symmetric, negative definite, five point central difference stencil, i.e.

$$(L\boldsymbol{\psi})_{i,j} = \frac{\psi_{i+1,j} - 2\psi_{i,j} + \psi_{i-1,j}}{\Delta x^2} + \frac{\psi_{i,j+1} - 2\psi_{i,j} + \psi_{i,j-1}}{\Delta z^2},$$

where the boundary terms are modified to satisfy (2.A.5). We define the discrete vorticity field by $\boldsymbol{q} = -L\boldsymbol{\psi}$.

For diagnostic purposes we also define the discrete velocity components at cell centers:

$$\boldsymbol{u} = -M_x^T D_z \boldsymbol{\psi}, \qquad \boldsymbol{w} = M_z^T D_x \boldsymbol{\psi}. \tag{2.A.7}$$

### 2.A.2 Hamiltonian semi-discretization

To construct a Hamiltonian semi-discretization with structure analogous to (2.A.2), we define a quadrature for $H$ and a skew-symmetric structure that approximates $\mathcal{J}$.

In terms of inner products on $\mathbf{U}$ and $\mathbf{V}$, the discrete Hamiltonian is defined by

$$\begin{aligned} H(\boldsymbol{q},\boldsymbol{b}) &= \frac{1}{2}\left(-\langle\boldsymbol{\psi},\boldsymbol{q}\rangle_{\mathbf{V}} + \frac{1}{N_f^2}\langle\boldsymbol{b},\boldsymbol{b}\rangle_{\mathbf{U}}\right) \\ &= \frac{1}{2}\left(-\langle\boldsymbol{q},L^{-1}\boldsymbol{q}\rangle_{\mathbf{V}} + \frac{1}{N_f^2}\langle\boldsymbol{b},\boldsymbol{b}\rangle_{\mathbf{U}}\right). \end{aligned} \tag{2.A.8}$$

The variational derivatives of $H$ are defined in the weak sense in these inner products by

$$\begin{aligned} \langle\frac{\delta H}{\delta \boldsymbol{q}},\boldsymbol{r}\rangle_{\mathbf{V}} &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(H(\boldsymbol{q}+\varepsilon\boldsymbol{r},\boldsymbol{b}) - H(\boldsymbol{q},\boldsymbol{b})\right) = \langle\boldsymbol{\psi},\boldsymbol{r}\rangle_{\mathbf{V}}, \quad \forall\,\boldsymbol{r}\in\mathbf{V}, \\ \langle\frac{\delta H}{\delta \boldsymbol{b}},\boldsymbol{a}\rangle_{\mathbf{U}} &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(H(\boldsymbol{q},\boldsymbol{b}+\varepsilon\boldsymbol{a}) - H(\boldsymbol{q},\boldsymbol{b})\right) = \langle\frac{1}{N_f^2}\boldsymbol{b},\boldsymbol{a}\rangle_{\mathbf{U}}, \quad \forall\,\boldsymbol{a}\in\mathbf{U}, \end{aligned}$$

i.e.

$$\frac{\delta H}{\delta \boldsymbol{q}} = \boldsymbol{\psi}, \qquad \frac{\delta H}{\delta \boldsymbol{b}} = \frac{1}{N_f^2}\boldsymbol{b}.$$

Next, we define a composite space $\boldsymbol{G} = \mathbf{V}\times\mathbf{U}$. A vector $\boldsymbol{g}\in\boldsymbol{G}$ takes the form

$$\boldsymbol{g} = \begin{pmatrix}\boldsymbol{g}_{\mathbf{V}}\\ \boldsymbol{g}_{\mathbf{U}}\end{pmatrix},$$

where $\boldsymbol{g}_{\mathbf{V}}\in\mathbf{V}$ and $\boldsymbol{g}_{\mathbf{U}}\in\mathbf{U}$. We also define a joint inner product on $\boldsymbol{G}$:

$$\langle\!\langle\boldsymbol{g},\boldsymbol{h}\rangle\!\rangle = \langle\boldsymbol{g}_{\mathbf{V}},\boldsymbol{h}_{\mathbf{V}}\rangle_{\mathbf{V}} + \langle\boldsymbol{g}_{\mathbf{U}},\boldsymbol{h}_{\mathbf{U}}\rangle_{\mathbf{U}},$$

and the variational derivative

$$\frac{\delta H}{\delta \boldsymbol{g}} = \begin{pmatrix}\frac{\delta H}{\delta \boldsymbol{g}_{\mathbf{V}}}\\ \frac{\delta H}{\delta \boldsymbol{g}_{\mathbf{U}}}\end{pmatrix}.$$

We approximate the structure operator (2.A.2) using our finite difference matrices:

$$J = -N_f^2\cos\theta\begin{bmatrix}0 & -D_x^T M_z\\ M_z^T D_x & 0\end{bmatrix} + N_f^2\sin\theta\begin{bmatrix}0 & -D_z^T M_x\\ M_x^T D_z & 0\end{bmatrix}.$$

Note that $J$ is skew-symmetric with respect to $\langle\!\langle\cdot,\cdot\rangle\!\rangle$.

Choosing $\boldsymbol{g} = (\boldsymbol{q},\boldsymbol{b})$, the Hamiltonian semi-discretization of the Euler-Boussinesq equations can now be defined by

$$\frac{\mathrm{d}\boldsymbol{g}}{\mathrm{d}t} = J\frac{\delta H}{\delta \boldsymbol{g}}$$

or, in terms of $\boldsymbol{q}$, $\boldsymbol{b}$ and $\boldsymbol{\psi}$,

$$\frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}t} = D_x^T M_z \boldsymbol{b} \cos\theta - D_z^T M_x \boldsymbol{b} \sin\theta, \tag{2.A.9}$$

$$\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}t} = -N_f^2 (M_z^T D_x \boldsymbol{\psi} \cos\theta - M_x^T D_z \boldsymbol{\psi} \sin\theta), \tag{2.A.10}$$

$$\boldsymbol{q} = -L\boldsymbol{\psi}. \tag{2.A.11}$$

By construction the discrete total energy $H$ is a first integral of the semi-discretization. Additionally, this system of ODEs is reversible and symplectic.

### 2.A.3 Time integration

We have shown that semi-discrete Euler-Boussinesq equations constitute a time-reversible Hamiltonian system. We solve the Hamiltonian system (2.A.9)–(2.A.11) in time with the symmetric and symplectic Störmer-Verlet method (Hairer et al. [37]; Leimkuhler & Reich [63]):

$$\boldsymbol{q}^{n+1/2} = \boldsymbol{q}^n + \frac{\tau}{2}(D_x^T M_z \boldsymbol{b}^n \cos\theta - D_z^T M_x \boldsymbol{b}^n \sin\theta), \tag{2.A.12}$$

$$\boldsymbol{\psi}^{n+1/2} = -L^{-1}\boldsymbol{q}^{n+1/2}, \tag{2.A.13}$$

$$\boldsymbol{b}^{n+1} = \boldsymbol{b}^n - \tau N_f^2 (M_z^T D_x \boldsymbol{\psi}^{n+1/2} \cos\theta - M_x^T D_z \boldsymbol{\psi}^{n+1/2} \sin\theta), \tag{2.A.14}$$

$$\boldsymbol{q}^{n+1} = \boldsymbol{q}^{n+1/2} + \frac{\tau}{2}(D_x^T M_z \boldsymbol{b}^{n+1} \cos\theta - D_z^T M_x \boldsymbol{b}^{n+1} \sin\theta), \tag{2.A.15}$$

such that the Hamiltonian function (2.A.8) will be conserved in time up to small fluctuations of second order amplitude. The method requires the solution of the Poisson equation once per time step, but is otherwise explicit. We solve the Poisson equation efficiently using a fast Poisson solver. The overall method is second order in space and time. Sparse discretization in space combined with a fast Poisson solver allows us to compute efficiently at high spatial resolution.

## 2.B Normal mode decomposition

We next consider the discrete model (2.16)–(2.17) with parametric forcing, written in terms of the stream function $\boldsymbol{\psi} \in \mathbf{R}^M$ and buoyancy $\boldsymbol{b} \in \mathbf{R}^N$:

$$\begin{bmatrix} -L & 0 \\ 0 & \frac{1}{N_f^2} I_N \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix} = \begin{bmatrix} 0 & \alpha(t)K \\ -K^T & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix}, \tag{2.B.16}$$

where $N = N_x N_z$, $M = (N_x - 1)(N_z - 1)$, $L \in \mathbf{R}^{M \times M}$ is the discrete approximation of the Laplacian (2.A.6), $K \in \mathbf{R}^{M \times N}$ is a finite difference matrix

$$K = D_x^T M_z \cos\theta - D_z^T M_x \sin\theta$$

and $I_N$ denotes the identity matrix on $\mathbf{R}^N$. The matrix $L$ is symmetric and negative definite, and hence possesses an orthogonal basis of eigenvectors, and we can write

$-L = QD_LQ^T$, where $Q^TQ = QQ^T = I_M$, $Q \in \mathbf{R}^{M\times M}$ and $D_L \in \mathbf{R}^{M\times M}$ is a diagonal matrix with positive entries. In matrix form we write

$$\begin{bmatrix} QD_LQ^T & 0 \\ 0 & \frac{1}{N_f^2} I_N \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix} = \begin{bmatrix} 0 & \alpha(t)K \\ -K^T & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix}.$$

We transform as follows:

$$\begin{bmatrix} QD_L^{1/2} & 0 \\ 0 & \frac{1}{N_f} I_N \end{bmatrix} \begin{bmatrix} D_L^{1/2}Q^T & 0 \\ 0 & \frac{1}{N_f} I_N \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix}$$
$$= \begin{bmatrix} 0 & \alpha(t)K \\ -K^T & 0 \end{bmatrix} \begin{bmatrix} QD_L^{-1/2} & 0 \\ 0 & N_f I_N \end{bmatrix} \begin{bmatrix} D_L^{1/2}Q^T & 0 \\ 0 & \frac{1}{N_f} I_N \end{bmatrix} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix}$$

or, defining $\hat{\boldsymbol{\psi}} = D_L^{1/2}Q^T\boldsymbol{\psi}$ and $\hat{\boldsymbol{b}} = \frac{1}{N_f}\boldsymbol{b}$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{b}} \end{pmatrix} = N_f \begin{bmatrix} 0 & \alpha(t)D_L^{-1/2}Q^TK \\ -K^TQD_L^{-1/2} & 0 \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{b}} \end{pmatrix}. \tag{2.B.17}$$

Now let $C = N_f D_L^{-1/2}Q^TK \in \mathbf{R}^{M\times N}$. The singular value decomposition of the real matrix $C$ is denoted

$$C = S\Omega R^T,$$

where $S \in \mathbf{R}^{M\times M}$ and $R \in \mathbf{R}^{N\times N}$ are orthogonal matrices and $\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_M)$ is an $\mathbf{R}^{M\times N}$ matrix whose off-diagonals are zero and whose diagonal contains the $M$ real, positive singular values of $C$. Hence (2.B.17) can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{b}} \end{pmatrix} = \begin{bmatrix} 0 & \alpha(t)S\Omega R^T \\ -R\Omega^TS^T & 0 \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{b}} \end{pmatrix}.$$

Transforming again with $\tilde{\boldsymbol{\psi}} = S^T\hat{\boldsymbol{\psi}}$ and $\tilde{\boldsymbol{b}} = R^T\hat{\boldsymbol{b}}$ yields the system of (forced) harmonic oscillators

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \tilde{\boldsymbol{\psi}} \\ \tilde{\boldsymbol{b}} \end{pmatrix} = \begin{bmatrix} 0 & \alpha(t)\Omega \\ -\Omega^T & 0 \end{bmatrix} \begin{pmatrix} \tilde{\boldsymbol{\psi}} \\ \tilde{\boldsymbol{b}} \end{pmatrix}. \tag{2.B.18}$$

Expressed in terms of components, the above system becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{\psi}_i = -\alpha(t)\omega_i^2\tilde{\psi}_i + \dot{\alpha}(t)\omega_i\tilde{b}_i, \qquad i = 1, \ldots, M, \tag{2.B.19}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{b}_i = -\alpha(t)\omega_i^2\tilde{b}_i, \qquad i = 1, \ldots, M, \tag{2.B.20}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\tilde{b}_i = 0, \qquad i = M+1, \ldots, N. \tag{2.B.21}$$

To summarize, let $X = QD_L^{-1/2}S \in \mathbf{R}^{M\times M}$ and $Y = \frac{1}{N_f}R \in \mathbf{R}^{N\times N}$. The columns of $X$ and $Y$, denoted $(X_1, \ldots, X_M)$ and $(Y_1, \ldots, Y_N)$, respectively, represent the normal modes of $\boldsymbol{\psi}$ and $\boldsymbol{b}$. Then the normal mode decomposition

$$\boldsymbol{\psi} = X\tilde{\boldsymbol{\psi}}, \qquad \boldsymbol{b} = Y\tilde{\boldsymbol{b}}, \tag{2.B.22}$$

yields a system of $M$ independent systems (2.B.19)–(2.B.20), plus the $N-M$ trivial dynamics (2.B.21).

**Remark.** Note that if viscosity is included in the model, with viscosity parameter $\nu$, then equation (2.B.16) takes the form

$$\begin{bmatrix} -L & 0 \\ 0 & \frac{1}{N_f^2} I_N \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix} = \begin{bmatrix} \nu L^2 & \alpha(t) K \\ -K^T & 0 \end{bmatrix} \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{b} \end{pmatrix}. \tag{2.B.23}$$

By inverting the matrix on the left, this system is again a linear nonautonomous differential equation of the form

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}t} = A(t)\boldsymbol{f},$$

for some time dependent matrix $A(t)$. Even if $A(t)$ can be diagonalized, the similarity transformation that achieves this will typically be local in time, $A(t) = X(t)D_A(t)X(t)^{-1}$, and so one would not expect there to be a change of variables for which the dynamics decouples for all time. We can carry through the transformations used above in the inviscid case for (2.B.23), and (2.B.18) becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \tilde{\boldsymbol{\psi}} \\ \tilde{\boldsymbol{b}} \end{pmatrix} = \begin{bmatrix} \nu S^T D_L R & \alpha(t)\Omega \\ -\Omega^T & 0 \end{bmatrix} \begin{pmatrix} \tilde{\boldsymbol{\psi}} \\ \tilde{\boldsymbol{b}} \end{pmatrix}, \tag{2.B.24}$$

where we observe that the oscillators have become fully coupled through the (viscous) diagonal term in general.

# Chapter 3

# Thermostats for Constrained Systems

## 3.1 Introduction

Constraints are used in diverse ways in molecular dynamics studies. They replace the stiffest bond stretches in biomolecular models, allowing simulation with larger timesteps than would otherwise be possible [13]; they are part of free-energy and reaction pathway techniques [12, 15, 25], and they are used to constrain normal modes in some enhanced sampling approaches [17]. In general, these methods are implemented in the setting of canonical sampling, i.e. with thermostats, or barostats. The proper treatment of constraints in combination with appropriate thermostatting devices is therefore of great importance and the neglect of their correct handling may lead to uncontrollable errors in computed observables.

When constraints are introduced as a modelling device, they effectively reduce the dimension of phase space. To achieve a good agreement in thermodynamical calculations, a free energy correction, representing the energy of the missing degrees of freedom, should be incorporated (e.g. through a Fixman potential [11, 29, 111]); this thermodynamic correction (which can only be realized in a canonical simulation, i.e. with a thermostat) has the potential to interfere with the calculation of dynamical properties such as diffusion rates. In this chapter we discuss the use of stochastic-dynamical techniques for treating constrained models in the context of these issues.

Consider a Hamiltonian system with generalized coordinates $q, p \in \mathbf{R}^n$ and Hamiltonian

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

where $M$ is a positive definite and symmetric (typically diagonal) mass matrix. The

equations of motion are

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \tag{3.1}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla V(q). \tag{3.2}$$

The Hamiltonian represents the total energy and is a first integral of (3.1)–(3.2). Consequently, a trajectory of this system with initial condition $(q_0, p_0)$ samples the constant energy surface $H(q,p) = H_0 \equiv H(q_0,p_0)$, and when the flow is sufficiently ergodic, we expect time averages to converge to ensemble averages in the micro-canonical ensemble

$$\rho_\mu(q,p) = Z_\mu^{-1}\delta(H(q,p) - H_0), \qquad Z_\mu = \int \delta(H(q,p) - H_0)\,\mathrm{d}q \wedge \mathrm{d}p,$$

where $\mathrm{d}q \wedge \mathrm{d}p = \mathrm{d}q_1 \wedge \cdots \wedge \mathrm{d}q_n \wedge \mathrm{d}p_1 \wedge \cdots \wedge \mathrm{d}p_n$ is the volume form on $\mathbf{R}^{2n}$.

In molecular dynamics one is often interested, not in the dynamics of an isolated system at constant energy, but in a system in thermal equilibrium with a reservoir at temperature $\beta^{-1} = k_B T$. In this case an ergodic system should sample the canonical (Gibbs) distribution

$$\rho_\beta(q,p) = Z^{-1}e^{-\beta H(q,p)}, \qquad Z = \int e^{-\beta H(q,p)}\,\mathrm{d}q \wedge \mathrm{d}p. \tag{3.3}$$

A wide variety of thermostatting devices have been proposed to perturb the Hamiltonian dynamics (3.1)–(3.2) in order to sample the canonical distribution (3.3). In Section 3.2 of this chapter, we will discuss various schemes and the relationships among them, and mention a recently proposed unified framework.

In Section 3.3 of this chapter we will generalize the discussion to the case in which the dynamics (3.1)–(3.2) is subjected to a holonomic constraint, i.e. we enforce an algebraic relation on the position variables $q$, i.e.

$$g(q) = 0, \qquad g : \mathbf{R}^n \to \mathbf{R}^m. \tag{3.4}$$

This constraint restricts the positions $q$ to an $n-m$ dimensional manifold $\mathcal{M}$, and implies a restriction of the velocities $M^{-1}p$ to the tangent space $T_q\mathcal{M}$, i.e.

$$\nabla g(q) M^{-1} p = 0,$$

which follows by taking the derivative of (3.4) with respect to $t$ along a trajectory.

The constraint is enforced by introducing a Lagrange multiplier $\lambda \in \mathbf{R}^m$

$$\frac{\mathrm{d}q}{\mathrm{d}t} = M^{-1}p, \tag{3.5}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\nabla V(q) - \nabla g(q)^T \lambda, \tag{3.6}$$

$$0 = g(q) \tag{3.7}$$

with the augmented Hamiltonian

$$\tilde{H}(q,p,\lambda) = \frac{1}{2}p^T M^{-1} p + V(q) + g(q)^T \lambda. \tag{3.8}$$

Taking the second derivative of (3.4) with respect to time and making use of (3.6) yields an explicit expression for the Lagrange multiplier

$$\lambda = \left(\nabla g(q) M^{-1} \nabla g(q)^T\right)^{-1} \left(\nabla g(q) M^{-1} \nabla V(q) - G(q)\left(M^{-1}p, M^{-1}p\right)\right),$$

where $G(q)$ is the symmetric three-tensor (Hessian) of partial derivatives of $\nabla g(q)$, whose contraction is denoted $G(q)(\cdot,\cdot)$.

When the flow of the constrained dynamical system (3.5)–(3.7) is sufficiently ergodic, we expect time averages to converge to ensemble averages in the microcanonical ensemble

$$\begin{aligned}\rho_{\mu,c} &= Z_{\mu,c}^{-1}\delta(H(q,p)-H_0)\delta(g(q))\delta(\nabla g(q)M^{-1}p),\\ Z_{\mu,c}^{-1} &= \int \delta(H(q,p)-H_0)\delta(g(q))\delta(\nabla g(q)M^{-1}p)\,\mathrm{d}q\wedge\mathrm{d}p.\end{aligned}$$

In the context of molecular dynamics when one is interested in a system in thermal equilibrium with a reservoir at temperature $\beta^{-1}$, an ergodic system should sample the hybrid (Gibbs) distribution

$$\begin{aligned}\rho_{\beta,c}(q,p) &= Z_c^{-1} e^{-\beta H(q,p)}\delta(g(q))\delta(\nabla g(q)M^{-1}p),\\ Z_c &= \int e^{-\beta H(q,p)}\delta(g(q))\delta(\nabla g(q)M^{-1}p)\,dq\wedge dp.\end{aligned} \tag{3.9}$$

Some numerical methods for implementing constrained sampling methods discussed in Section 3.3 are provided in the appendix of this chapter.

Thermostats are, by their very nature, artificial devices. The purposes of thermostatting are varied, including the efficient decorrelation of sampling trajectories and the correction of temperature perturbations due to numerical drift [48, 85] or even applied forcing [50]. A strong motivation for some of the recent proposals (in particular [7, 8, 104]) for thermostats has been the desire to control temperature while exerting the least influence on the dynamics of the system, i.e. staying as close as possible to microcanonical dynamics. This topic has been studied in detail in a recent article [60]. If the convergence to equilibrium of two methods is similar, then the problem is to compare the accuracy of autocorrelation functions produced by the methods, as a measure of the *efficiency* of the thermostat. That is, for a given rate of convergence to the equilibrium measure, a more efficient thermostat is one that least perturbs the dynamics (measured in terms of autocorrelation functions). Similarly, we say that a thermostat is *gentle* if its effect on dynamics is relatively mild for a given rate of convergence of the measure. In this vein, we here demonstrate in Section 3.4 that results of [60] on the smaller autocorrelation error of the Nosé-Hoover-Langevin method compared to Langevin dynamics carry over to the constrained setting.

In Section 3.5 we consider the situation in which constraints are introduced as modelling devices derived as limits of strong restraints. By *restraints* we mean stiffly oscillatory forces or soft constraints. We have in mind applications in molecular dynamics where the constraints are used as models for chemical bonds which should, in a somewhat more accurate model, be allowed to stretch. The suppression of these fast bond vibrations has an advantage for numerical integration: the fast vibration necessitates a small timestep which is not in fact needed to resolve the expensive components of the molecular force field (such as Coulombic interactions). However, this simple approach has a fundamental problem: the thermodynamic properties of a system with strong restraint are not equivalent to those of the constrained system. We can see this by considering the fact that the model with restraint, however stiff, still has momenta which sample a Boltzmann distribution, i.e. they are normally distributed, whereas the constrained system cannot have this property due to the associated tangent space constraint. The energy which would be equidistributed into the transverse components to the constraint manifold must be accounted for in the model using a Fixman biasing potential [11, 29, 40, 96, 111, 112].

The methods are applied to a small chain of 4 atoms with Lennard-Jones forces. Our results suggest that Langevin dynamics performs in a reliable and robust manner for the computation of (stationary) thermodynamic averages, but it is unable to recover autocorrelation functions accurately. In our example, the Nosé-Hoover-Langevin method and Stochastic Velocity Rescaling method prove superior to the Langevin method when the goal is the calculation of dynamics.

## 3.2 Stochastic-dynamical thermostats

In this section, we discuss and compare a variety of methods for achieving canonical sampling in the unconstrained setting. All of these can be written in a simple unified framework [58].

Thermostats come in many different varieties, designed for a range of different purposes. Sometimes thermal control is effected by means of a randomized step with a Metropolis (Monte-Carlo) accept/reject step. In this chapter we are only concerned with methods that generate sampling paths by discretization of a suitable stochastic differential equation obtained as a perturbation of the original dynamics. One of the most popular methods is Langevin dynamics defined by

$$\mathrm{d}q = M^{-1}p\,\mathrm{d}t, \tag{3.10}$$

$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t - \frac{\beta}{2}\sigma\sigma^T M^{-1}p\,\mathrm{d}t + \sigma\,\mathrm{d}W, \tag{3.11}$$

where $W(t)$ is a vector of independent Wiener processes in $\mathbf{R}^n$ and $\sigma \in \mathbf{R}^{n\times n}$.

Nosé-Hoover-Langevin (NHL) dynamics [61, 104] is defined by

$$\mathrm{d}q = M^{-1}p\,\mathrm{d}t, \tag{3.12}$$

$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t + \xi p\,\mathrm{d}t, \tag{3.13}$$

$$\mathrm{d}\xi = \frac{1}{\alpha}\left(n - \beta p^T M^{-1}p\right)\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w, \tag{3.14}$$

where $\xi \in \mathbf{R}$ is an auxiliary thermostat variable and $w(t)$ is a scalar Wiener process, $\sigma \in \mathbf{R}$ and $\gamma = \alpha\sigma^2/2$. The NHL method is constructed such that the extended measure

$$\hat{\rho}(q, p, \xi) = \rho_\beta(q, p)\rho_\alpha(\xi) \tag{3.15}$$

is stationary under the phase space flow, where $\rho_\alpha(\xi)$ is the mean-zero normal distribution with variance $\alpha^{-1}$:

$$\rho_\alpha(\xi) = \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\alpha\frac{\xi^2}{2}\right). \tag{3.16}$$

The NHL dynamics can furthermore be shown to be ergodic in the measure (3.15) whenever the Lie algebra generated by $p$ and $\nabla V(q)$ spans $\mathbf{R}^n$. Whence the projected dynamics on $\mathbf{R}^{2n}$ ergodically samples (3.3).

In [60] it was shown that the NHL dynamics allows a more accurate computation of velocity autocorrelation functions (VAF) in the asymptotic limit of small correlation times.

Other schemes have been suggested recently for sampling purposes. Like NHL, the Stochastic Velocity Rescaling (SVR) method of Bussi et al. [7, 8] has been suggested to provide for thermostatting with a weak perturbation of dynamics. This claim was verified analytically by [60] who generalized the method to:

$$\mathrm{d}q = \nabla_p H \,\mathrm{d}t, \tag{3.17}$$

$$\mathrm{d}p = -\nabla_q H \,\mathrm{d}t - \Psi(K)p \,\mathrm{d}t + \sqrt{2k_B T\Phi(K)}p \,\mathrm{d}W, \tag{3.18}$$

where $W(t)$ is a (scalar) Wiener process, and $\Phi$, $\Psi$ are related by

$$\Psi(K) = (2K - (1+n)k_B T)\Phi(K) - 2k_B T K\frac{\mathrm{d}\Phi}{\mathrm{d}K}.$$

With these choices, the method can be shown to preserve the Gibbs distribution $\rho_\beta$. For the SVR method to be well defined, one also assumes

$$K\Phi(K) \text{ is bounded as } K \to 0,$$

$$\Phi(K) \text{ grows at most polynomially as } K \to \infty.$$

Frank & Gottwald [30] have shown that the NHL method converges to SVR (3.17)–(3.18) in an appropriate strong perturbation limit $\alpha \to 0$.

We mention that all the various methods described in this section have been unified into a single general formulation [58] which can be viewed as including any canonical measure-preserving deterministic extensions of the equations of motion coupled with measure-preserving stochastic perturbation. In general, one augments the system by some degrees of freedom $\xi_1, \xi_2, \ldots, \xi_k$ and designs an extended dynamics so that the density $\rho_\beta\hat{\rho}(\xi_1, \xi_2, \ldots, \xi_k)$ is preserved, for some suitable choice of $\hat{\rho}$. Then, if ergodic, the extended system can be used to compute canonical phase space averages with respect to $\rho_\beta$ (essentially by averaging out over the auxiliary degrees of freedom).

## 3.3 Extension to holonomic constraints

In this section we discuss various methods for treating the equations of motion with holonomic constraints, including Langevin dynamics, the Nosé-Hoover-Langevin dynamics and the Stochastic Velocity Rescaling thermostats. Besides the added constraints and associated Lagrange multiplier, the main difference in the methods is the reduction of degrees of freedom from $n$ to $n-m$, which appears explicitly in the thermostat relations. The derivations are included in Appendix 3.A.

The positions of the system (3.5)–(3.7) are constrained to the configuration manifold $\mathcal{M}$ of co-dimension $m$:

$$\mathcal{M} = \{q \in \mathbf{R}^n \,|\, g(q) = 0\},$$

and the associated phase space is the tangent bundle denoted by

$$\mathcal{TM} = \left\{q, p \in \mathbf{R}^n \,\middle|\, q \in \mathcal{M},\, \nabla g(q) M^{-1} p = 0\right\}.$$

For a given $q \in \mathcal{M}$, the tangent space is defined by

$$\mathcal{T}_q\mathcal{M} = \left\{p \in \mathbf{R}^n \,\middle|\, \nabla g(q) M^{-1} p = 0\right\}.$$

Given a measure (3.3) on the base space $\mathbf{R}^{2n}$, the associated measure on the tangent bundle $T\mathcal{M}$ is obtained by restricting the volume form $\mathrm{d}q \wedge \mathrm{d}p$ to $T\mathcal{M}$. Following [62] we introduce a local chart $(\zeta, \eta)$, where $\zeta, \eta \in \mathcal{D} \subset \mathbf{R}^{n-m}$ and a mapping $\phi(\zeta) : \mathcal{D} \to \mathbf{R}^n$ satisfying $g(\phi(\zeta)) = 0$ and $\nabla g(q) \nabla \phi = 0$. We parametrize $T_q\mathcal{M}$ using the relations

$$q = \phi(\zeta), \tag{3.19}$$

$$p = \nabla\phi \left(\nabla\phi^T \nabla\phi\right)^{-1} \eta. \tag{3.20}$$

Here we assume that the square matrix $\nabla\phi^T\nabla\phi$ has full rank. This map is a canonical transformation. The Hamiltonian (3.8) is transformed to

$$\hat{H}(\zeta, \eta) = \frac{1}{2}\eta^T \left(\nabla\phi^T\nabla\phi\right)^{-T} \left(\nabla\phi^T M^{-1}\nabla\phi\right) \left(\nabla\phi^T\nabla\phi\right)^{-1} \eta + V(\phi(\zeta)), \tag{3.21}$$

and the projected volume form transforms as

$$\mathrm{d}q \wedge \mathrm{d}p = \mathrm{d}\zeta \wedge \mathrm{d}\eta.$$

This means that expectations of a function $f(q, p)$ can be evaluated (locally) as

$$\mathbf{E}\{f\} = \int_{\mathcal{D}} f(q(\zeta), p(\zeta, \eta)) e^{-\beta \hat{H}(\zeta,\eta)}\, \mathrm{d}\zeta \wedge \mathrm{d}\eta,$$

where the integration is understood as an integral over nonoverlapping local coordinate charts. Hence we consider the projected distribution

$$\rho(\zeta, \eta) = Z^{-1} \exp\left(-\beta \hat{H}(\zeta, \eta)\right), \qquad Z = \int_{\mathcal{D}} \exp\left(-\beta \hat{H}(\zeta, \eta)\right) \mathrm{d}\zeta \wedge \mathrm{d}\eta. \tag{3.22}$$

For future reference we note from (3.22) and (3.21) that $\eta$ is mean-zero distributed in $\rho$, i.e. $\langle \eta_i \rangle = 0$.

The generalization of the Langevin dynamics (3.10)–(3.11) to the constrained system (3.5)–(3.7) has been treated in [65]. Considering the phase space measure $\mu_{\mathcal{TM}}$ of $\mathcal{TM}$. The system which admits this measure as an invariant equilibrium measure is the following Langevin process with holonomic constraints:

$$\mathrm{d}q = M^{-1}p\,\mathrm{d}t, \tag{3.23}$$

$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t - \nabla g(q)^T\lambda\,\mathrm{d}t - \gamma(q)M^{-1}p\,\mathrm{d}t + \sigma(q)\,\mathrm{d}W, \tag{3.24}$$

$$0 = g(q), \tag{3.25}$$

where $W(t)$ is n-dimensional Wiener process, and $\gamma(q)$, $\sigma(q)$ are $n \times n$ real matrices. The standard fluctuation-dissipation identity

$$\sigma(q)\sigma(q)^T = \frac{2}{\beta}\gamma(q)$$

has to be imposed such that the canonical distribution on the tangent bundle $\mathcal{TM}$ with the phase space measure $\mu_{\mathcal{TM}}$ is invariant under the dynamics (3.23)–(3.25).

In local coordinates, the Langevin dynamics (3.23)–(3.25) takes the following form:

$$\mathrm{d}\zeta = \nabla_\eta \hat{H}\,\mathrm{d}t, \tag{3.26}$$

$$\mathrm{d}\eta = -\nabla_\zeta \hat{H}\,\mathrm{d}t - \Gamma(\zeta)\nabla_\eta \hat{H}\,\mathrm{d}t + \Sigma(\zeta)\,\mathrm{d}W, \tag{3.27}$$

where $\Sigma(\zeta) = \nabla\phi^T\sigma(\phi(\zeta))$ and the standard fluctuation-dissipation identity

$$\Sigma(\zeta)\Sigma(\zeta)^T = \frac{2}{\beta}\Gamma(\zeta)$$

is satisfied in order for the projected distribution (3.22) to be invariant under the dynamics of (3.26)–(3.27).

The Nosé-Hoover-Langevin dynamics extended with holonomic constraint read:

$$\mathrm{d}q = M^{-1}p\,\mathrm{d}t, \tag{3.28}$$

$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t - \nabla g(q)^T\lambda\,\mathrm{d}t + \xi p\,\mathrm{d}t, \tag{3.29}$$

$$\mathrm{d}\xi = h(p)\,\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w, \tag{3.30}$$

$$0 = g(q), \tag{3.31}$$

where $\xi$ is an auxiliary thermostat variable, $w(t)$ is scalar Wiener process, $\sigma \in \mathbf{R}$, $\gamma = \alpha\sigma^2/2$ and the function $h(p) : \mathbf{R}^n \to \mathbf{R}$ has to be determined.

To find the function $h(\zeta, \eta)$ we ask that the extended projected distribution

$$\hat{\rho}(q, p, \xi) = \rho(\zeta, \eta)\rho_\alpha(\xi), \tag{3.32}$$

where $\rho_\alpha$ is defined in (3.16), be invariant under the Fokker-Planck equation. The calculation is given in Appendix 3.A. We find that

$$h(p) = \frac{1}{\alpha}\left(n - m - \beta p^T M^{-1} p\right)$$

in generalized coordinates $(q, p)$. Note the difference between the constants $n$ of $h(p)$ in NHL dynamics without constraint (3.12)–(3.14) and $n - m$ of $h(p)$ in the NHL dynamics with constraint (3.28)–(3.31). This form is also applicable to the original Hoover thermostat applied to constrained systems.

The Stochastic Velocity Rescaling thermostat method with holonomic constraints reads:

$$\mathrm{d}q = M^{-1}p\,\mathrm{d}t, \tag{3.33}$$

$$\mathrm{d}p = -\nabla V(q)\,\mathrm{d}t - \nabla g(q)^T\lambda\,\mathrm{d}t - \Psi(K)p\,\mathrm{d}t + \sqrt{2k_BT\Phi(K)}p\,\mathrm{d}W, \tag{3.34}$$

$$0 = g(q), \tag{3.35}$$

where $W(t)$ is a scalar Wiener processes, and $\Phi$, $\Psi$ are related by (see Appendix 3.A):

$$\Psi(K) = (2K - (1 + n - m)k_BT)\Phi(K) - 2k_BTK\frac{\mathrm{d}\Phi}{\mathrm{d}K}.$$

Note the difference in the constants $1 + n$ in the SVR dynamics without constraints (3.17)–(3.18) and $1 + n - m$ for the dynamics with constraints (3.33)–(3.35). For the SVR dynamics with constraints the original proposal of Bussi et al. [7, 8] transforms to

$$\Phi(K) = \frac{\gamma''}{2K}, \quad \text{so that} \quad \Psi(K) = \left(1 - \frac{n - m - 1}{2K}k_BT\right)\gamma''. \tag{3.36}$$

### 3.3.1 Numerical methods

All of the methods mentioned above are easily implemented in the constrained setting using ideas of geometric integration (splitting methods). For a discussion of numerical methods for Langevin dynamics, see [82]. The numerical implementations are discussed in Appendix 3.B.

## 3.4 Relative efficiencies of NHL and Langevin

Our approach to investigating the relative efficiencies of the different schemes is to determine the Maclaurin expansion of the velocity autocorrelation function (VAF), comparing the asymptotic convergence of this expansion in the limit of small correlation time $\tau$. All methods are expected to recover the correct (de-)correlation in the limit $\tau \to \infty$. The intermediate time $0 \ll \tau \ll \infty$ is also interesting, but does not easily yield to analysis. However it is hard to imagine that a method can be accurate for intermediate correlation times if it is inaccurate in the limit $\tau \to 0$ studied here. The analysis closely follows that of [60] and the results are analogous to the unconstrained case. For that reason we consider here only the NHL and Langevin methods, and refer the reader to [60] for the SVR method.

Following Leimkuhler et al. [60],

$$F(t) := \frac{1}{F_0}\mathbf{E}_{eq}\left\{p(0)^TM^{-1}p(t)\right\} = \frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^TM^{-1}p\right\},$$
$$F_0 := \mathbf{E}_{eq}\left\{p(0)^TM^{-1}p(0)\right\},$$

in the measure (3.9) for both constrained Langevin dynamics (3.23)–(3.25) and constrained NHL dynamics (3.28)–(3.31). These are compared with the expansion arising from the microcanonical dynamics (3.5)–(3.7). In [60] it was shown that in the limit $t \searrow 0$, Langevin dynamics decorrelates linearly in $t$ whereas NHL dynamics decorrelates as $t^2$, as does the microcanonical dynamics. That is, in the limit $t \searrow 0$, NHL dynamics approaches microcanonical dynamics asymptotically.

We expand the VAF in Maclaurin series,

$$F(t) = 1 + t\frac{\mathrm{d}F(t)}{\mathrm{d}t}\bigg|_{t=0+} + \frac{t^2}{2}\frac{\mathrm{d}^2F(t)}{\mathrm{d}t^2}\bigg|_{t=0+} + O(t^3) \qquad (t > 0),$$

and for comparison we compute the first and the second derivatives of $F$ for the Hamiltonian, Langevin and NHL dynamics.

### 3.4.1 Hamiltonian dynamics

We compute the first derivatives of the VAF for the constrained Hamiltonian dynamics (3.5)–(3.7) (without a thermostat). Multiplying equation (3.6) by $M^{-1}p(0)$ and taking the expectation with respect to the equilibrium measure (3.9) we obtain

$$\begin{aligned}\frac{\mathrm{d}F_{Ham}(t)}{\mathrm{d}t} = \frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\frac{\mathrm{d}p}{\mathrm{d}t}\right\} &= -\frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\nabla V(q)\right\} \\ &\quad -\frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\nabla g(q)^T\lambda\right\} \qquad (t > 0).\end{aligned}$$

Using the canonical transformation (3.19)–(3.20) the first expectation value is

$$\begin{aligned}&\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\nabla V(q)\right\} = \\ &\mathbf{E}^t_{eq}\left\{\eta(0)^T\left(\nabla\phi(0)^T\nabla\phi(0)\right)^{-T}\nabla\phi(0)^T M^{-1}\nabla\phi^T\nabla V(\phi(\zeta))\right\} = 0,\end{aligned}$$

since $\eta$ in each component is mean-zero distributed in (3.22). Taking the limit $t \searrow 0$, the second expectation value is equal to zero since $M^{-1}p(0)$ belongs to the tangent space $T_{q(0)}\mathcal{M}$. Hence we obtain

$$\frac{\mathrm{d}F_{Ham}(t)}{\mathrm{d}t}\bigg|_{t=0+} = 0.$$

We compute the second derivative of the VAF. Differentiating equation (3.6) with respect to $t$, multiplying by $M^{-1}p(0)$ and taking the expectation value with respect to the equilibrium measure (3.9) we obtain

$$\begin{aligned}\frac{\mathrm{d}^2F_{Ham}(t)}{\mathrm{d}t^2} = \frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\frac{\mathrm{d}^2p}{\mathrm{d}t^2}\right\} &= -\frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\mathrm{Hess}_q(\tilde{H})\frac{\mathrm{d}q}{\mathrm{d}t}\right\} \\ &= -\frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\mathrm{Hess}_q(\tilde{H})M^{-1}p\right\} \qquad (t > 0),\end{aligned}$$

where $\mathrm{Hess}_q(\tilde{H})$ stands for Hessian matrix of the constrained Hamiltonian function (3.8) with respect to $q$. In the limit $t \searrow 0$, the term in braces is positive definite, so

$$\left.\frac{\mathrm{d}^2 F_{Ham}(t)}{\mathrm{d}t^2}\right|_{t=0+} \neq 0$$

(in fact this term is strictly negative). Hence the Maclaurin expansion of VAF of the Hamiltonian dynamics

$$F_{Ham}(t) = 1 + \frac{t^2}{2}\left.\frac{\mathrm{d}^2 F_{Ham}(t)}{\mathrm{d}t^2}\right|_{t=0+} + O(t^3) \qquad (t > 0)$$

decorrelates quadratically for small times $t$.

### 3.4.2 Langevin dynamics

We compute the first derivative of the VAF for the Langevin dynamics. Multiplying equation (3.24) by $M^{-1}p(0)$, taking the expectation with respect to the equilibrium measure (3.9) we obtain, since $p(0)$ and $\mathrm{d}W$ are statistically independent if $t > 0$,

$$\begin{aligned}\frac{\mathrm{d}F_{LD}(t)}{\mathrm{d}t} &= \frac{1}{F_0}\mathbf{E}^t_{eq}\left\{p(0)^T M^{-1}\frac{\mathrm{d}p}{\mathrm{d}t}\right\} \\ &= \frac{\mathrm{d}F_{Ham}(t)}{\mathrm{d}t} - \frac{1}{F_0}\mathbf{E}^t_{eq}\left\{\frac{\beta}{2}p(0)^T M^{-1}\sigma(q)\sigma(q)^T M^{-1}p\right\} \qquad (t > 0).\end{aligned}$$

Taking the limit $t \searrow 0$, we obtain

$$\left.\frac{\mathrm{d}F_{LD}(t)}{\mathrm{d}t}\right|_{t=0+} = -\frac{1}{F_0}\mathbf{E}_{eq}\left\{\frac{\beta}{2}p(0)^T M^{-1}\sigma(q(0))\sigma(q(0))^T M^{-1}p(0)\right\} = -\hat{\gamma} \neq 0$$

and the Maclaurin expansion of the VAF for Langevin dynamics is

$$F_{LD}(t) = 1 - \hat{\gamma}t + O(t^2) \qquad (t > 0)$$

with explicit dependence on the parameter $\hat{\gamma}$. Since Hamiltonian dynamics is the same as Langevin dynamics with $\sigma(q) = 0$, which in turn implies $\hat{\gamma} = 0$, the error in $F(t)$ due to the use of Langevin dynamics rather than Hamiltonian dynamics is

$$\Delta_{LD}F(t) := F_{LD}(t) - F_{Ham}(t) = -\hat{\gamma}t + O(t^2) \qquad (t > 0).$$

Thus for small $t$ the magnitude of the error is $\hat{\gamma}t$.

### 3.4.3 The NHL dynamics

The NHL thermostat depends on the variable $\xi$. The following computations are done with respect to the extended equilibrium hybrid (Gibbs) distribution (3.9). We begin by computing the first derivative of the VAF for the NHL dynamics.

Multiplying the equation (3.24) by $M^{-1}p(0)$, taking the equilibrium expectation and dividing by $\mathrm{d}t$ we obtain

$$\begin{aligned}\frac{\mathrm{d}F_{NHL}(t)}{\mathrm{d}t} &= \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{p(0)^T M^{-1}\frac{\mathrm{d}p}{\mathrm{d}t}\right\}\\ &= \frac{\mathrm{d}F_{Ham}(t)}{\mathrm{d}t} + \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{\xi p(0)^T M^{-1}p\right\} \qquad (t>0).\end{aligned}$$

Taking the limit $t \searrow 0$, we find that

$$\left.\frac{\mathrm{d}F_{NHL}(t)}{\mathrm{d}t}\right|_{t=0+} = \frac{1}{F_0}\mathbf{E}_{eq}\left\{\xi(0)p(0)^T M^{-1}p(0)\right\} = 0.$$

The result follows from the fact that $\xi$ is mean-zero normally distributed in the extended measure (3.32). Thus, the Maclaurin series of VAF for NHL dynamics begins with a quadratic term, for which we need the second derivative of $F(t)$ at $t = 0$.

We define the function

$$y := -\nabla_q \tilde{H} + \xi p$$

such that the equation (3.29) for $\mathrm{d}p$ can be written $\mathrm{d}p = y\,\mathrm{d}t$. Differentiating $y$ by the Itō-Doeblin formula and using the equations (3.28)–(3.30) we obtain

$$\begin{aligned}\mathrm{d}y &= -\operatorname{Hess}_q(\tilde{H})\,\mathrm{d}q + \xi\,\mathrm{d}p + p\,\mathrm{d}\xi\\ &= -\operatorname{Hess}_q(\tilde{H})\nabla_p\tilde{H}\,\mathrm{d}t + \xi(-\nabla_q\tilde{H}+\xi p)\,\mathrm{d}t\\ &\quad + p\left(h(p)\,\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w\right) \qquad (t>0).\end{aligned}$$

Hence

$$\frac{\mathrm{d}^2F_{NHL}(t)}{\mathrm{d}t^2} = \frac{1}{F_0}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}_{eq}^t\left\{p(0)^T M^{-1}y\right\} = \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{p(0)^T M^{-1}\frac{\mathrm{d}y}{\mathrm{d}t}\right\} \qquad (t>0).$$

Since $p(0)$ and $\mathrm{d}w(t)$ are statistically independent if $t > 0$,

$$\begin{aligned}\frac{\mathrm{d}^2F_{NHL}(t)}{\mathrm{d}t^2} =& \frac{\mathrm{d}^2F_{Ham}(t)}{\mathrm{d}t^2} - \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{\xi p(0)^T M^{-1}(\nabla_q\tilde{H}+\gamma p)\right\}\\ &+ \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{\xi^2 p(0)^T M^{-1}p\right\} + \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{p(0)^T M^{-1}ph(p)\right\} \qquad (t>0).\end{aligned}$$

Taking the limit $t \searrow 0$ and omitting the terms which turn out to be zero we find that

$$\begin{aligned}\left.\frac{\mathrm{d}^2F_{NHL}(t)}{\mathrm{d}t^2}\right|_{t=0+} &= \left.\frac{\mathrm{d}^2F_{Ham}(t)}{\mathrm{d}t^2}\right|_{t=0+} + \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{\xi(0)^2 p(0)^T M^{-1}p(0)\right\}\\ &\quad + \frac{1}{F_0}\mathbf{E}_{eq}^t\left\{p(0)^T M^{-1}p(0)h(p(0))\right\} = \left.\frac{\mathrm{d}^2F_{Ham}(t)}{\mathrm{d}t^2}\right|_{t=0+} + C(\alpha),\end{aligned}$$

where $C(\alpha) \to 0$ when $\alpha \to \infty$. This thermostat reduces to the Hamiltonian dynamics in the limit $\alpha \to \infty$, since in this limit $h(p) \to 0$ and the Gaussian

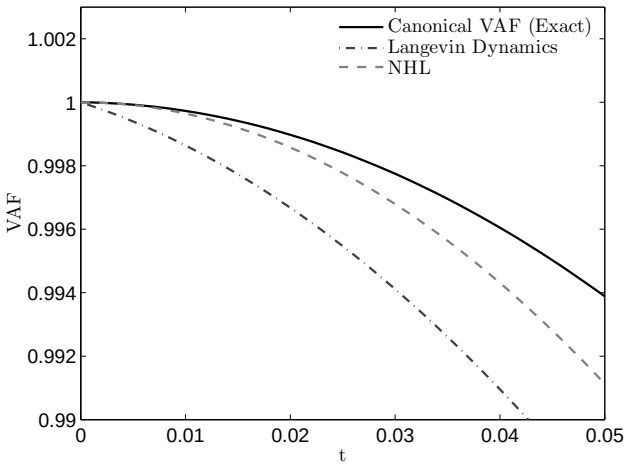



Figure 3.1: Convergence of velocity autocorrelation functions for: microcanonical simulation, NHL method, and Langevin method, in the limit of small t.

distribution for $\xi$ converges to the delta function $\delta(\xi)$, such that the expectation value of $\xi(0)^2$ is equal to zero.

Thus

$$\Delta_{NHL}F(t) := F_{NHL}(t) - F_{Ham}(t) = \frac{1}{2}C(\alpha)t^2 + O(t^3) \qquad (t > 0),$$

i.e. in the limit $t \searrow 0$, NHL dynamics approaches microcanonical dynamics asymptotically. This asymptotic behaviour at small $t$ is illustrated in Figure 3.1 for a double pendulum.

## 3.5 Treatment of a flexible constraint

Let us briefly recount the observations of [28, 29, 33, 34, 96, 111, 112] regarding the statistical mechanics of systems in which a stiff restraining term is replaced by a holonomic constraint. To illustrate the discussion, consider, as in [96], a Hamiltonian

$$H = H_0(q,p) + U_\varepsilon(q), \qquad H_0(q,p) = \frac{p^T M^{-1} p}{2} + U(q), \qquad U_\varepsilon(q) = \frac{1}{2\varepsilon}g^2(q),$$

where $g : \mathbf{R}^n \to \mathbf{R}$ is a function of the position variables and $\varepsilon$ is a (small) parameter. The equations of motion are

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}t} &= M^{-1}p, \\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\nabla U(q) - \varepsilon^{-1}\nabla g(q)^T g(q). \end{aligned}$$

In the limit $\varepsilon \to 0$, the equations reduce to constrained Euler-Lagrange equations

$$\begin{aligned}\frac{\mathrm{d}q}{\mathrm{d}t} &= M^{-1}p,\\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\nabla U(q) - \nabla g(q)^T\lambda,\\ 0 &= g(q).\end{aligned}$$

If we denote the solution of the flexible system by $(q_\varepsilon, p_\varepsilon)$ the suggestion is that

$$\varepsilon^{-1} g(q_\varepsilon(t)) \sim \lambda.$$

This simple analysis appears to justify replacing the stiff restraint by the constrained alternative, but the situation is a little more complicated. Let us assume that our original restrained system is modelled at a prescribed temperature $T$. We expect, assuming ergodicity, that some energy is present in the degree of freedom corresponding to the transverse (vibrational) motion. In the linearly restrained case, i.e. if $g(q) = \gamma \cdot q - \delta$, for some vector $\gamma$ and scalar $\delta$, the energy of restraint is quadratic and we easily justify

$$\varepsilon^{-1} \left\langle g^2(q_\varepsilon) \right\rangle \sim k_B T.$$

(It might be assumed that a similar relation holds for more general systems as long as the constraints are sufficiently smooth.) Thus some energy is present in the restraint, of fixed amount and independent of $\varepsilon$. In the constrained case there is no transverse energy at all. Thus there is a gap between the two models, and this will lead to incorrect calculation of statistical quantities when the constrained model is substituted for the unconstrained one. In essence, this means that the stiffer the restraint, the faster the restraint oscillates.

The idea of Van Kampen [111] and Fixman [29] was to "average out" over the fast vibrational motion, computing the free energy of the remaining degrees of freedom in the presence of this rapidly fluctuating auxiliary variable. Then it turns out that the modification of configurational statistics needed in order to compensate for the vibrational degrees of freedom can be modelled by the incorporation of the simple potential energy correction term, often termed the *Fixman potential*. The modified constrained system is then simply

$$\begin{aligned}\frac{\mathrm{d}q}{\mathrm{d}t} &= M^{-1}p,\\ \frac{\mathrm{d}p}{\mathrm{d}t} &= -\nabla U(q) - \nabla U_{\mathrm{Fix}}(q) - \nabla g(q)^T\lambda,\\ 0 &= g(q),\end{aligned}$$

where

$$U_{\mathrm{Fix}}(q) = k_B T \ln \|\nabla g\|. \tag{3.37}$$

The observation is that the canonical statistical mechanics of this system will provide configurational averages which are corrected for the constraining approximation. That is, canonical averages of functions of the positions of the Fixman- adjusted constrained system will correspond, in the limit $\varepsilon \to 0$, to the corresponding

averages taken in the unconstrained system. (This relationship has recently been explored in detail by C. Hartmann [40].) Hence the corrected model can be used as a foundation for configurational sampling. It is important to note, however, that even the stationary averages of functions of momenta—let alone autocorrelation functions or diffusion constants—will be incorrect with or without the Fixman term.

This raises an interesting question. If our goal is to compute some dynamical quantities, how can we achieve this in the setting of constrained dynamics? Clearly we have no hope of calculating accurate dynamics that heavily depends on the vibrational (transverse to the constraint manifold) degrees of freedom, unless we are prepared to properly model this. But what if the function of interest is, for example, a long term rearrangement involving some coarsened degrees of freedom such as backbone dihedral angles or tertiary structural characteristics in a biomolecule, or order parameters or end-to-end stretch in a polymer? Then one may still hope that the dynamics of the Fixman system will reflect some of the dynamical properties of interest. However, there is an additional complication: the thermostat! The Fixman system itself only makes sense if it is implemented within a framework of canonical molecular dynamics, implying the use of a thermostat. The thermostat will itself complicate the picture in general, and distort the dynamics of the model. Thus we see an added motivation for a gentle thermostat in the setting of soft constraints.

## 3.6 Numerical experiment

In this section, we compare a number of the mentioned methods for the problems of calculating equilibrium distributions and dynamics of a small planar constraint chain. We begin with the model of an $N$-particle chain defined by the Hamiltonian

$$H = \frac{1}{2}\sum_{i=1}^{N} \|p_i\|^2 + \frac{1}{2\varepsilon}\sum_{i=1}^{N}(\|q_i - q_{i-1}\| - 1)^2 + \sum_{i=0}^{N}\sum_{j=i+2}^{N} \phi_{LJ}(\|q_i - q_j\|), \qquad (3.38)$$

where $q_i \in \mathbf{R}^2$, $\|\cdot\|$ represents Euclidean 2-norm and $\phi_{LJ}$ is the Lennard-Jones potential, acting between all atom pairs except those sharing a bond. We define $q_0 \equiv 0$ in all summations. For appropriately scaled initial conditions, when $\varepsilon$ is driven to zero the system assumes the constrained form with $N-1$ constraints of the form $g_i(q) = \|q_{i+1} - q_i\|^2 - 1 = 0$ [112]. Fixman forces were calculated for this model and are given (for pedagogical purposes) in Appendix 3.C. We were interested in the comparison of both sampling and dynamics of the different constrained methods with those of the unconstrained model. We used a small value $\varepsilon = 10^{-4}$ for the restraint, making a stiff spring which introduced an additional numerical challenge due to stability of the numerical method. For simplicity we worked with a chain of length 4 which gave sufficiently interesting behaviour.

We first compare the equilibrium distributions obtained by the various methods. We chose to compute the end-to-end distance (chain extension) $R = \|q_N - q_0\|$ as observable. The distributions (at $k_B T = 2$) for $R$ were computed with each method. These are shown in Figure 3.2. For the calculation in the unconstrained case we used small stepsizes $\Delta t = 10^{-4}$ and in the constrained case $\Delta t = 10^{-2}$. All long time simulations were run on the interval $t \in [0, 10^6]$.

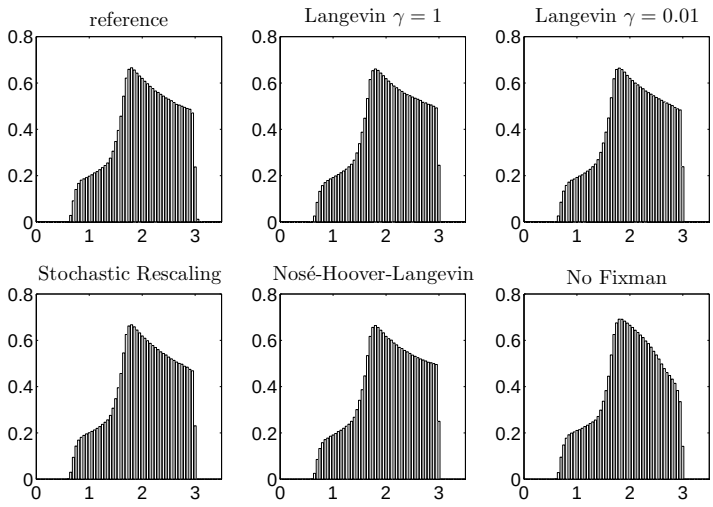



Figure 3.2: Probability density functions of the end-to-end distance $R = \|q_N - q_0\|$ for the chain model (3.38), using stiff restraints (top left) and constraints: Langevin with strong $\gamma = 1$ (top center) and weak $\gamma = 0.01$ (top right) thermostatting, SVR (bottom left) and NHL (bottom center) methods. The bottom right pdf illustrates the necessity of the Fixman correction.

As shown in Figure 3.2 all thermostat methods (Langevin, SVR, and NHL) produced pdfs that were essentially identical to the reference distribution, over a wide range of parameter values $\gamma$. This indicates that the methods are ergodic in the desired measure, and that the Fixman force is effective in correcting the distribution to that of the stiff restrained case. On the contrary, the lower right subplot in Figure 3.2 includes a distribution computed without Fixman correction (using NHL). The distribution is altered, especially for large extension lengths $R$, illustrating the necessity of the correction term.

Next, we considered the approximation of autocorrelation functions using the constrained thermostat methods. The analysis of Section 3.4 showed that the NHL method reproduces the velocity autocorrelation function to second order in $\tau$ as $\tau \to 0$. It is clear from the derivation that this analysis is specific to velocity autocorrelations. In this section we instead consider a different autocorrelation function, i.e. the relaxation of the difference of the end-to-end distance from its mean value (calculated by averaging over a trajectory),

$$\varphi(\tau) = \mathbf{E}_{eq} \left\{ \left( R(\tau) - \bar{R} \right) \left( R(0) - \bar{R} \right) \right\}.$$

We will investigate numerically the accuracy of the NHL, SVR and Langevin dynamics for this function, following the common practice of evaluating the expectations through long time averaging, relying on the assumption of ergodicity. Leimkuhler

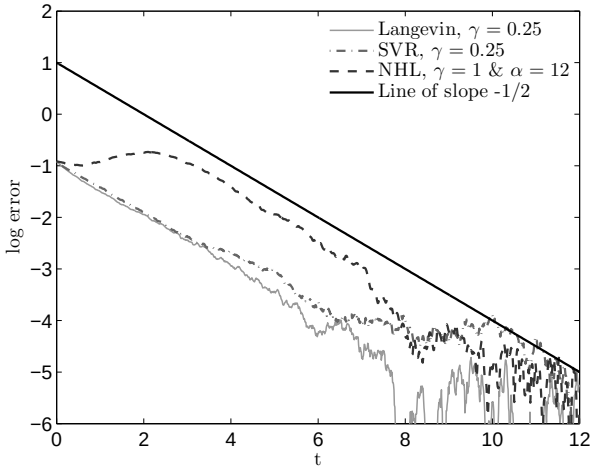



Figure 3.3: Convergence to temperature $k_BT = 2$ for the Langevin, SVR, and NHL methods, averaged over a $10^4$-member ensemble with initial temperature $k_BT = 2.4$.

et al. [60] derive a first order analysis of the rate of convergence to the canonical measure for NHL, SVR and Langevin dynamics. The analysis in [60] indicates a rate of convergence for all methods proportional to the dissipation parameter $\gamma$ and to the number of degrees of freedom. In our simulations we choose, for the Langevin and SVR thermostats, $\gamma = 0.25$; and for the NHL thermostat, $\gamma = 1$, $\alpha = 12$. For these values, all methods approach the desired temperature at approximately the same rate (slope $-1/2$), as shown in Figure 3.3. We then investigate the degree to which the associated autocorrelation function for $R$ approximates that of the reference curve. The reference solution was computed using a $10^6$-member ensemble, canonically distributed initial conditions, and Hamiltonian (constant energy) dynamics with a stiff restraint ($\varepsilon = 10^{-4}$).

In Figure 3.4 we see that Langevin dynamics with $\gamma = 0.25$, although giving a good sampling of the equilibrium state, completely misses the dynamics of the system beyond the first trough. For smaller values of $\gamma$ the results can be improved somewhat, but in no case was the autocorrelation function well approximated on the given interval. Both the SVR and NHL methods capture the qualitative shape of the autocorrelation function, with NHL approximating the reference solution very closely over the whole interval.

## 3.7 Conclusion

In this chapter, we have presented an overview of stochastic-dynamical thermostatting methods for constrained molecular modelling. We have shown that these methods have properties analogous to those of the unconstrained case. The Nosé-

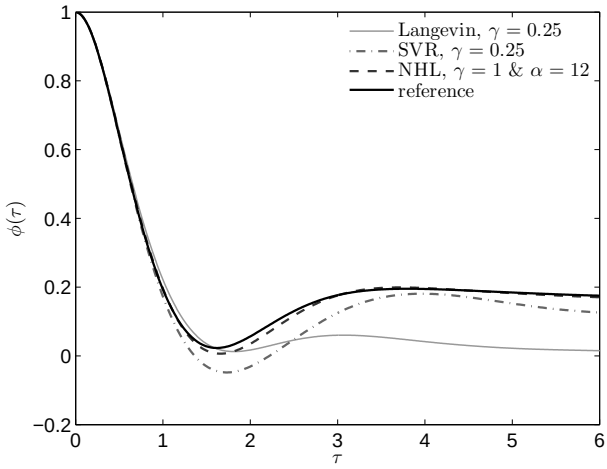



Figure 3.4: Autocorrelation of the end-to-end distance as a function of time using Langevin, SVR, and NHL thermostats. The reference curve was computing using constant energy simulations from a canonically distributed ensemble.

Hoover-Langevin method and the Stochastic Velocity Rescaling method were shown to weakly perturb the dynamics of the system. An application where thermostats are probably essential is in the evolution of constrained systems in the presence of a thermodynamic correction, and for these problems we have shown that the NHL and SVR thermostats with Fixman correction can provide improved accuracy in the autocorrelation function compared to a stiffly restrained model.

# 3.A Constrained stochastic thermostats

In this appendix, we provide the derivations of the constrained forms of thermostat dynamics for the NHL and SVR methods, by introducing local coordinates on the constraint manifold.

We write down the equations (3.28)–(3.29) in local chart coordinates by differentiating the relations (3.19)–(3.20).

$$\begin{aligned} \mathrm{d}q &= \nabla\phi\,\mathrm{d}\zeta, \\ \nabla\phi\,\mathrm{d}\zeta &= \left(M^{-1}\nabla\phi\right)\left(\nabla\phi^T\nabla\phi\right)^{-1}\eta\,\mathrm{d}t. \end{aligned}$$

We multiply both sides by $\nabla\phi^T$ and invert matrix $\nabla\phi^T\nabla\phi$:

$$\begin{aligned} \left(\nabla\phi^T\nabla\phi\right)\,\mathrm{d}\zeta &= \left(\nabla\phi^T M^{-1}\nabla\phi\right)\left(\nabla\phi^T\nabla\phi\right)^{-1}\eta\,\mathrm{d}t, \\ \mathrm{d}\zeta &= \left(\nabla\phi^T\nabla\phi\right)^{-1}\left(\nabla\phi^T M^{-1}\nabla\phi\right)\left(\nabla\phi^T\nabla\phi\right)^{-1}\eta\,\mathrm{d}t, \\ \mathrm{d}\zeta &= \nabla_\eta\hat{H}\,\mathrm{d}t. \end{aligned} \tag{3.A.1}$$

Similarly,

$$\mathrm{d}p = \nabla\phi\big(\nabla\phi^T\nabla\phi\big)^{-1}\,\mathrm{d}\eta + \mathrm{d}\big(\nabla\phi\big(\nabla\phi^T\nabla\phi\big)^{-1}\big)\eta.$$

We multiply both sides by $\nabla\phi^T$ and use property $\nabla\phi^T p = \eta$ to find $\mathrm{d}\eta$:

$$\begin{aligned}
\nabla\phi^T\,\mathrm{d}p &= \mathrm{d}\eta + \nabla\phi^T\,\mathrm{d}\left(\nabla\phi\left(\nabla\phi^T\nabla\phi\right)^{-1}\right)\eta,\\
\mathrm{d}\eta &= -\nabla\phi^T\,\mathrm{d}\left(\nabla\phi\left(\nabla\phi^T\nabla\phi\right)^{-1}\right)\eta + \nabla\phi^T\,\mathrm{d}p,\\
\mathrm{d}\eta &= -\nabla\phi^T\,\mathrm{d}\left(\nabla\phi\left(\nabla\phi^T\nabla\phi\right)^{-1}\right)\eta - \nabla\phi^T\nabla V(\phi(\zeta))\,\mathrm{d}t + \xi\eta\,\mathrm{d}t,\\
\mathrm{d}\eta &= \left[-D(\zeta)\left(\eta, M^{-1}\nabla\phi\left(\nabla\phi^T\nabla\phi\right)^{-1}\eta\right) - \nabla\phi^T\nabla V(\phi(\zeta)) + \xi\eta\right]\mathrm{d}t,\\
\mathrm{d}\eta &= -\nabla_\zeta\hat{H}\,\mathrm{d}t + \xi\eta\,\mathrm{d}t, \qquad (3.A.2)
\end{aligned}$$

where $D(\zeta)$ is the symmetric three-tensor (Hessian) of partial derivative of $\nabla\phi\left(\nabla\phi^T\nabla\phi\right)^{-1}$, whose contraction is denoted $D(\zeta)(\cdot,\cdot)$. Equation (3.30) simply takes the following form:

$$\mathrm{d}\xi = h(\zeta,\eta)\,\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w. \qquad (3.A.3)$$

To find the function $h(\zeta,\eta)$ we ask that the extended projected distribution (3.32) be invariant under the Fokker-Planck equation. We find that

$$h(\zeta,\eta) = \frac{1}{\alpha}\left(\nabla\cdot\eta - \beta\nabla_\eta\hat{H}\cdot\eta\right) = \frac{1}{\alpha}\left(n - m - \beta\nabla_\eta\hat{H}\cdot\eta\right).$$

The NHL method for the equations in the local chart coordinates (3.A.1), (3.A.2) and (3.A.3) is ergodic in the extended projected measure (3.32) whenever the Lie algebra generated by $\eta$ and $\nabla_\zeta\hat{H}$ spans $\mathbf{R}^{n-m}$.

Since

$$\begin{aligned}
\nabla_\eta\hat{H}\cdot\eta &= \left(\left(\nabla\phi^T\nabla\phi\right)^{-1}\left(\nabla\phi^T M^{-1}\nabla\phi\right)\left(\nabla\phi^T\nabla\phi\right)^{-1}\eta\right)\cdot\eta\\
&= \left(M^{-1}\nabla\phi\left(\nabla\phi^T\nabla\phi\right)^{-1}\eta\right)\cdot\left(\nabla\phi\left(\nabla\phi^T\nabla\phi\right)^{-1}\eta\right)\\
&= p^T M^{-1} p,
\end{aligned}$$

we find that

$$h(p) = \frac{1}{\alpha}\left(n - m - \beta p^T M^{-1} p\right)$$

in generalized coordinates $(q,p)$.

For the SVR thermostat (3.33)–(3.35), to find the relation between $\Phi$ and $\Psi$, we re-write the equations in local chart coordinates:

$$\begin{aligned}
\mathrm{d}\zeta &= \nabla_\eta\hat{H}\,\mathrm{d}t,\\
\mathrm{d}\eta &= -\nabla_\zeta\hat{H}\,\mathrm{d}t - \Psi(\hat{K})\eta\,\mathrm{d}t + \sqrt{2k_B T\Phi(\hat{K})}\eta\,\mathrm{d}W,
\end{aligned}$$

where $\hat{K}$ is the kinetic energy in local coordinates. We ask that the projected distribution (3.22) be invariant under the Fokker-Planck equation. We find that functions $\Psi(\hat{K})$ and $\Phi(\hat{K})$ are related by

$$\Psi(\hat{K}) = (2\hat{K} - (1+n-m)k_BT)\Phi(\hat{K}) - 2k_BT\hat{K}\frac{\mathrm{d}\Phi}{\mathrm{d}\hat{K}}.$$

In coordinates on $\mathbf{R}^{2n}$ this relation reads:

$$\Psi(K) = (2K - (1+n-m)k_BT)\Phi(K) - 2k_BTK\frac{\mathrm{d}\Phi}{\mathrm{d}K}.$$

The original proposal of Bussi et al. [7, 8] for the SVR dynamics without constraints corresponds to the choice

$$\Phi(K) = \frac{\gamma''}{2K}, \quad \text{so that} \quad \Psi(K) = \left(1 - \frac{n-1}{2K}k_BT\right)\gamma'',$$

where $\gamma''$ is a positive constant (the constant $1/2\gamma''$ is termed the 'relaxation time'). For the SVR dynamics with constraints the original proposal of Bussi et al. transforms to

$$\Phi(K) = \frac{\gamma''}{2K}, \quad \text{so that} \quad \Psi(K) = \left(1 - \frac{n-m-1}{2K}k_BT\right)\gamma''.$$

## 3.B Aspects of time integration

In this appendix, we describe the numerical implementations of the Langevin, NHL and SVR thermostats with holonomic constraints. For all thermostat methods we adapt the RATTLE algorithm by splitting the system into deterministic and stochastic parts. RATTLE is a symmetric method and symplectic in the Hamiltonian limit. The stochastic part can be then solved, depending on the equations, analytically or numerically.

We denote the time index with a superscript, and we consider a single time step, i.e. the map $(q^0, p^0) \mapsto (q^1, p^1)$.

We consider the Langevin thermostat, the Nosé-Hoover-Langevin thermostat and the Stochastic Velocity Rescaling thermostat equations with holonomic constraints (3.23)–(3.25), (3.28)–(3.31) and (3.33)–(3.35), respectively.

We split the right hand side vector field of (3.23)–(3.25) into a Hamiltonian part and a fluctuation-dissipation part acting only on the momentum. For simplicity, we restrict ourselves to constant, scalar $\sigma$. Hence the fluctuation-dissipation part reduces to the simple Ornstein-Uhlenbeck process, and since the mass matrix is typically diagonal the computation of the analytic solution of the Ornstein-Uhlenbeck process is cheap. Generalizations to constant and position dependent matrix $\sigma$ are straightforward by adapting the approach of [65].

The numerical method for the Langevin dynamics with holonomic constraints reads:

$$\begin{cases} \tilde{p} = \exp\left(-\gamma M^{-1}\dfrac{\tau}{2}\right)p^0 + \sigma\sqrt{\dfrac{1-\exp(-\gamma M^{-1}\tau)}{2\gamma}M}\,\Delta W^0 - \dfrac{\tau}{2}\nabla g(q^0)^T\lambda_0, \\ 0 = \nabla g(q^0)M^{-1}\tilde{p}, \end{cases}$$

$$
\left\{
\begin{array}{l}
q^1 = q^0 + \tau M^{-1} p^{1/2}, \\
p^{1/2} = \tilde{p} - \dfrac{\tau}{2}\nabla V(q^0) - \dfrac{\tau}{2}\nabla g(q^0)^T \lambda_1, \\
0 = g(q^1), \\
\hat{p} = p^{1/2} - \dfrac{\tau}{2}\nabla V(q^1) - \dfrac{\tau}{2}\nabla g(q^1)^T \lambda_2, \\
0 = \nabla g(q^1) M^{-1} \hat{p},
\end{array}
\right.
$$

$$
\left\{
\begin{array}{l}
p^1 = \exp\left(-\gamma M^{-1}\dfrac{\tau}{2}\right)\hat{p} + \sigma\sqrt{\dfrac{1-\exp(-\gamma M^{-1}\tau)}{2\gamma}M}\,\Delta W^1 - \dfrac{\tau}{2}\nabla g(q^1)^T \lambda_3, \\
0 = \nabla g(q^1) M^{-1} p^1,
\end{array}
\right.
$$

where $\tau$ is a time step and $\Delta W^0$ and $\Delta W^1$ are independently and identically distributed Gaussian random variables of mean 0 and covariance matrix $\mathrm{Id}_n$.

For the Nosé-Hoover-Langevin thermostat equations with holonomic constraints (3.28)–(3.31) we split the right hand side vector field in three parts, i.e. a Hamiltonian part, an external forcing and an Ornstein-Uhlenbeck process. Each resulting vector field is solved exactly, i.e. we solve

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \xi p$$

for fixed value of $\xi$, and the scalar Ornstein-Uhlenbeck process

$$\mathrm{d}\xi = \gamma(\mu - \xi)\,\mathrm{d}t + \sigma\,\mathrm{d}w$$

is also solved exactly for given Wiener increments. The numerical method for the Nosé-Hoover-Langevin thermostat equations with holonomic constraints reads:

$$
\left\{ \tilde{p} = \exp\left(\frac{\tau}{2}\xi^0\right) p^0, \right.
$$

$$
\left\{
\begin{array}{l}
q^1 = q^0 + \tau M^{-1} p^{1/2}, \\
p^{1/2} = \tilde{p} - \dfrac{\tau}{2}\nabla V(q^0) - \dfrac{\tau}{2}\nabla g(q^0)^T \lambda_1, \\
0 = g(q^1),
\end{array}
\right.
$$

$$
\left\{ \xi^1 = \exp(-\gamma\tau)\xi^0 + \frac{1}{\gamma}h(p^{1/2})(1-\exp(-\gamma\tau)) + \sigma\sqrt{\frac{1-\exp(-2\gamma\tau)}{2\gamma}}\,\Delta w, \right.
$$

$$
\left\{
\begin{array}{l}
\hat{p} = p^{1/2} - \dfrac{\tau}{2}\nabla V(q^1) - \dfrac{\tau}{2}\nabla g(q^1)^T \lambda_2, \\
0 = \nabla g(q^1) M^{-1} \hat{p},
\end{array}
\right.
$$

$$
\left\{ p^1 = \exp\left(\frac{\tau}{2}\xi^1\right)\hat{p}, \right.
$$

where $\tau$ is a time step and $\Delta w \sim \mathcal{N}(0,1)$. Since the velocities $p^0$ and $\hat{p}$ belong to the tangent spaces $T_{q^0}\mathcal{M}$ and $T_{q^1}\mathcal{M}$, respectively, it follows that the velocities $\tilde{p}$ and $\hat{p}$ belong to the tangent spaces $T_{q^0}\mathcal{M}$ and $T_{q^1}\mathcal{M}$, respectively. Hence we do not need to perform additional projection of the velocities onto the tangent spaces.

The stochastic part of the Stochastic Velocity Rescaling thermostat equations (3.33)–(3.35) reads:

$$\mathrm{d}p = -\Psi(K)p\,\mathrm{d}t + \sqrt{2k_B T\Phi(K)}p\,\mathrm{d}W. \tag{3.B.4}$$

Shortly we will show that the solution of this differential equation only changes the scaling of $p$ and not its direction. Since the RATTLE step ensures $p \in T_q\mathcal{M}$, it is not necessary to introduce a Lagrange multiplier into (3.B.4).

We make the ansatz $p(t) = \alpha(t)p^0$, where $\alpha(t)$ is a scalar function. Substituting this solution into (3.B.4) gives the SDE

$$p^0 \mathrm{d}\alpha = -\Psi(\alpha^2 K_0)\alpha p^0 \,\mathrm{d}t + \sqrt{2k_B T \Phi(\alpha^2 K_0)}\alpha p^0 \,\mathrm{d}W. \tag{3.B.5}$$

Since each term contains a factor $p^0$, we can omit it, leaving a scalar SDE for $\alpha$, and proving our assertion.

For simplicity, we consider the relation (3.36) between functions $\Psi(K)$ and $\Phi(K)$ of the original proposal of Bussi et al. for the systems with constraints such that SDE (3.B.5) takes the particular form

$$\mathrm{d}\alpha = -\left(\alpha - \frac{n-m-1}{\alpha K_0} k_B T\right)\gamma \,\mathrm{d}t + \sqrt{\frac{2k_B T\gamma}{K_0}}\,\mathrm{d}W,$$

where $K_0 = p_0^T M^{-1} p_0$. We note that this SDE has additive noise, making it more amenable to numerical integration than the equation for $K$ proposed in [7]. We solve it by splitting into a nonlinear term and an Ornstein-Uhlenbeck process, applied symmetrically about the RATTLE step. Alternatively one could use the exact solution given in [7], but for our experiments the splitting method with single Wiener process was found to be a cheap alternative. Furthermore, we observed no adverse effects from splitting errors. The numerical method reads:

$$\begin{cases} K_0 = {p^0}^T M^{-1} p^0, \\ \tilde{\alpha} = \sqrt{\dfrac{n-m-1}{K_0} k_B T\gamma\tau + 1}, \\ \alpha = \exp\left(-\gamma\dfrac{\tau}{2}\right)\tilde{\alpha} + \sqrt{\dfrac{1-\exp(-\gamma\tau)}{K_0} k_B T}\,\Delta W^0, \\ \tilde{p} = \alpha p^0, \end{cases}$$

$$\begin{cases} q^1 = q^0 + \tau M^{-1} p^{1/2}, \\ p^{1/2} = \tilde{p} - \dfrac{\tau}{2}\nabla V(q^0) - \dfrac{\tau}{2}\nabla g(q^0)^T \lambda_1, \\ 0 = g(q^1), \\ \hat{p} = p^{1/2} - \dfrac{\tau}{2}\nabla V(q^1) - \dfrac{\tau}{2}\nabla g(q^1)^T \lambda_2, \\ 0 = \nabla g(q^1) M^{-1}\hat{p}, \end{cases}$$

$$\begin{cases} K_0 = \hat{p}^T M^{-1}\hat{p}, \\ \tilde{\alpha} = \exp\left(-\gamma\dfrac{\tau}{2}\right) + \sqrt{\dfrac{1-\exp(-\gamma\tau)}{K_0} k_B T}\,\Delta W^1, \\ \tilde{\alpha} = \sqrt{\dfrac{n-m-1}{K_0} k_B T\gamma\tau + \tilde{\alpha}^2}, \\ p^1 = \alpha\hat{p}, \end{cases}$$

where $\tau$ is a time step and $\Delta W^0, \Delta W^1 \sim \mathcal{N}(0,1)$.

## 3.C Fixman forces for the chain model

In this appendix, we give the detailed description of the Fixman potential (3.37) and force for the chain model considered in Section 3.6.

The Fixman potential for the chain model is defined by

$$U_{\mathrm{Fix}}(q) = \frac{k_B T}{2} \ln \det \left( \nabla g(q) \nabla g(q)^T \right),$$

where $\nabla g(q)$ is the Jacobian matrix of constraint $g(q)$. The matrix product $A := \nabla g(q) \nabla g(q)^T$ is a tridiagonal symmetric matrix. To find the Fixman force

$$F_{\mathrm{Fix}} = -\nabla U_{\mathrm{Fix}} = -\frac{k_B T}{2 \det \left( \nabla g(q) \nabla g(q)^T \right)} \nabla \det \left( \nabla g(q) \nabla g(q)^T \right)$$

we need to compute the gradient of the determinant [28]. Let $\det A$ be the determinant of the matrix $A$, then the determinant can be expressed as

$$\det A = \sum_{i=1}^{N} a_{i,j} C_{i,j}$$

for any $j = 1, \ldots, N$, where $C_{i,j} = (-1)^{i+j} M_{i,j}$ is a so-called cofactor and $M_{i,j}$ is a minor. Note that for the symmetric matrices $C_{i,j} = C_{j,i}$.

The derivative of the determinant $\det A$ with respect to each component of position vector $q_k$ is

$$\frac{\partial \det A}{\partial q_k} = \sum_{i,j} \frac{\partial \det A}{\partial a_{i,j}} \frac{\partial a_{i,j}}{\partial q_k} = \sum_{i,j} C_{i,j} \frac{\partial a_{i,j}}{\partial q_k}.$$

For the chain model special care has to be taken when $k = 1, 2, N-2, N$. The Fixman forces for the chain model can be computed by the following formulas:

$$\begin{aligned}
F^1_{\mathrm{Fix}} &= 2C_{1,1}q_1 + 4C_{2,2}(q_1 - q_2) + 2C_{1,2}(2q_1 - q_2) + 2C_{2,3}(q_3 - q_2), \\
F^2_{\mathrm{Fix}} &= 4C_{2,2}(q_2 - q_1) - 2C_{2,1}q_1 + 4C_{3,3}(q_2 - q_3) \\
&\quad - 2C_{2,3}(q_1 - 2q_2 + q_3) + 2C_{3,4}(q_4 - q_3), \\
F^k_{\mathrm{Fix}} &= 4C_{k,k}(q_k - q_{k-1}) + 2C_{k,k-1}(q_{k-2} - q_{k-1}) + 4C_{k+1,k+1}(q_k - q_{k+1}) \\
&\quad - 2C_{k,k+1}(q_{k-1} - 2q_k + q_{k+1}) \\
&\quad + 2C_{k+1,k+2}(q_{k+2} - q_{k+1}), \quad k = 2 \ldots N-2, \\
F^{N-1}_{\mathrm{Fix}} &= 4C_{N-1,N-1}(q_{N-1} - q_{N-2}) + 2C_{N-1,N-2}(q_{N-3} - q_{N-2}) \\
&\quad + 4C_{N,N}(q_{N-1} - q_N) - 2C_{N-1,N}(q_{N-2} - 2q_{N-1} + q_N), \\
F^N_{\mathrm{Fix}} &= 4C_{N,N}(q_N - q_{N-1}) + 2C_{N,N-1}(q_{N-2} - q_{N-1}),
\end{aligned}$$

where each vector $F^k_{\mathrm{Fix}}$ must be multiplied by $-k_B T/2/\det \left( \nabla g(q) \nabla g(q)^T \right)$.

# Chapter 4

# Weakly Coupled Heat Bath Models for PDEs

## 4.1 Introduction

Thermal bath models such as Langevin dynamics or Nosé-Hoover dynamics are widely used techniques for maintaining the canonical distribution in molecular simulation. Simple thermal baths allow the simulation of bidirectional energy flow, whereas more complicated methods can be designed to provide momentum transfer (barostats) or mimic relaxation processes (generalized Langevin dynamics). As there are natural parallels between turbulent fluids and molecular dynamics, it is interesting to adapt these techniques to hydrodynamics applications. In this chapter, as a first step, we consider an artificial thermal bath for semi-discretized partial differential equations, specifically the Burgers-Hopf and KdV equations.

Molecular dynamics (in the common use of the term) has the structure of a finite dimensional Hamiltonian system, with a total energy function that is a function of positions and momenta. Under typical conditions (the so-called *NVT ensemble*), the volume of the simulation cell is restricted and the number of atoms is fixed, and these may be assumed to share energy equally (equipartition). The system is assumed to be immersed within a larger system (and freely exchanging energy with it) and the energy of the entire system including thermal bath is assumed to remain fixed. In this situation, Gibbs proposed that the microstates of the isolated system will be distributed according to the law

$$\rho_\beta \propto e^{-\beta H},$$

where $H$ is the Hamiltonian (total energy function) of the subsystem, meaning that the invariant measure of the extended system has an associated density which, when integrated out with respect to the bath degrees of freedom, is proportional to $\rho_\beta$. The Gibbs (canonical) distribution is only rigorous for special systems in the so-called thermodynamic limit ($N \to \infty$, $V \to \infty$, $N/V$ fixed); for typical systems such as molecular liquids or proteins, the Gibbs distribution is often assumed and

is the starting point for simulation. In order to maintain the canonical distribution in simulation, various devices are used. The *sampling* problem refers to the calculation of averages of given functions with respect to a specified invariant (equilibrium) distribution. Molecular models may involve constraints (for example fixing the distance between two atoms) or modifications such as those required to model an imposed environmental pressure, so the form of the Gibbs distribution is often modified in practice to reflect such considerations.

In the case of the Gibbs distribution, or, more generally, any distribution defined by a suitably bounded smooth, positive density function, we have a few choices for the mechanism by which sampling is achieved. The Monte-Carlo method [9] is an iteration strategy that combines a randomly generated step with a Metropolis-Hastings accept/reject condition in order to guarantee that the points generated have the desired distribution. In some cases, for example with steep molecular potentials, Monte-Carlo methods may experience large numbers of rejected steps, which can lead to an inefficient sampling of the phase space. Moreover, the sequence of points generated by a Monte-Carlo method has no temporal correlation. For these reasons, molecular modellers often rely on dynamical approaches or the use of stochastic differential equations. These techniques generate paths in phase space which can be used to calculate thermodynamic averages under an ergodic hypothesis: the assumption that the path emanating from any particular initial condition densely covers the relevant portion of phase space with an appropriate probability density. The ergodicity of stochastic dynamics sampling methods such as Brownian or Langevin dynamics can be demonstrated by showing that the adjoint *generator* (i.e. the *Fokker-Planck*, or *Kolmogorov forward* operator) is elliptic or, more generally, hypoelliptic [38, 39, 52, 79, 97, 101].

An alternative to Langevin dynamics often used in molecular simulations is the Nosé-Hoover thermostat [46, 88, 89] which modifies Newtonian dynamics to include an auxiliary variable that provides partial control of the molecular dynamics ensemble; when applied to a sufficiently strongly mixing dynamical system, such deterministic schemes can be effective in practice, although in order to have a rigorous ergodic property it is necessary to incorporate an additional stochastic perturbation. Generalized thermostat methods which combine auxiliary dynamics with stochastic perturbation are studied in [7, 58, 61, 104].

### 4.1.1 Thermostats and PDE models

The foundation for studying the motions of a fluid dynamics model by reference to an invariant distribution has been considered by a variety of authors [4, 10, 27, 55, 69, 83, 84, 92, 99, 100, 103], and there is numerical evidence that these systems typically evolve near thermodynamic equilibrium [2, 20, 21, 78]. Thus it is also natural to consider adapting the thermostatting methodologies to partial differential equations (or their semi-discrete analogues). Equilibrium statistical mechanics is largely dictated by the conservation laws of the system. These constrain the probability space and enter directly into the invariant measure. For partial differential equations, the discretization in space destroys or modifies some or all of the conservation laws, and thereby the invariant measure, resulting in numerical bias [20, 21]. This

creates a potential application for thermostats which is distinct from their motivation in the molecular dynamics setting: they may allow the correction of defects in the distribution due to spatial discretization.

For example, for 2D ideal fluids, the most comprehensive mean field equilibrium theory yields the Miller-Robert-Sommeria (MRS) measure [83, 84, 99, 100], which is grounded in the conservation of the full infinite family of vorticity invariants of the Euler equations. By contrast, standard numerical methods preserve total energy, and at most two of the vorticity invariants. Consequently, when the dynamics of such a system is ergodic, its invariant measure is necessarily significantly different from the MRS measure. (Possible exceptions are the sine-bracket truncation [114] and particle methods [21].)

In addition to perturbing the invariant measure, models for fluids involve dynamics at a range of spatio-temporal scales, and in particular, there may be no clear scale separation. Additionally, there is usually a downscale cascade of vorticity and in some cases kinetic energy, i.e. a secular tendency to excite motion on ever smaller scales: the phenomena known as turbulence. Spatial discretization must arrest this cascade and some sort of closure model (either implicit in the discretization or explicitly modelled and parameterized) is necessary. The choice of closure has consequences for statistical mechanics, and it may be desirable to restore the invariant distribution to correct for the numerical bias. Hence, solely for the purpose of correcting thermodynamic calculations for discretization effects, there is a need to study thermostatting methods in the context of partial differential equations. In Section 4.2 of this chapter, we describe a general framework for treating semi-discrete PDEs using a reasonably general thermostatting methodology.

### 4.1.2 Weak thermostats and accurate dynamical approximation

In the setting of fluids modelling, there can be an additional issue in play. While it can be said that much of molecular modelling is solely focused on the recovery of Gibbs averages, the purpose of simulation in fluids is more often to model dynamics in the vicinity of a Gibbs state. The thermostats used in PDEs may thus be viewed as model corrections to maintain the environment for a dynamical simulation. The requirements of: (1) fast convergence to the invariant distribution (as needed to efficiently compute ensemble averages) and (2) minimal disturbance of the short term time dynamics (as needed for accurately computing correlations) are mutually competing ones, and the design of a good thermostat implies a choice in the tradeoff between these. For this reason, we discuss the concept of a *weak* thermostat.

In the meteorology literature, DelSole [16] observed that the covariance matrix of a variable satisfying a smooth deterministic ordinary differential equation must take the form

$$C(\tau) = C_0 + \tau S + \tau^2 A + \cdots$$

with $S$ a skew-symmetric and $A$ a symmetric matrix. By comparison, the covariance matrix of a variable satisfying a multivariate Ornstein-Uhlenbeck process must take the form

$$C(\tau) = C_0 \exp(\tau \tilde{A}),$$

where $\tilde{A}$ is a (different) symmetric matrix with nonpositive eigenvalues.

In particular, a stochastic process $\eta(t)$ is mean-square differentiable if there exists a function $\dot{\eta}(t)$ such that the expectation

$$\lim_{\varepsilon \to 0} \left\langle \left[ \frac{\eta(t+\varepsilon) - \eta(t)}{\varepsilon} - \dot{\eta}(t) \right]^2 \right\rangle = 0$$

holds in mean-square sense. This derivative is consistent with the deterministic concept. The solution of a smooth dynamical system is differentiable, whereas that of a stochastic differential equation is not so. For accurate computation of dynamical quantities, mean-square differentiability is a desirable property for thermostated dynamics.

Recently, Leimkuhler, Noorizadeh & Penrose [60] have proposed a criterion for assessing the efficiency of a thermostat as a function of the above two criteria. Their analysis in the context of Hamiltonian dynamics showed that the velocity auto-correlation function (VAF) of Nosé-Hoover-Langevin (NHL) dynamics [61] scales as $c(\tau) = 1 - \kappa_2 \tau^2$ in the limit of small correlation times $\tau$, just as the unperturbed dynamics. By comparison, for Langevin dynamics the VAF scales as $c(\tau) = 1 - \kappa_1 \tau$ in this limit. In particular, this implies that VAFs under Langevin dynamics have the wrong curvature at $\tau = 0$, making accurate computation of auto-correlations impossible. For NHL dynamics the noise process is only present in the differential equation for the auxiliary thermostat variable; hence it is integrated once before influencing the momenta variables (and twice before influencing the positions). Consequently the noise in the NHL dynamics takes the form of a memory term or colored noise process and allows for a more accurate computation of correlations.

We mention in passing that another potential application in which the trade-off between fast sampling and accurate dynamics can be expected to play a prominent role is the application of the fluctuation-dissipation theorem to determine the sensitivity of an invariant measure to perturbations in the underlying dynamics [64, 74, 75]. The non-equilibrium response of a system to a small change in its vector field is computed from correlation functions in the unperturbed equilibrium measure. To do so it is necessary both to ensure that the numerical simulation samples the correct measure, and at the same time to perturb the system as little as possible, while allowing the accurate computation of temporal autocorrelations.

### 4.1.3 Results for the Burgers-Hopf and KdV equations

Herein, we apply stochastic-dynamical thermostats to truncated PDE models, in the form of discretized nonlinear wave equations, under the restrictions: (1) that the finite dimensional phase flow is divergence-free, and that (2) the invariant measure is a smooth function of the conserved quantities of the finite dimensional flow (possibly conditional on $\delta$-function measures involving additional conserved quantities).

We demonstrate weakly coupled thermal regulation techniques in the setting of the inviscid Burgers-Hopf (BH) and Korteweg-de Vries (KdV) models

$$u_t + u u_x + \mu u_{xxx} = 0, \tag{4.1}$$

where $\mu = 0$ for BH and $\mu > 0$ for KdV. Both equations are one-dimensional models inheriting the quadratic nonlinearity of fluid motion. The models share a bi-Hamiltonian structure and are formally integrable. However, classical solutions of the Burgers-Hopf equation fail to exist for all time, whereas solutions of the the KdV equation remain smooth. Finite truncation of the BH and KdV models typically breaks integrability. Truncated BH models exhibit chaos and decorrelation of modes on a range of different time scales, and as such it has been used in the literature as a highly simplified model representative of certain aspects of climate [1, 76, 77]. In contrast, truncated solutions of the KdV model may be supposed to retain KAM tori, obstructing ergodicity, and making it a good test model for thermostatting. The truncated BH/KdV models preserve discrete approximations to the first three integrals of the bi-Hamiltonian hierarchy, i.e. the momentum $\mathcal{M} = \int u \, \mathrm{d}x$, kinetic energy $\mathcal{E} = \int \frac{1}{2} u^2 \, \mathrm{d}x$, and Hamiltonian $\mathcal{H} = \int \frac{1}{6} u^3 - \frac{\mu}{2} u_x^2 \, \mathrm{d}x$.

In a series of papers, [1, 76, 77] Majda and co-workers studied the equilibrium statistical mechanics of finite difference and spectral discretizations of the Burgers-Hopf equation ($\mu = 0$), discussing the associated conservation laws and weak invariant sets, and their relation to ergodicity. They computed pdfs of the spectral coefficients, mean spectra, and time-correlation functions. Since the BH equation can be written as a Hamiltonian system in two distinct forms, the definition of the Gibbs measure $\rho \propto \exp(-\beta H)$ depends on the choice of Hamiltonian. Abramov et al. [1] choose the linear Poisson bracket and cubic Hamiltonian, for which the associated Gibbs measure is unbounded. However, since their deterministic dynamics also preserves a quadratic invariant, the resulting product measure (the Gibbs measure restricted to a level set of a hypersphere) does define a probability measure. In this chapter we introduce a perturbation to the BH/KdV model to ensure ergodic sampling of this measure, $\rho \propto \exp(-\beta H)\delta(E - E_0)\delta(M)$, where $H \approx \mathcal{H}$, $E \approx \mathcal{E}$ and $M \approx \mathcal{M}$, and $E_0$ is the initial energy.

### 4.1.4 Ergodicity

For the truncated incompressible Navier-Stokes equation, E & Mattingly [24] proved ergodicity under highly degenerate stochastic forcing of just two modes in the low wave number range, with viscous damping at the large wave number end of the spectrum. Their proof requires establishing a Lyapunov function and verifying the Hörmander condition for their drift and diffusion vector fields. Our concept of thermostatting is meant to provide a realistic model for the interaction of a semidiscrete PDE with the unresolved high modes of the full (infinite) representation, thus we introduce thermostatting only in the high modes and have in effect a situation opposite to that of E and Mattingly. Nevertheless we show that their method based on commutators could in principle be applied in the present instance, were the phase space is flat. In fact our vector fields (in the case of a $N$-mode truncation) are confined to the tangent space of the $(2N-1)$-dimensional hypersphere, so that the calculation of high order brackets becomes extremely involved. We therefore rely on numerical experiments to verify the ergodic property and show that the expected density is obtained with a high degree of accuracy.

The remainder of this chapter is laid out as follows. In Section 4.2 we introduce

the thermostat techniques and discuss the relevant theory. In Section 4.3 we describe the pseudospectral truncations of the BH and KdV equations, present their equilibrium statistical mechanics, discuss ergodicity in the context of thermostatting, and propose some perturbation vector fields. Results with the thermostated dynamics of the BH and KdV are presented in Section 4.4. Discussion and conclusions are given in Section 4.5.

## 4.2 Thermostats

In this section we discuss thermostats in the context of finite-dimensional Hamiltonian systems. In particular we encounter noncanonical Hamiltonian systems with multiple conserved quantities. We discuss the statistical mechanics of general Hamiltonian systems by introducing microcanonical, canonical and mixed canonical distribution functions. To sample the mixed canonical distribution function we discuss the use of a generalized thermostat method for the Hamiltonian system with conserved quantities and consider its theoretical foundation (in particular the ergodicity property).

### 4.2.1 Finite-dimensional Hamiltonian dynamics and statistical mechanics

Consider a Hamiltonian system on $\mathbf{R}^d$, i.e. an initial value problem of the form

$$\frac{\mathrm{d}X}{\mathrm{d}t} = f(X) \equiv J\,\nabla H(X), \qquad X(t) \in \mathcal{D} \subset \mathbf{R}^d, \quad X(0) = X_0, \tag{4.2}$$

where $J = -J^T$ is a constant skew-symmetric matrix, $H(X) : \mathcal{D} \to \mathbf{R}$ is the Hamiltonian, and $\nabla$ denotes the vector of partial derivatives with respect to $X$. The Poisson bracket is an abstract geometrical object associated with the form $J$ and defined by

$$\{F, G\} := \nabla F(X)^T J\,\nabla G(X), \tag{4.3}$$

for arbitrary functions $F(X)$, $G(X) : \mathcal{D} \to \mathbf{R}$. Note that the time derivative of a function $F(X(t)) : \mathcal{D} \to \mathbf{R}$ along a solution to (4.2) is given by

$$\frac{\mathrm{d}F}{\mathrm{d}t} = \{F, H\}.$$

Evident from the antisymmetry of the Poisson bracket, $H(X)$ is invariant under the flow, since $\mathrm{d}H/\mathrm{d}t = \{H, H\} = 0$. In fact, it can be easily checked that any function $\mu(H(X))$ is also invariant. More generally, a first integral of the system is a function $I(X)$ such that

$$\{I, H\} = \nabla I(X)^T J\,\nabla H(X) = 0.$$

The Hamiltonian vector field $f(X)$ defines a flow on $\mathbf{R}^d$. A probability density function $\rho(X,t) : \mathcal{D} \times \mathbf{R} \to \mathbf{R}$, satisfying $\rho(X,t) \geq 0$, $\int \rho(X,t)\,\mathrm{d}X = 1, \forall t$, is transported under the Hamiltonian flow according to

$$\frac{\partial}{\partial t}\rho(X,t) + \nabla \cdot \rho(X,t) f(X) = 0. \tag{4.4}$$

Equilibrium statistical mechanics is concerned with stationary solutions of (4.4). An equilibrium pdf is a solution of

$$\nabla \cdot \rho(X) f(X) = 0.$$

It may be readily checked that the vector field $f(X)$ associated with (4.2) is divergence-free, $\nabla \cdot f(X) = 0$, in which case the above relation simplifies to

$$f(X) \cdot \nabla \rho(X) = 0. \tag{4.5}$$

It follows that $\rho(X)$ is itself a first integral of the flow ($\{\rho, H\} = 0$). In particular, if (4.2) admits precisely $J + 1$ independent first integrals $H$, $I_1$, $\ldots$, $I_J$, then any equilibrium pdf must be a function of these:

$$\rho(X) = \rho(H(X), I_1(X), \cdots, I_J(X)). \tag{4.6}$$

On the other hand, it is clear that any such function $\rho$ that depends on $X$ only through its invariants is stationary under (4.4).

For a system of particles in thermal contact with a heat reservoir, such that energy is exchanged at constant temperature, volume and mass, the likelihood of states is given by the canonical Gibbs density

$$\rho(X) \propto \exp(-\beta H(X)), \tag{4.7}$$

where $\beta$ is the inverse temperature. When more invariants are present, this pdf may be generalized to

$$\rho(X) \propto \exp(-\beta H(X) - \beta_1 I_1(X) - \cdots - \beta_J I_J(X)). \tag{4.8}$$

For the Gibbs measure to define a pdf, it has been assumed that the function is normalizable, i.e. there exists a finite proportionality constant such that $\int \rho(X)\, \mathrm{d}X = 1$.

More generally, one can define an equilibrium pdf as a generalized function, such as the singular measure

$$\rho(X) \propto \delta(H(X) - H^0)\delta(I_1(X) - I_1^0) \cdots \delta(I_J(X) - I_J^0), \tag{4.9}$$

where $\delta$ is the Dirac distribution. In statistical physics this pdf is referred to as the microcanonical ensemble and specifies the relative probabilities of various microstates of a system at fixed values of energy, volume and mass (as well as the other first integrals). It is a stationary solution to (4.4) only in a weak sense.

In some cases it is useful to define a mixed canonical-microcanonical measure such as

$$\rho(X) \propto \exp\left(-\beta H(X)\right) \delta(I_1(X) - I_1^0) \cdots \delta(I_J(X) - I_J^0). \tag{4.10}$$

For example, [1] investigated the statistics of finite-truncations of the Burgers-Hopf equation in a pdf of the form

$$\rho(X) \propto \exp(-\beta H(X))\delta(E(X) - E_0)\delta(M(X)), \tag{4.11}$$

where the level sets of a quadratic invariant $E$ define compact subspaces upon which the Gibbs measure (in the cubic Hamiltonian, see Appendix 4.A) can be normalized.

The expectation of an observable $F(X)$ under the measure $\rho(X)$ is defined as the ensemble average

$$\langle F \rangle = \int_{\mathcal{D}} F(X)\rho(X)\,\mathrm{d}X = \int_{\mathcal{D}} F(X)\,\nu(\mathrm{d}X)$$

for some proper measure $\nu$ such that $\nu \geq 0$ and $\int_{\mathcal{D}} \nu(\mathrm{d}X) = 1$. In general the approximation of such an integral by numerical quadrature is prohibitively expensive due to the large dimension of $X$ encountered in practical applications. Instead Metropolis Monte-Carlo methods are frequently used to compute expectation, despite their slow convergence rate. Such methods give us no information about the dynamics of (4.2) however.

An equilibrium distribution $\rho$ is practically meaningful when it is the density of the unique invariant measure $\nu$ under (4.4). Let $\Phi_t(X)$ denote the time-$t$ flow map of (4.2), and denote by $\Phi_t^n(X)$, its $n$th iterate. We say the flow of (4.2) *samples* the distribution $\rho$ if the iterates $\{\Phi_t^n(X), n \in \mathbf{Z}\} \sim \rho$, for almost all $t$ and almost all $X$. In particular, if the flow is ergodic with respect to $\rho(X)$, then for almost every initial condition $X_0$, the solution to (4.2) samples the equilibrium density $\rho$, and the time average

$$\bar{F} = \lim_{T\to\infty} \frac{1}{T} \int_0^T F(X(t))\,\mathrm{d}t$$

equals the ensemble average $\bar{F} = \langle F \rangle$.

We remark that solutions to the transport equation (4.4) starting from a smooth, nonstationary initial density function $\rho(X, 0)$ do not asymptotically approach a steady state in the sense of classical solutions, due to lack of diffusion. However, they may converge weakly to an equilibrium measure (for example, a uniform measure with compact support on a proper subset of the kinetic energy manifold may converge weakly to the uniform measure on the whole manifold).

The autocorrelation function $c(\tau)$ of observable $F(X)$ is defined by

$$c(\tau) = \frac{\langle F(\Phi_\tau X)F(X)\rangle}{\langle F(X)^2\rangle}.$$

If the flow is ergodic, the autocorrelation can be computed from the time average according to

$$c(\tau) = c_0^{-1} \lim_{T\to\infty} \frac{1}{T} \int_0^T F(X(t))F(X(t+\tau))\,\mathrm{d}t, \quad c_0 = \lim_{T\to\infty} \frac{1}{T} \int_0^T F(X(t))^2\,\mathrm{d}t.$$

### 4.2.2 Generalized Bulgac-Kusnezov thermostats

A typical trajectory of (4.2) cannot ergodically sample a distribution like (4.7) due to preservation of the Hamiltonian $H$. Therefore, in molecular dynamics a number of mechanisms have been introduced to model the thermal exchange with the reservoir, so perturbing the Hamiltonian vector field that typical trajectories of the perturbed dynamics do ergodically sample (4.7).

One such approach is Langevin dynamics, in which balanced stochastic noise and dissipation are added to the Hamiltonian flow, such that the desired measure

becomes the unique, globally attracting invariant measure of the associated Fokker-Planck equations. A generalized form of Langevin dynamics that perturbs (4.2) such that it samples the Gibbs distribution (4.7) is

$$\mathrm{d}X = f(X)\,\mathrm{d}t - \frac{\beta\sigma^2}{2}\nabla H(X)\,\mathrm{d}t + \sigma\,\mathrm{d}W,$$

where $W(t)$ is a vector of independent Wiener processes. One limitation of this approach is that it destroys all invariants of the original system. In order to retain some of these it would be necessary to introduce constraint projections which would also create significant difficulties in discretization. It is well known that additive noise is much easier to treat accurately in discretization than multiplicative noise.

Another approach, proposed by Nosé [88, 89] and Hoover [46], involves the introduction of an auxiliary variable, embedding the Hamiltonian flow in a higher dimensional phase space, such that the projected dynamics on the original phase space is (one hopes) ergodic. The deterministic approach is often non-ergodic, however, motivating the inclusion of Langevin forcing of the auxiliary variable [104]. Nosé-Hoover type schemes can be expanded to include multiple auxiliary variables and more general couplings than originally conceived; a broadened framework was proposed in [58] and termed Generalized Bulgac-Kusnezov (GBK) thermostatting. In the simplest form of a GBK thermostat, we augment the system (4.2) with a small number of additional variables $\xi_k$, $k = 1, \ldots, d_T$, and perturbation vector fields which for our purposes may be assumed to be linear in the $\xi_k$. Let $g_k(X) : \mathcal{D} \to \mathbf{R}^d$, $k = 1, \ldots, d_T$, be smooth vector fields. The complete system is then a set of coupled ordinary and stochastic differential equations of the form:

$$\mathrm{d}X = f(X)\,\mathrm{d}t + \sum_{k=1}^{d_T} \xi_k g_k(X)\,\mathrm{d}t, \tag{4.12}$$

$$\mathrm{d}\xi_k = h_k(X)\,\mathrm{d}t - \gamma\xi_k\,\mathrm{d}t + \sigma\,\mathrm{d}w_k, \quad k = 1, \ldots, d_T, \tag{4.13}$$

where the $w_k(t)$ are independent scalar Wiener processes. The number of thermostat variables $d_T$ is typically small, say $d_T = 1$ or $d_T = 2$, so the computational cost of simulating the thermostated system is essentially equivalent to that of simulating the physical model.

Recall that the Ornstein-Uhlenbeck (OU) process

$$\mathrm{d}\xi = -\gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w \tag{4.14}$$

has analytical solution

$$\xi(t) = e^{-\gamma t}\xi(0) + \sigma\sqrt{\frac{1 - e^{-2\gamma t}}{2\gamma}}\Delta w,$$

where $\Delta w \sim \mathcal{N}(0, 1)$. Choosing $\gamma = \alpha\sigma^2/2$, the normal distribution with mean zero and variance $\alpha^{-1}$, i.e.

$$\vartheta(\xi) = \sqrt{\frac{\alpha}{2\pi}}\exp\left(\frac{-\alpha}{2}\xi^2\right), \tag{4.15}$$

satisfies the stationary Fokker-Planck equation

$$\gamma \frac{\partial}{\partial \xi} \left( \xi \vartheta(\xi) \right) + \frac{1}{2} \sigma^2 \frac{\partial^2}{\partial \xi^2} \vartheta(\xi) = 0. \tag{4.16}$$

In particular, it is well known that the density (4.15) is the unique, globally attracting, steady state solution of the Fokker-Planck equation associated to (4.14). Hence, solutions of (4.14) ergodically sample (4.15).

Of course our interest is not in the simple Ornstein-Uhlenbeck equation but in (4.12)–(4.13). Given a desired distribution $\rho(X)$, we seek $h_k(X) : \mathbf{R}^d \to \mathbf{R}$, $k = 1, \ldots, d_T$, such that the product distribution

$$\pi(X, \xi) = \rho(X) \vartheta(\xi) \tag{4.17}$$

is a stationary solution of the Fokker-Planck equation associated with (4.12)–(4.13), i.e.

$$\begin{aligned} \nabla \cdot \pi(X, \xi) & \left( f(X) + \sum_k \xi_k g_k(X) \right) \\ & + \sum_k \left[ \frac{\partial}{\partial \xi_k} \left( \pi(X, \xi)(h_k(X) - \gamma \xi_k) \right) - \frac{\sigma^2}{2} \frac{\partial^2}{\partial \xi_k^2} \pi(X, \xi) \right] = 0. \end{aligned} \tag{4.18}$$

We proceed formally, assuming a smooth density of the general form (4.6), but note that for singular measures such as (4.10), the above requirement must be satisfied in an appropriate weak sense. The case when the measure depends on a subset of $I_j$ via a Dirac distribution will be handled later.

For concreteness, let $\rho(X) = \exp(-F(X))$, where

$$F(X) = F(H(X), I_1(X), \ldots, I_J(X))$$

is differentiable with respect to all of its arguments, and denote $\beta_0(X) = \partial F / \partial H$ and $\beta_j(X) = \partial F / \partial I_j$, $j = 1, \ldots, J$. The expression (4.18) simplifies under the conditions $\nabla \cdot f = 0$ and $\nabla H \cdot f = \nabla I_j \cdot f = 0$. Additionally using the fact that the terms of the OU process (4.14) satisfy the stationary Fokker-Planck equation (4.16), the relation (4.18) reduces to

$$\begin{aligned} 0 &= \sum_k \xi_k \nabla \cdot \pi(X, \xi) g_k(X) + h_k(X) \frac{\partial}{\partial \xi_k} \pi(X, \xi) \\ &= \sum_k \xi_k \pi(X, \xi) \nabla \cdot g_k(X) - \xi_k \pi(X, \xi) \left( \beta_0 \nabla H + \sum_{j=1} \beta_j \nabla I_j \right) \cdot g_k(X) \\ &\quad - \alpha \xi_k \pi(X, \xi) h_k(X) \\ &= \sum_k \xi_k \left( \nabla \cdot g_k(X) - \left( \beta_0 \nabla H + \sum_{j=1} \beta_j \nabla I_j \right) \cdot g_k(X) - \alpha h_k(X) \right). \end{aligned}$$

Hence it is sufficient to take

$$h_k(X) = \frac{1}{\alpha}\left(\nabla \cdot g_k(X) - \left(\beta_0 \nabla H + \sum_{j=1} \beta_j \nabla I_j\right) \cdot g_k(X)\right)$$

for a given vector field $g_k(X)$. For the Gibbs distribution (4.7), $h_k(X)$ reduces to

$$h_k(X) = \frac{1}{\alpha}\left(\nabla \cdot g_k(X) - \beta \nabla H \cdot g_k(X)\right). \tag{4.19}$$

We have yet to specify the vector fields $g_k$. The construction of this section ensures that the target distribution is invariant under the thermostated Fokker-Planck operator for any choice of $g_k$.

### 4.2.3 The ergodic property

The previous derivation of the GBK method ensures that the augmented probability distribution $\pi$ is invariant under the Fokker-Planck flow associated with the GBK dynamics (4.12)–(4.13). To ensure correct sampling, one must also show that $\pi$ is the density of the unique ergodic invariant measure. By construction, (4.12)–(4.13) define a phase flow under which the density $\pi$ is invariant. The associated measure is positive for all open sets on the phase space. Hence, to show uniqueness and thereby ergodicity, it suffices to show that the Fokker-Planck operator associated to (4.12)–(4.13) is hypoelliptic, which follows from the controllability condition due to Hörmander [38, 39, 52, 79, 97].

Hörmander's condition can be tailored slightly for the GBK thermostat, as demonstrated next. Let $\mathcal{L}(V_0, V_1, \ldots, V_{d_T})$ denote the ideal of the vector fields $V_k$ with $k > 0$ within the Lie algebra generated by all of the $V_k$:

$$\mathcal{L}(V_0, V_1, \ldots, V_{d_T}) = \{V_{k_0}, [V_{k_0}, V_{k_1}], [[V_{k_0}, V_{k_1}], V_{k_2}], \ldots\},$$

where $[\cdot, \cdot]$ denotes the commutator of vector fields, $k_0$ takes values in the set $\{1, \ldots, d_T\}$, and $k_1$, $k_2$, etc. take values in $\{0, \cdots, d_T\}$. Denoting by $\partial_{\xi_k}$ the unit vector in $\mathbf{R}^{d+d_T}$ corresponding to the variable $\xi_k$, Hörmander's condition [101] to ensure a smooth probability measure for this system is

$$\mathbf{R}^{d+d_T} \subset \operatorname{span} \mathcal{L}(F, \partial_{\xi_1}, \ldots, \partial_{\xi_{d_T}}),$$

where

$$F = \begin{pmatrix} f(X) + \sum_k \xi_k g_k(X) \\ h_1(X) - \gamma \xi_1 \\ \vdots \\ h_{d_T}(X) - \gamma \xi_{d_T} \end{pmatrix}$$

denotes the deterministic vector field of (4.12)–(4.13). Defining

$$G_k = [F, \partial_{\xi_k}] = \begin{pmatrix} g_k(X) \\ -\gamma \partial_{\xi_k} \end{pmatrix}, \quad k = 1, \ldots, d_T, \tag{4.20}$$

we find that

$$[F, G_k] = \begin{pmatrix} [f, g_k] \\ 0 \end{pmatrix} + c_1 G_k + c_2(X)\partial_{\xi_k}. \tag{4.21}$$

Since the unit vectors $\partial_{\xi_k}$ form a globally defined basis for the auxiliary space of the thermostat variables $\xi_k$, it remains to construct a basis for the original space $\mathbf{R}^d$. Eliminating the $\xi_k$ and the $G_k$ from (4.21), shows that the following reduced Hörmander condition holds:

**Lemma 4.2.3.1.** *The GBK method (4.12)–(4.13) satisfies Hörmander's condition at a point $(X, \xi_1, \ldots, \xi_{d_T}) \in \mathbf{R}^{d+d_T}$ if the related Hörmander condition on $\mathbf{R}^d$ holds at $X$:*

$$\mathbf{R}^d \subset \operatorname{span} \mathcal{L}(f, g_1, g_2, \ldots, g_{d_T}).$$

When choosing appropriate vector fields $g_k$, it is important to ensure that $f$ and the $g_k$ do not all share an invariant manifold of co-dimension one. For example, in the case $d_T = 1$, $g = g_1$, if $\mathcal{N} = \{X \in \mathcal{D} \,|\, \eta(X) = 0\}$ defines a smooth invariant manifold such that $\nabla\eta(X) \cdot f(X) = \nabla\eta(X) \cdot g(X) = 0$ for all $X \in \mathcal{N}$, then it follows that

$$f, g \in T_X\mathcal{N} \quad \Rightarrow \quad [f, g] \in T_X\mathcal{N},$$

and consequently, the Lie algebra will be rank deficient on $\mathcal{N}$, and Hörmander's condition will fail there. Furthermore, if $\mathcal{N}$ is of co-dimension one, it may partition the phase space.

When constructing thermostats for a mixed measure such as (4.10) we take advantage of the just noted symmetry of the Lie algebra. That is, we choose the perturbation vector fields $g_k(X)$ to satisfy $\nabla I_j \cdot g_k(X) = 0$, $\forall j, k$, and subsequently determine the $h_k(X)$ to ensure the invariance of the smooth part of the measure (4.10), according to (4.19).

To choose the $g_k(X)$, one can either appeal to underlying symmetries of the Hamiltonian vector field (4.2), or make use of a projector onto the tangent bundle of the manifold defined by intersection of the conditions $I_j(X) = I_j^0$, $j = 1, \ldots, J$. Let $A(X) \in \mathbf{R}^{d \times d_T}$ denote the matrix whose columns are the gradients of the first integrals $I_j(X)$:

$$A(X) = (\nabla I_1, \cdots, \nabla I_J),$$

and assume $A$ has full column rank. Then for a given perturbation vector field $\tilde{g}(X)$, the projected vector field

$$g(X) = (I - A(A^T A)^{-1} A^T)\tilde{g}(X) \tag{4.22}$$

preserves the invariants $I_j$.

## 4.3 Semidiscrete PDE models

To illustrate the application of thermostats to PDEs, we select two related model problems, the inviscid Burgers-Hopf (BH) and Korteweg-De Vries (KdV) equations.

We choose these models as simple one-dimensional problems with features in common with more sophisticated fluid models, i.e. quadratic nonlinearity, multiple conserved quantities, a tendency to generate fine scale dynamics from smooth initial conditions, and slow (fast) decorrelation times for low (high) wave numbers.

The BH/KdV model (4.1) is discretized using a pseudospectral truncation (see Appendix 4.A), resulting in an equation of the form (cf. 4.A.9)

$$\frac{\partial u_N}{\partial t} + \frac{1}{2}\frac{\partial \mathcal{P}_N u_N^2}{\partial x} + \mu \frac{\partial^3 u_N}{\partial x^3} = 0, \tag{4.23}$$

where $u_N := \mathcal{P}_N u(x)$ is the projection of the function $u$ onto $N$ Fourier modes.

The truncated model retains as first integrals the discrete analogs of $\mathcal{M}$, $\mathcal{E}$ and $\mathcal{H}$, respectively [76]:

$$M = \int_0^{2\pi} u_N \,\mathrm{d}x, \tag{4.24}$$

$$E = \frac{1}{2}\int_0^{2\pi} u_N^2 \,\mathrm{d}x, \tag{4.25}$$

$$H = \int_0^{2\pi} \left( \frac{1}{6}u_N^3 - \frac{\mu}{2}\left(\frac{\partial u_N}{\partial x}\right)^2 \right) \mathrm{d}x. \tag{4.26}$$

### 4.3.1 Statistical mechanics of the truncated model

Abramov et al. [1] proposed a statistical mechanics for the pseudospectral truncation of the Burgers-Hopf equation, which carries over to the KdV equation. The spectral representation of (4.23) is (cf. 4.A.10)

$$\frac{\mathrm{d}\hat{u}_n}{\mathrm{d}t} = \hat{f}_n(\hat{u}) = -\frac{in}{2}\left( \sum_{|n-m|\le N} \hat{u}_{n-m}\hat{u}_m \right) + in^3\mu\hat{u}_n, \quad |n|,|m| \le N, \tag{4.27}$$

where $\hat{u}, \hat{f} \in \mathbf{C}^{2N+1}$, $\hat{u}_{-n} = \hat{u}_n^*$. First note that the vector field $\hat{f}(\hat{u})$ is divergence-free:

$$\nabla \cdot \hat{f}(\hat{u}) = 2\mathrm{Re} \sum_{|n|\le N} \frac{\partial \hat{f}_n}{\partial \hat{u}_n} = 2\mathrm{Re} \sum_{|n|\le N} \left(-in\hat{u}_0 + in^3\mu\right) = 0,$$

since $\hat{u}_0 \in \mathbf{R}$ for a real smooth $2\pi$-periodic function $u(x)$. This implies that an equilibrium density is a function of the conserved quantities

In [1] it is noted that a Gibbs-like density $\rho(\hat{u}) = \exp(-\beta H(\hat{u}) - \gamma E(\hat{u}))$ cannot be normalized due to the unboundedness of level sets of the highest order terms in $H$. Noting that level sets of $E$ are hyperspheres when $M = 0$, and hence compact, [1] instead propose a mixed ensemble that is microcanonical in $E$ and $M$, and canonical in $H$, i.e.

$$\rho(\hat{u}) \propto \exp(-\beta H(\hat{u}))\delta(E(\hat{u}) - E_0)\delta(M(\hat{u})).$$

We adopt this density here. Since the phase space is compact, the system supports both positive and negative regimes for the statistical temperature $\beta^{-1}$ [92].

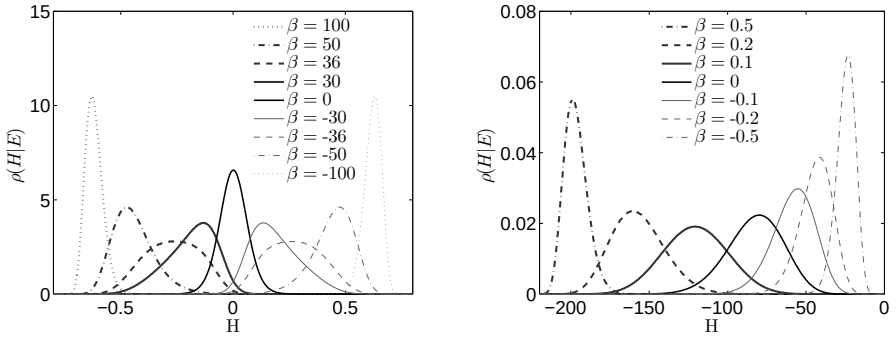



Figure 4.1: Probability density functions of the Hamiltonian $H(\hat{u})$ for different values of $\beta$, $E_0 = 1$. Left: BH equation. Right: KdV equation.

We used the Metropolis-Hastings algorithm to compute probability density functions of the Hamiltonian $H$ for $E_0 = 1$ and different values of $\beta$. The pdfs shown in Figure 4.1 were obtained using $10^8$ samples and $N = 15$. Because the phase space is compact the temperature assumes both positive and negative values. For the Burgers-Hopf equation we note that the skewness varies in a nonlinear way as a function of $\beta$, but that the pdfs are anti-symmetric with respect to $\beta = 0$. For the KdV equation the pdf with $\beta = 0$ has negative skewness and it changes to positive near the value $\beta = 0.1$.

In Figure 4.2 we plot expectation values of the kinetic energy spectrum $|\hat{u}_n|^2$ as a function of wave number $n$. Note that the energy is equipartitioned for $\beta = 0$ which corresponds to the case of a uniform distribution on the sphere $\delta(E(\hat{u}) - E_0)$. For the Burgers-Hopf equation we observe significant tilt in the spectrum for values $\beta \neq 0$. More energy resides in the large scales (small wave numbers). Furthermore, the spectra are identical for opposite signed $\beta$. For the KdV equation we observe opposite tilt in the spectrum depending on the sign of $\beta$, with more energy at low wave numbers for $\beta < 0$ and at high wave numbers for $\beta > 0$.

### 4.3.2 Ergodicity of stochastic hydrodynamics models

For the truncated incompressible Navier-Stokes equation, E & Mattingly [24] proved ergodicity under highly degenerate stochastic forcing of just two modes in the low wave number range, with viscous damping at the large wave number end of the spectrum. The proof of [24] requires establishing a Lyapunov function and verifying the Hörmander condition for the drift and diffusion vector fields.

In this chapter we use GBK thermostats [58] to effect a simple non-dissipative closure model, with forcing implemented at the small scales/large wave numbers. Because the thermostats control the flux of energy into and out of the system, they do not require a separate dissipation term to maintain stability. Here we illustrate through analysis that thermostatting the small scales (essentially through "backscatter") can be effective, i.e. we show the Hörmander condition for this type

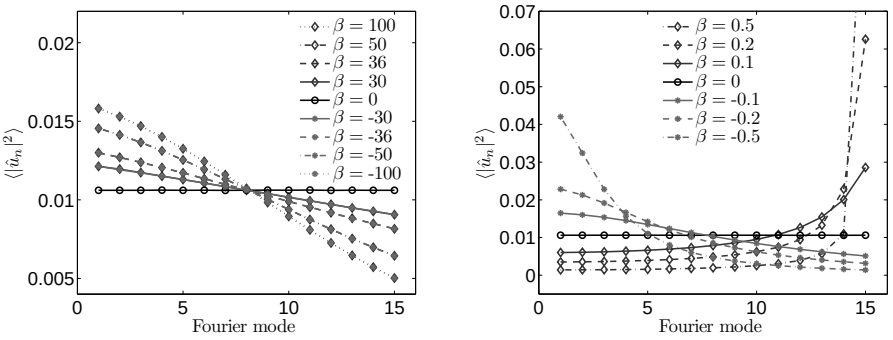



Figure 4.2: Mean kinetic energy spectrum, for different values of $\beta$. Left: BH equation. Right: KdV equation.

of forcing. Our starting point is the GBK method (4.12)–(4.13) on $\mathbf{C}^{2N+1} \times \mathbf{R}^{d_T}$.

The form of Lemma 4.2.3.1 suggests adapting the analysis of E and Mattingly to Burgers-Hopf equation, and forcing at large wave numbers. We next derive a suitable set of perturbation vector fields $\hat{g}_1, \ldots, \hat{g}_M$ that ensure Hörmander's condition.

The truncated BH/KdV model (4.27) is derived in Appendix 4.A. For the rest of this section we restrict our attention to the case $\mu = 0$ of Burgers-Hopf equation, because the formulas are simpler, and the dispersion term does not contribute to mixing between distinct wave numbers. Following [24], define $\hat{u}_n = a_n + ib_n$, and denote the unit vectors in the respective real coordinates by $\partial_{a_n}$ and $\partial_{b_n}$. Then

$$\begin{aligned}\hat{f}_n &= -\frac{in}{2} \sum_{|m-n|<N} (a_{n-m} + ib_{n-m})(a_m + ib_m) \\ &= \frac{n}{2} \sum_{|m-n|<N} (a_{n-m}b_m + a_m b_{n-m})\partial_{a_n} + (b_{n-m}b_m - a_{n-m}a_m)\partial_{b_n}.\end{aligned}$$

Since the solution $u(x,t)$ of the BH/KdV model is real valued, the Fourier modes $\hat{u}_n$ satisfy $\hat{u}_{-n} = \hat{u}_n^*$, which in turn implies the conditions $a_{-n} = a_n$ and $b_{-n} = -b_n$. We also assume $\hat{u}_0 \equiv \hat{f}_0 \equiv 0$.

Fixing $n > 0$ for the moment, define the index sets $N_n^+ = \{n+1, \ldots, N\}$ and $N_n^- = \{1, \ldots, n-1\}$, and note that

$$\begin{aligned} m \in N_n^+ \quad &\Rightarrow \quad m > 0, n - m < 0, \\ m \in N_n^- \quad &\Rightarrow \quad m > 0, n - m > 0, \\ m \in (n - N_n^+) \quad &\Rightarrow \quad m < 0, n - m > 0. \end{aligned}$$

With this in mind, the vector field $\hat{f}_n$ is written as

$$\begin{aligned}\hat{f}_n =& \frac{n}{2} \sum_{m \in N_n^-} (a_{n-m}b_m + a_m b_{n-m})\partial_{a_n} + (b_{n-m}b_m - a_{n-m}a_m)\partial_{b_n} \\ &+ \frac{n}{2} \sum_{m \in N_n^+} (a_{m-n}b_m - a_m b_{m-n})\partial_{a_n} + (-b_{m-n}b_m - a_{m-n}a_m)\partial_{b_n}\end{aligned}$$

$$+\frac{n}{2}\sum_{m\in(N_n^+-n)}(-a_{n+m}b_m+a_mb_{n+m})\partial_{a_n}+(-b_{n+m}b_m-a_{n+m}a_m)\partial_{b_n},$$

where now all indices are positive. Furthermore, it can be checked that the last two sums are equivalent, so the formula simplifies to

$$\begin{aligned}\hat{f}_n=&\frac{n}{2}\sum_{m\in N_n^-}(a_{n-m}b_m+a_mb_{n-m})\partial_{a_n}+(b_{n-m}b_m-a_{n-m}a_m)\partial_{b_n}\\&+n\sum_{m\in N_n^+}(a_{m-n}b_m-a_mb_{m-n})\partial_{a_n}+(-b_{m-n}b_m-a_{m-n}a_m)\partial_{b_n}.\end{aligned}$$

Next we compute commutators with the canonical unit vectors, for future reference. (These are the columns of the Jacobian matrix of $\hat{f}$.) We find:

$$\begin{aligned}X_\ell=[\hat{f},\partial_{a_\ell}]&=n(b_{n-\ell}-b_{\ell-n}+b_{\ell+n})\partial_{a_n}+n(-a_{n-\ell}-a_{\ell-n}-a_{\ell+n})\partial_{b_n},\\Y_\ell=[\hat{f},\partial_{b_\ell}]&=n(a_{n-\ell}+a_{\ell-n}-a_{\ell+n})\partial_{a_n}+n(b_{n-\ell}-b_{\ell-n}-b_{\ell+n})\partial_{b_n},\end{aligned}$$

where henceforth it is understood that the index of each term is either an element of the set $\{1,\dots,N\}$, or the term itself is neglected, meaning that each expression in parentheses above has at least one and at most two (when $\ell+m\le N$) nontrivial terms.

The commutators of $X_\ell$ and $Y_\ell$ with respect to generic unit vectors $\partial_{a_m}$ and $\partial_{b_m}$ are:

$$\begin{aligned}[X_\ell,\partial_{a_m}]&=n(-\delta_{n-\ell,m}-\delta_{\ell-n,m}-\delta_{\ell+n,m})\partial_{b_n}\\&=-(m+\ell)\partial_{b_{m+\ell}}-(\ell-m)\partial_{b_{\ell-m}}-(m-\ell)\partial_{b_{m-\ell}},\\[X_\ell,\partial_{b_m}]&=(m+\ell)\partial_{a_{m+\ell}}-(\ell-m)\partial_{a_{\ell-m}}+(m-\ell)\partial_{a_{m-\ell}},\\[Y_\ell,\partial_{a_m}]&=(m+\ell)\partial_{a_{m+\ell}}+(\ell-m)\partial_{a_{\ell-m}}-(m-\ell)\partial_{a_{m-\ell}},\\[Y_\ell,\partial_{b_m}]&=(m+\ell)\partial_{b_{m+\ell}}-(\ell-m)\partial_{b_{\ell-m}}-(m-\ell)\partial_{b_{m-\ell}},\end{aligned}$$

where $\delta_{m,\ell}$ is the Kronecker delta.

If the unit vector $\partial_{a_\ell}$ is an element of the Lie algebra $\mathcal{L}(\hat{f},\hat{g}_1,\dots,\hat{g}_K)$ (hereafter simply denoted by $\mathcal{L}$), then so is $X_\ell$. Likewise, inclusion of $\partial_{b_\ell}$ implies that of $Y_\ell$. As a result, we have the following inclusions:

$$\begin{aligned}\partial_{a_\ell},\partial_{a_m}\in\mathcal{L}&\Rightarrow(\partial_{b_{|\ell-m|}}\pm\partial_{b_{\ell+m}})\in\mathcal{L},\\\partial_{a_\ell},\partial_{b_m}\in\mathcal{L}&\Rightarrow(\partial_{a_{|\ell-m|}}\pm\partial_{a_{\ell+m}})\in\mathcal{L},\\\partial_{b_\ell},\partial_{b_m}\in\mathcal{L}&\Rightarrow(\partial_{b_{|\ell-m|}}\pm\partial_{b_{\ell+m}})\in\mathcal{L},\end{aligned}$$

where the second term on the right in each relation is present only if $\ell+m\le N$. From the last of these three recursions, it immediately follows that if $\partial_{b_1}$ is in $\mathcal{L}$, so are all of the $\partial_{b_\ell}$. If additionally $\partial_{a_1}\in\mathcal{L}$, then from the second implication above, all of the $\partial_{a_\ell}$ also follow, and the Hörmander condition is satisfied. Hence, to demonstrate the Hörmander condition, it suffices to thermostat only the lowest wave number, taking $\hat{g}_1=\partial_{a_1}$, $\hat{g}_2=\partial_{b_1}$.

On the other hand, directly thermostatting the low wave numbers is likely to be intrusive in the dynamics. Instead we wish to thermostat the highest wave numbers, which constitute an uncertain component in the solution anyway. If $\partial_{a_N}$ and $\partial_{a_{N-1}}$ are in $\mathcal{L}$, then we obtain $\partial_{b_1} \in \mathcal{L}$ from the commutator $[X_N, \partial_{a_{N-1}}]$, and subsequently all of the $\partial_{b_\ell}$ and associated $Y_\ell$. Finally, the commutator $[Y_N, \partial_{a_{N-1}}]$ yields $\partial_{a_1} \in \mathcal{L}$, subsequently all of the $\partial_{a_\ell}$, and Hörmander is again satisfied. Therefore, we can construct a GBK thermostat satisfying Hörmander's condition for Burgers-Hopf equation and perturbations only to the real parts of the two highest wave numbers, taking $\hat{g}_1 = \partial_{a_N}$, $\hat{g}_2 = \partial_{a_{N-1}}$. Combining this with a Lyapunov function would ensure ergodicity in the measure $\rho(X) = \exp(-\beta E(X))$, where $E$ is the quadratic invariant of BH/KdV.

The above approach will not allow sampling of a mixed measure (4.11), however, since the perturbation vector fields so defined do not lie in the tangent bundle to the hypersphere of constant kinetic energy $E$. Instead, we may choose a single perturbation vector field $\hat{g}$ that is a rotation about one or more coordinate axes, for example,

$$\hat{g} = b_N \partial_{a_N} - a_N \partial_{b_N}. \tag{4.28}$$

Since both $\hat{f}$ and $\hat{g}$ are defined in the tangent space to the manifold of constant $E$, the Lie algebra generated by these vectors also preserves the first integral. The phase space is compact, and ergodicity follows from the Hörmander condition. However, with quadratic $\hat{f}$ and linear $\hat{g}$, the commutators are all quadratic or higher in order, making this condition difficult to check. Instead we include numerical experiments to assess ergodicity.

### 4.3.3 Thermostated dynamics for the semidiscrete model

In this section we specify the thermostated dynamics in the context of the truncated BH/KdV equation (4.23) and the mixed canonical distribution (4.11). The GBK thermostat for equation (4.23) and a single thermostat variable $\xi$ is:

$$\mathrm{d}u_N = f_N(u_N)\,\mathrm{d}t + \xi g_N(u_N)\,\mathrm{d}t, \tag{4.29}$$

$$\mathrm{d}\xi = 2\,\mathrm{Re}\,h(u_N)\,\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w, \tag{4.30}$$

where $f_N(u_N) = -\frac{1}{2}\frac{\partial \mathcal{P}_N(u_N^2)}{\partial x} - \mu\frac{\partial^3 u_N}{\partial x^3}$. The function $g(u_N)$ is chosen such that its projection $g_N(u_N) := \mathcal{P}_N g(u_N)$ satisfies the constraints

$$\int_0^{2\pi} \frac{\delta M}{\delta u_N} g_N(u_N)\,\mathrm{d}x = 0, \qquad \int_0^{2\pi} \frac{\delta E}{\delta u_N} g_N(u_N)\,\mathrm{d}x = 0. \tag{4.31}$$

These relations constrain the dynamics to the Dirac distributions on $M$ and $E$. Taking into account that $\mathcal{P}_N$ is symmetric, the constraints (4.31) reduce to

$$\int_0^{2\pi} g(u_N)\,\mathrm{d}x = 0, \qquad \int_0^{2\pi} u_N g(u_N)\,\mathrm{d}x = 0. \tag{4.32}$$

Without loss of generality one may assume $M = 0$, since the nonzero case may be handled with a change of variables. Furthermore, in spectral representation this

condition takes the simple form $\hat{u}_0 \equiv 0$, which can be easily enforced by simply neglecting the constant mode in the spectral representation, taking $\hat{u} = (\hat{u}_n; 1 \leq |n| \leq N)$.

To preserve the kinetic energy constraint we either choose $g(u_N)$ to respect the rotation symmetry, as in (4.28), or use a projection such as (4.22).

For a given function $\tilde{g}(u_N)$, using the definitions (4.24) and (4.25) of $M$ and $E$, and taking $M = 0$, we observe that the function

$$g(u_N) = \tilde{g}(u_N) - \frac{1}{2\pi}\int_0^{2\pi} \tilde{g}(u_N)\,\mathrm{d}x - \frac{1}{2E}u_N \int_0^{2\pi} u_N \tilde{g}(u_N)\,\mathrm{d}x \tag{4.33}$$

satisfies the constraints (4.32). In spectral representation, $\hat{g}_0 = 0$ and

$$\hat{g}_n(\hat{u}) = \frac{1}{2\pi}\int_0^{2\pi} g(u_N) e^{-inx}\,\mathrm{d}x, \quad 1 \leq |n| \leq N.$$

Using (4.19) we compute $h(\hat{u})$ from

$$h(\hat{u}) = \frac{1}{\alpha}\left(\nabla_{\hat{u}} \cdot \hat{g}(\hat{u}) - \beta \nabla_{\hat{u}} H(\hat{u}) \cdot \hat{g}(\hat{u})\right).$$

The gradient of the Hamiltonian $H(\hat{u})$ is

$$\frac{\partial H(\hat{u})}{\partial \hat{u}_n} = \int_0^{2\pi} \frac{\delta H}{\delta u_N}\frac{\partial u_N}{\partial \hat{u}_n}\,\mathrm{d}x = \int_0^{2\pi} \frac{\delta H}{\delta u_N} e^{inx}\,\mathrm{d}x = 2\pi \left(\widehat{\frac{\delta H}{\delta u_N}}\right)^*_n, \quad 1 \leq |n| \leq N,$$

which yields

$$\begin{aligned}\nabla_{\hat{u}} H(\hat{u}) \cdot \hat{g}(\hat{u}) &= \sum_{1 \leq |n| \leq N} \frac{\partial H(\hat{u})}{\partial \hat{u}_n}\hat{g}_n(\hat{u}) \\ &= 2\pi \sum_{1 \leq |n| \leq N} \left(\widehat{\frac{\delta H}{\delta u_N}}\right)^*_n \hat{g}_n(\hat{u}) = \int_0^{2\pi} \frac{\delta H}{\delta u_N} g_N(u_N)\,\mathrm{d}x.\end{aligned}$$

The spectral representation of $h(u_N)$ follows:

$$h(\hat{u}) = \frac{1}{\alpha}\left(\nabla_{\hat{u}} \cdot \hat{g}(\hat{u}) - 2\pi\beta \sum_{1 \leq |n| \leq N} \left(\widehat{\frac{\delta H}{\delta u_N}}\right)^*_n \hat{g}_n(\hat{u})\right).$$

For each value of $1 \leq |n| \leq N$:

$$\frac{\partial \hat{g}_n(\hat{u})}{\partial \hat{u}_n} = \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial g(u_N)}{\partial u_N}\frac{\partial u_N}{\partial \hat{u}_n} e^{-inx}\,\mathrm{d}x = \frac{1}{2\pi}\int_0^{2\pi} \frac{\partial g(u_N)}{\partial u_N}\,\mathrm{d}x.$$

This gives us

$$h(u_N) = \frac{1}{\alpha}\left(\frac{N}{\pi}\int_0^{2\pi} \frac{\partial \tilde{g}(u_N)}{\partial u_N}\,\mathrm{d}x\ - \frac{N}{E}\int_0^{2\pi} u_N \tilde{g}(u_N)\,\mathrm{d}x - \beta \int_0^{2\pi} \frac{\delta H}{\delta u_N} g_N(u_N)\,\mathrm{d}x\right).$$

## 4.4 Numerical study

We rely on a series of numerical simulations to test the performance of the thermostats mentioned above in the setting of the Burgers-Hopf and KdV equations. Our interest here is in two crucial issues: (i) the ergodic nature of the extended SDE models, even under limited contact with the stochastic heat bath, and (ii) the degree to which thermodynamic corrections alter dynamic observables (e.g. temporal correlation functions). In evaluating the experimental results, we use the terminology from Subsection 4.2.1 and explicitly define the autocorrelation functions of the real part of the Fourier modes, i.e.

$$c_n(\tau) = C \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathrm{Re}\left\{\hat{u}_n(t+\tau)\right\} \mathrm{Re}\left\{\hat{u}_n(t)\right\} \,\mathrm{d}t, \quad n = 1, \ldots, N, \tag{4.34}$$

where $C$ is a suitable normalization constant so that $c_n(0) = 1$.

### 4.4.1 Thermostated Burgers-Hopf equation

In our computations we set $N = 15$, $E_0 = 1$ and $\beta = -30$. We solve equations (4.29)–(4.30) in time by applying a Strang splitting method, thus dividing the calculation into two steps: (1) the solution of the equation for the auxiliary variable, and (2) the solution of the equations governing Fourier coefficients of the solution. The stochastic differential equation for the auxiliary variable can be solved exactly when $u_N$ is fixed, whereas in step (2) the system for $u_N$ is treated using the implicit midpoint rule (a scheme which preserves quadratic first integrals, i.e. the hypersphere).

The numerical method is

$$\begin{aligned}
\xi^* &= e^{-\gamma\frac{\tau}{2}}\xi^0 + \frac{2\mathrm{Re}h(u_N^0)}{\gamma}\left(1 - e^{-\gamma\frac{\tau}{2}}\right) + \sigma\sqrt{\frac{1-e^{-\gamma\tau}}{2\gamma}}\,\Delta w^0,\\
u_N^1 &= u_N^0 + \tau f_N(u_N^{1/2}) + \tau\xi^* g_N(u_N^{1/2}), \quad u_N^{1/2} := \frac{u_N^1 + u_N^0}{2},\\
\xi^1 &= e^{-\gamma\frac{\tau}{2}}\xi^* + \frac{2\mathrm{Re}h(u_N^1)}{\gamma}\left(1 - e^{-\gamma\frac{\tau}{2}}\right) + \sigma\sqrt{\frac{1-e^{-\gamma\tau}}{2\gamma}}\,\Delta w^1,
\end{aligned}$$

where $\Delta w^0, \Delta w^1 \sim \mathcal{N}(0,1)$ and $\tau$ is a time step.

The first question concerns the ergodic sampling of the target distribution. This can depend on the choice of $g$. We let $\tilde{g}(u_N) = u_N^2$. With this particular choice of function $\tilde{g}(u_N)$ from expression (4.33) we find

$$g(u_N) = u_N^2 - \frac{1}{2\pi}\int_0^{2\pi} u_N^2 \,\mathrm{d}x - \frac{1}{2E} u_N \int_0^{2\pi} u_N^3 \,\mathrm{d}x$$

and compute

$$h(u_N) = -\frac{1}{\alpha}\left(\frac{N}{E}\int_0^{2\pi} u_N^3 \,\mathrm{d}x + \beta \int_0^{2\pi} \frac{\delta H}{\delta u_N} g_N(u_N)\,\mathrm{d}x\right).$$

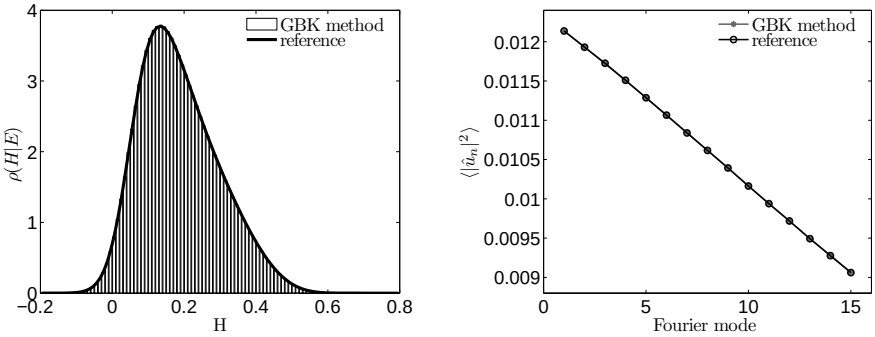



Figure 4.3: Gentle thermostatting of BH equation with GBK method, taking $\tilde{g}(u_N) = u_N^2$, $\alpha = 30$ and $\gamma = 20$. Left: probability density function of $H(\hat{u})$. Right: mean kinetic energy spectrum.

Numerical results are presented in Figure 4.3. In the computations we used $10^9$ data points and $\tau = 0.001$. In Figure 4.3 we compare the numerically computed histogram of $H(\hat{u})$ and the spectrum to the Monte Carlo simulations using the Metropolis-Hastings algorithm. Since a single trajectory produces what is essentially a perfect Hamiltonian pdf and spectrum we infer that the method is ergodic.

We then set about constructing a thermostat that controls the invariant measure using only forcing at high wave numbers. To this end we work with a spectral representation of $g(u_N)$. For any skew-Hermitian matrix $B(\hat{u})$ the vector field

$$\hat{g}(\hat{u}) = B(\hat{u})\hat{u}$$

is norm preserving and therefore retains the first integral $E$. We choose matrix $B$ such that it only acts on the large wave number Fourier coefficients, i.e.

$$\hat{g}(\hat{u})_n = \begin{cases} 0, & |n| < n^*, \\ i\,\mathrm{sign}(n)\hat{u}_n, & |n| \geq n^*, \end{cases} \tag{4.35}$$

and refer to this method as GBK($n^*$). In this case the effect of the perturbation is to directly modify the phase of only the $(N-n^*+1)$ highest Fourier modes. This can be contrasted directly with the approach of E & Mattingly [24], who stochastically force the lowest modes of a truncated Navier-Stokes model using a Langevin approach. Here we thermostat at the finest scales, effectively controlling the measure through backscatter.

The results for $n^* = 11$ are shown in Figure 4.4 using $10^9$ data points. All computations are done with $\tau = 0.01$. These results again suggest that the method is ergodic. Not only can we get away with thermostatting directly only the highest five wave numbers, it is in fact possible to control the distribution using only a *single mode*. In Figure 4.5, the pdfs are shown for the real parts of the Fourier coefficients 1, 5, 10 and 15 when only the highest wave number $\hat{u}_{15}$ is directly coupled to the stochastic auxiliary variable $\xi$. Note that while $E_0 = 1$, the dynamics is constrained

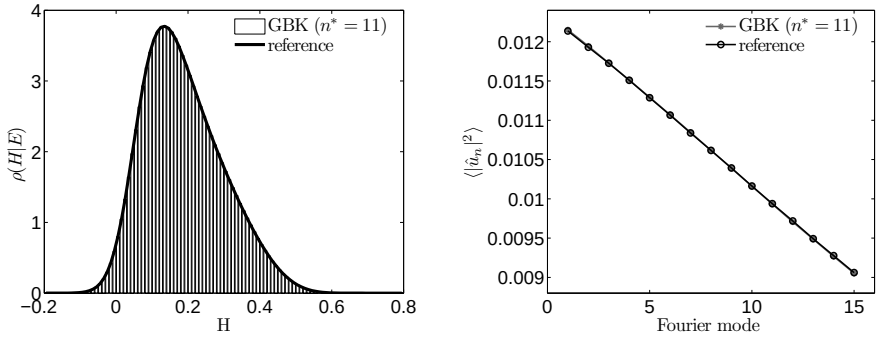



Figure 4.4: Gentle thermostatting of BH equation with GBK($n^* = 11$) method, $\alpha = \gamma = 1$. Left: probability density function of $H(\hat{u})$. Right: mean kinetic energy spectrum.

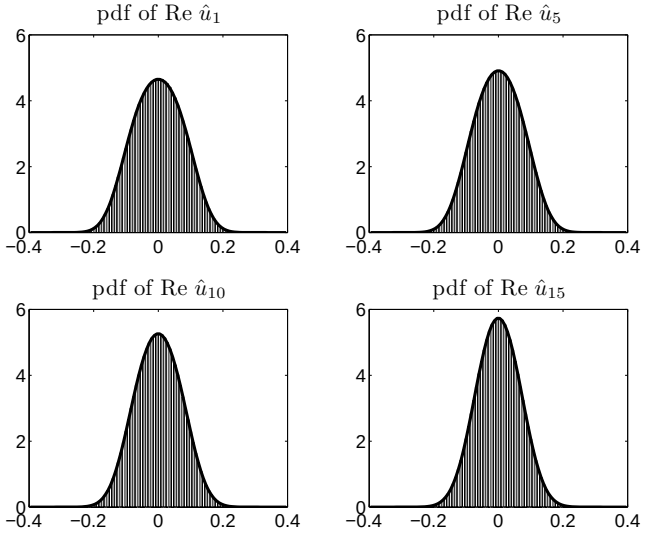



Figure 4.5: Probability density function of real parts of Fourier coefficients 1, 5, 10 and 15. GBK($n^* = 15$) thermostat method compared to reference.

to the hypersphere with radius $1/\sqrt{2\pi} \approx 0.4$, and this number bounds the support of the pdfs.

Thermostatting only the high wave numbers leads to a reduced rate of convergence of averages compared to a thermostat that acts directly on all components. This effect can be seen in Figure 4.6. Slope values are approximate. When plotted

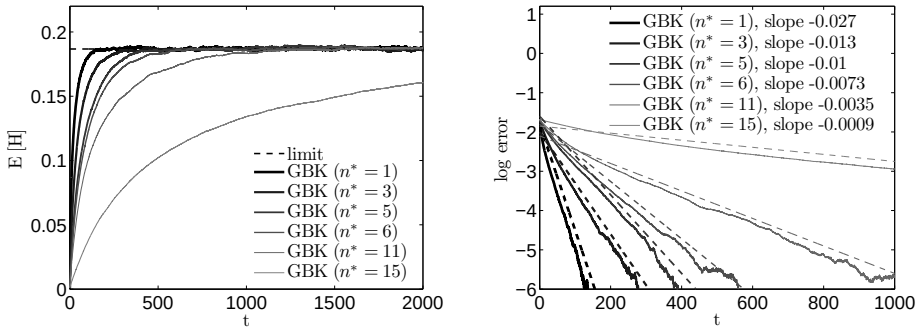



Figure 4.6: Convergence of the expected value of Hamiltonian for an ensemble of 20 000 initial conditions, $\alpha = \gamma = 1$.

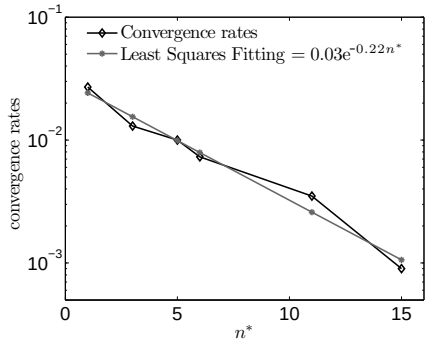



Figure 4.7: Convergence rates from Figure 4.6 as a function of $n^*$ and least squares fitting to the exponential function.

as a function of $n^*$ and compared to the least square fitted exponential function in Figure 4.7 on a logarithmic scale we observe good agreement. This suggests that the rate of convergence decreases exponentially with respect to $n^*$.

On the other hand, although the convergence rate of averages may be reduced by using a weak thermostat, the perturbation of slow dynamics is simultaneously reduced, meaning that where the dynamics of slow variables is relevant, these methods are likely to be of greatest value. As we noted in the introduction, the advantage of the GBK thermostat over direct Langevin thermostatting is that the stochastic forcing only influences the original dynamics after integration—as a memory or red noise term—leading to a second order perturbation of autocorrelation functions *of the fast modes* $\hat{u}_n$, $n \geq n^*$. In fact, a straightforward calculations shows that we would expect only third order or higher perturbations to autocorrelations of the slow modes $\hat{u}_n$, $n < n^*$, which are not directly thermostated.

In Figure 4.8(a) we plot the $L^2$ error of the pdf of the Hamiltonian as a function

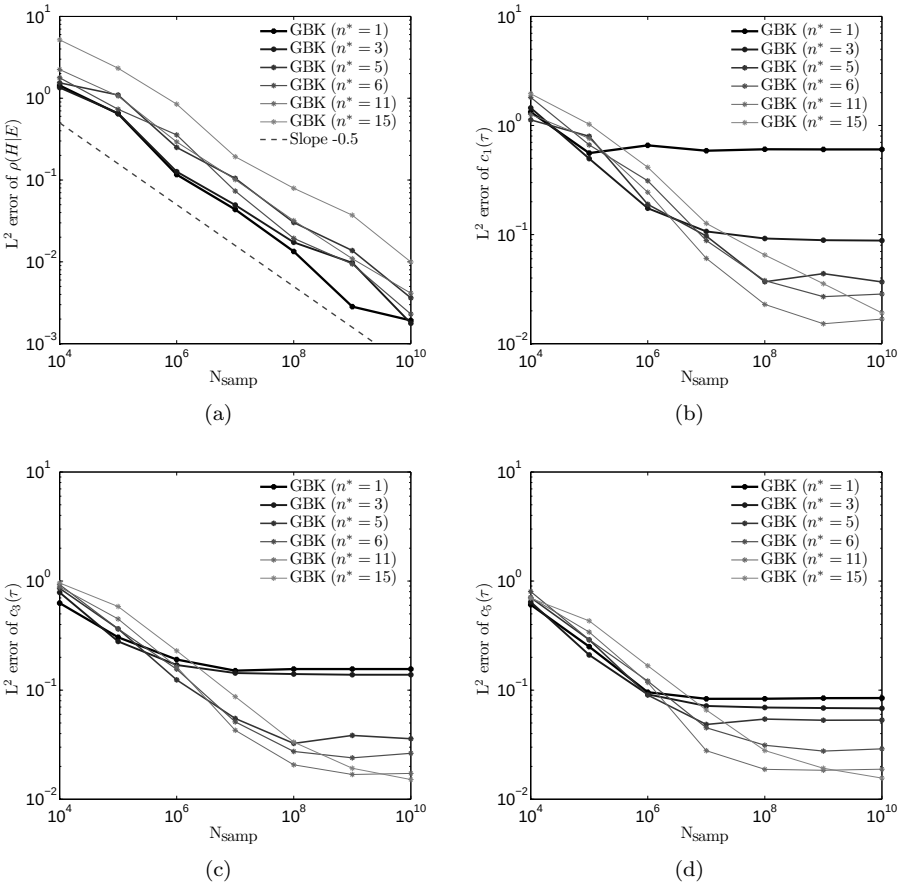
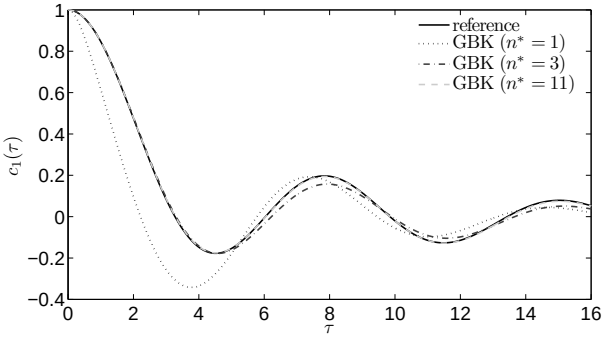


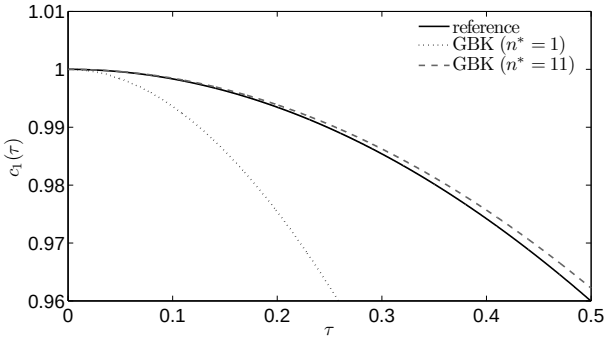

Figure 4.8: $L^2$ errors as a function of sampling time with small scale forcing (4.35) of wave numbers $n \geq n^*$. Top left: evolving pdf of the Hamiltonian, Top right and bottom: autocorrelation functions for $\mathrm{Re}\,\hat{u}_1$, $\mathrm{Re}\,\hat{u}_3$ and $\mathrm{Re}\,\hat{u}_5$ in the evolving pdf of the Hamiltonian, $\alpha = \gamma = 1$.

of sampling time, showing the convergence to the reference pdf. We observe the expected sampling convergence rate, 1/2. In Figures 4.8(b), 4.8(c) and 4.8(d) we plot $L^2$ errors, computed on the interval $\tau \in [0, 50]$, as a function of sampling time of the autocorrelation functions $c_1(\tau)$, $c_3(\tau)$ and $c_5(\tau)$, respectively. Observe that the graphs level off indicating a convergence to a limiting value of the net perturbation. (To see that the graph for the method GBK($n^* = 15$) eventually stabilizes, we would have to integrate even longer in time.) Complementary to Figure 4.8 we plot in Figures 4.9 and 4.10 the same autocorrelation functions and visually compare them to reference curves. The reference curves are computed using constant Hamiltonian simulations from a mixed canonically distributed ensemble of $10^6$ initial conditions.

Note the big differences in errors between GBK($n^* = 1$) and the others in Fig-




(a)

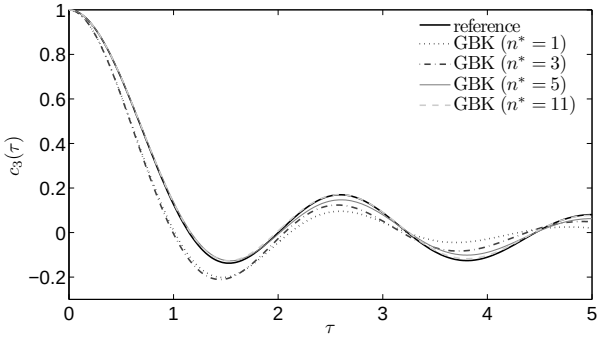



(b)

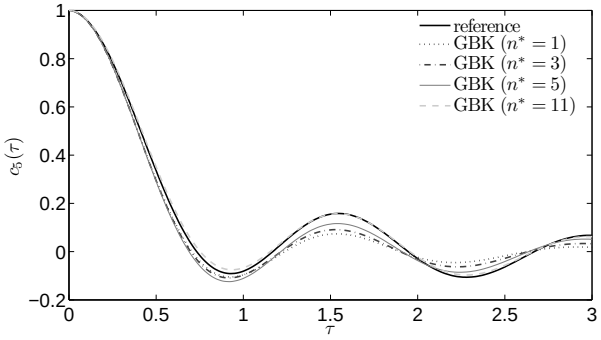
Figure 4.9: Gentle thermostatting of BH equation with GBK($n^*$) method, $\alpha = \gamma = 1$. (a) autocorrelation function $c_1(\tau)$. (b) autocorrelation function $c_1(\tau)$ for small correlation times $\tau$. The reference curves were computed using constant Hamiltonian simulations from a mixed canonically distributed ensemble of $10^6$ initial conditions.

ure 4.8(b). Only in the case of GBK($n^* = 1$) is the equation for $\mathrm{Re}\,\hat{u}_1$ directly perturbed. This leads to the second order perturbation of autocorrelation function while the other methods, GBK($n^* > 1$), lead to third order or higher perturbations of autocorrelation functions. This can be seen in Figure 4.9(b) where we compare the autocorrelation functions $c_1(\tau)$ for two methods, GBK($n^* = 1$) and GBK($n^* = 11$), with the reference curve for the small correlation times $\tau$. It is clearly evident that the method GBK($n^* = 11$) is more accurate than GBK($n^* = 1$).

Interestingly errors in autocorrelation functions also depend on the value of $n^*$. For larger value of $n^*$ we observe smaller errors, see Figures 4.8(b), 4.8(c) and 4.8(d). In Figure 4.9(a) it is easy to see that two methods GBK($n^* = 3$) and GBK($n^* = 11$), which do not directly perturb the equation for $\mathrm{Re}\,\hat{u}_1$, have better autocorrelation functions compared to method GBK($n^* = 1$). But it is also notable that the more gentle method, i.e. the GBK($n^*$) method with larger value of $n^*$, has the more accurate autocorrelation function.




(a)




(b)

Figure 4.10: Gentle thermostatting of BH equation with GBK($n^*$) method, $\alpha = \gamma = 1$. (a) autocorrelation function $c_3(\tau)$. (b) autocorrelation function $c_5(\tau)$. The reference curves were computed using constant Hamiltonian simulations from a mixed canonically distributed ensemble of $10^6$ initial conditions.

Similar effects of perturbation order to autocorrelation functions can be seen in Figures 4.8(c) and 4.8(b). In both methods, GBK($n^* = 1$) and GBK($n^* = 3$), the thermostat variable $\xi$ is directly coupled to the equation for $\mathrm{Re}\,\hat{u}_3$. This gives larger errors compared to the other methods, GBK($n^* > 3$), as seen in Figure 4.8(c). We observe similar trends in the results in Figure 4.8(d). And these translate also to Figures 4.10(a) and 4.10(b).

Numerical results presented in Figures 4.8, 4.9 and 4.10 show that the direct coupling to the thermostat $\xi$ in the equations for the slow modes can significantly effect the errors in autocorrelation functions of these modes and vice versa. Since errors in autocorrelation functions decrease with larger value of $n^*$ while, at the same time the convergence rate decreases with larger value of $n^*$ (Figure 4.6), it suggests seeking $n^*$ to obtain the optimal trade-off between rate of convergence and rate of perturbation to dynamics. (Of course the choice would also depend on the goal of simulation.)

### 4.4.2 Thermostated KdV equation

In the case of the KdV equation we note that the Hamiltonian function $H(\hat{u})$ can be written as the sum of two parts $H = H_1 + H_2$, where

$$H_1(u_N) = \int_0^{2\pi} u_N^3 \,\mathrm{d}x, \qquad H_2(u_N) = -\frac{1}{2}\int_0^{2\pi} \left(\frac{\partial u_N}{\partial x}\right)^2 \,\mathrm{d}x.$$

The Hamiltonian systems generated by $H_1(u_N)$ and $H_2(u_N)$ each conserves the first integrals $M$ and $E$ individually. Now we consider the following GBK thermostated KdV equation:

$$\begin{aligned}
\frac{\partial}{\partial t}u_N &= -\frac{\partial}{\partial x}\mathcal{P}_N\frac{\delta H}{\delta u_N} - \xi\frac{\partial}{\partial x}\mathcal{P}_N\frac{\delta H_2}{\delta u_N} = -\frac{\partial}{\partial x}\mathcal{P}_N\frac{\delta H_1}{\delta u_N} - (1+\xi)\frac{\partial}{\partial x}\mathcal{P}_N\frac{\delta H_2}{\delta u_N},\\
\mathrm{d}\xi &= 2\frac{\beta}{\alpha}\mathrm{Re}\int_0^{2\pi}\frac{\delta H_1}{\delta u_N}\frac{\partial}{\partial x}\mathcal{P}_N\frac{\delta H_2}{\delta u_N}\,\mathrm{d}x\,\mathrm{d}t - \gamma\xi\,\mathrm{d}t + \sigma\,\mathrm{d}w,
\end{aligned}$$

where $\mathrm{d}w$ is scalar Wiener process. This approach was suggested in a slightly different form (and for finite dimensional systems only) in [58] and is referred to as a force-perturbation thermostat since it perturbs the 'natural forces' of the system, or rather the balance between these, to realize a thermal control. For the KdV equation we effectively thermostat the system by controlling the strength and direction of dispersion.

To integrate the dynamics numerically in time, we use the following splitting method, which generates a map $(u_N^0, \xi^0) \mapsto (u_N^1, \xi^1)$ with time step $\tau$:

$$\begin{aligned}
\xi^{1/2} &= e^{-\gamma\frac{\tau}{2}}\xi^0 + \frac{2}{\gamma}\mathrm{Re}\,h(u_N^0)\left(1 - e^{-\gamma\frac{\tau}{2}}\right) + \sigma\sqrt{\frac{1-e^{-\gamma\tau}}{2\gamma}}\,\Delta w^0,\\
u_N^* &= e^{-(1+\xi^{1/2})\frac{\tau}{2}\frac{\partial^3}{\partial x^3}}u_N^0,\\
u_N^{**} &= u_N^* + \tau f_N(u_N^{1/2}), \qquad u_N^{1/2} := \frac{1}{2}(u_N^* + u_N^{**}),\\
u_N^1 &= e^{-(1+\xi^{1/2})\frac{\tau}{2}\frac{\partial^3}{\partial x^3}}u_N^{**},\\
\xi^1 &= e^{-\gamma\frac{\tau}{2}}\xi^{1/2} + \frac{2}{\gamma}\mathrm{Re}\,h(u_N^1)\left(1 - e^{-\gamma\frac{\tau}{2}}\right) + \sigma\sqrt{\frac{1-e^{-\gamma\tau}}{2\gamma}}\,\Delta w^1,
\end{aligned}$$

where $\Delta w^0, \Delta w^1 \sim \mathcal{N}(0,1)$. Numerical results with $\tau = 0.001$ are presented in Figures 4.11–4.12 (showing, in Figure 4.11, the convergence of the Hamiltonian probability density function and the spectrum, and, in Figure 4.12, the autocorrelation functions). $10^9$ data points were used to compute the graphs in Figure 4.11 and 4.12.

Results from Figure 4.11 demonstrate that a single trajectory produces what is essentially a perfect Hamiltonian pdf and spectrum, hence we infer that the method is ergodic. On the other hand Figure 4.12 shows that, for the particular values of $\alpha$ and $\gamma$, the GBK method has only a small impact on dynamics as measured by autocorrelation functions.

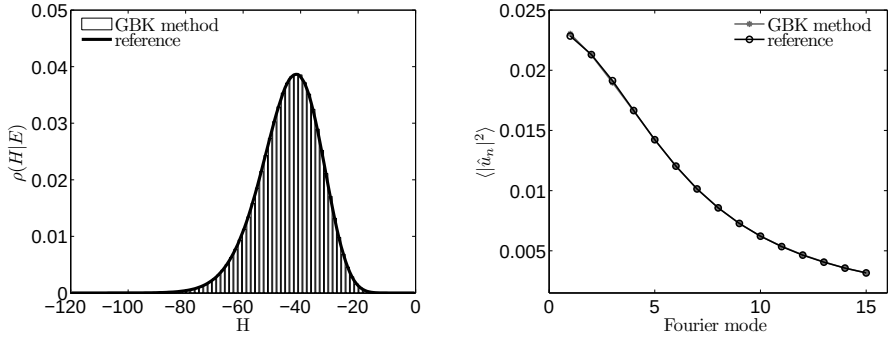



Figure 4.11: Gentle thermostatting of KdV equation with GBK method, $\alpha = 15$ and $\gamma = 40$. Left: probability density function of $H(\hat{u})$. Right: mean kinetic energy spectrum.

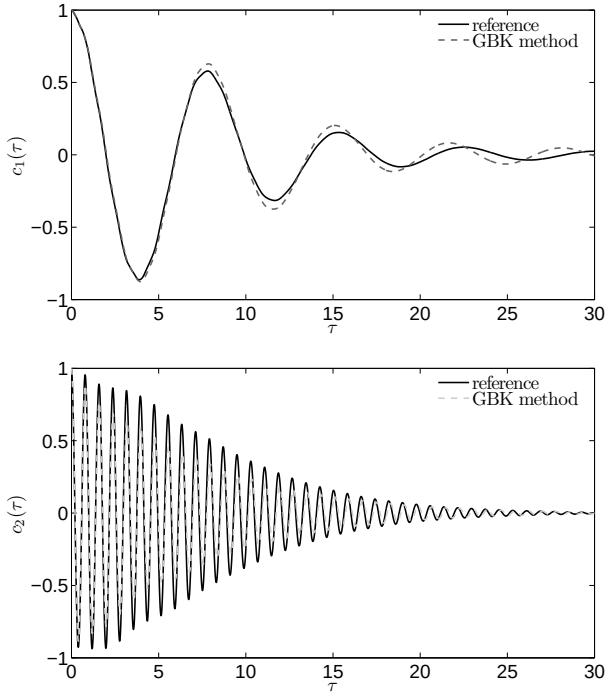



Figure 4.12: Gentle thermostatting of KdV equation with GBK method, $\alpha = 15$ and $\gamma = 40$. Top: autocorrelation function $c_1(\tau)$. Bottom: autocorrelation function $c_2(\tau)$. The reference curves were computed using constant Hamiltonian simulations from a mixed canonically distributed ensemble of $10^6$ initial conditions.

## 4.5 Conclusions

In this chapter we have outlined a framework for controlling the invariant measures of discretized PDEs, via coupling to one or more thermostat variables. Ergodicity is guaranteed if the perturbation vector fields satisfy Hörmander's condition.

The convergence rate of the time averages to the ensemble average in the desired measure depends on the strength of the thermostat, which can be controlled either through method parameters, or through choosing the degree of coupling between dynamical and thermostat wave numbers. The strength of the thermostat must be weighed against the degree of perturbation of dynamical quantities such as correlations.

For the example of the Burgers-Hopf equation, we have demonstrated that effective sampling can be achieved with perturbation only to the Fourier mode with largest wave number, providing a simple model of energetic exchange with unresolved modes. In this way the invariant measure is controlled by perturbing the least accurate component of the solution, and without introducing explicit viscous terms, which might suppress an inverse cascade. A test for this framework will come in extending the results to the 2D Euler equations, which is the subject of current work.

## 4.A Burgers-Hopf/Korteweg-de Vries model

The BH and KdV equations can be written in unified form

$$u_t + uu_x + \mu u_{xxx} = 0, \tag{4.A.1}$$

where the dispersion constant $\mu$ is zero for the BH equation and nonzero for the KdV equation. The classical KdV equation is obtained for $\mu = 1/6$ by rescaling time by the same factor. The BH and KdV equation are Hamiltonian PDEs with similar structure, as we briefly review in the next subsection.

### 4.A.1 Hamiltonian structure and conserved quantities

We consider Hamiltonian PDEs on a function space $\mathcal{U}$ of smooth, $2\pi$-periodic functions equipped with an inner product. The Poisson bracket (4.3) generalizes to an integral

$$\{\mathcal{F}, \mathcal{G}\} := \int_0^{2\pi} \frac{\delta \mathcal{F}}{\delta u} \mathcal{J} \frac{\delta \mathcal{G}}{\delta u} \, \mathrm{d}x, \qquad \mathcal{F}, \mathcal{G} : \mathcal{U} \to \mathbf{R},$$

i.e. a skew-symmetric, bilinear form acting on functionals on $\mathcal{U}$ and satisfying the Jacobi identity [91]. Here, $u(x,t) \in \mathcal{U}$ is a (possibly vector-valued) function, $\frac{\delta}{\delta u}$ denotes the variational derivative with respect to $u$, and $\mathcal{J}$ is a (matrix) differential operator, skew-symmetric with respect to the inner product on $\mathcal{U}$. The Hamiltonian PDE is given by

$$\frac{\partial u}{\partial t} = \{u, \mathcal{H}\}. \tag{4.A.2}$$

Hence the evolution of any functional $\mathcal{F}$ under the dynamics of a Hamiltonian PDE (4.A.2) obeys the equation

$$\frac{\partial \mathcal{F}}{\partial t} = \{\mathcal{F}, \mathcal{H}\}.$$

A functional $\mathcal{I}$ satisfying $\{\mathcal{I}, \mathcal{H}\} = 0$ is a first integral and constant along classical solutions to (4.A.2), as long as these exist.

Equation (4.A.1) has a bi-Hamiltonian structure [91] and is therefore integrable. The Hamiltonian structure we will use here is defined by

$$\mathcal{J} = -\frac{\partial}{\partial x}, \qquad \mathcal{H} = \int_0^{2\pi} \left( \frac{1}{6} u^3 - \frac{\mu}{2} \left( \frac{\partial u}{\partial x} \right)^2 \right) \mathrm{d}x. \tag{4.A.3}$$

Hence the corresponding Poisson bracket is

$$\{\mathcal{F}, \mathcal{G}\} := -\int_0^{2\pi} \frac{\delta \mathcal{F}}{\delta u} \frac{\partial}{\partial x} \frac{\delta \mathcal{G}}{\delta u} \, \mathrm{d}x. \tag{4.A.4}$$

One conserved quantity of (4.A.2) is the linear momentum

$$\mathcal{M} = \int_0^{2\pi} u \, \mathrm{d}x, \tag{4.A.5}$$

which we can assume to be zero up to Galilean change of coordinates.

Equation (4.A.1) has infinitely many conserved quantities, as mentioned above, depending on the value of $\mu$. For the BH equation ($\mu = 0$), the integral of any function of $u$ is conserved, and in particular the moments

$$\mathcal{I} = \int_0^{2\pi} u^p \, \mathrm{d}x, \quad p = 1, 2, \ldots \tag{4.A.6}$$

The first of these is the momentum mentioned earlier and assumed to be zero. The second moment represents the kinetic energy

$$\mathcal{E} = \frac{1}{2} \int_0^{2\pi} u^2 \, \mathrm{d}x. \tag{4.A.7}$$

The third moment is the Hamiltonian $\int u^3$, i.e. (4.A.3) with $\mu = 0$.

For the KdV equation, the first integrals of the infinite class are given by

$$\mathcal{I}_n = \int_0^{2\pi} P_{2n-1} \left( u, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}, \ldots \right) \mathrm{d}x, \quad n = 1, 2, \ldots,$$

where the polynomials $P_n$ are defined recursively [86] by

$$\begin{aligned} P_1 &= u, \\ P_n &= -\frac{\partial}{\partial x} P_{n-1} + \sum_{m=1}^{n-2} P_m P_{n-m-1}, \quad n \geq 2. \end{aligned}$$

The even-indexed polynomials $P_{2n}$ are exact differentials and thus trivially preserved. The polynomial $P_1(u)$ corresponds to momentum (4.A.5), $P_3(u)/2$ to the kinetic energy (4.A.7) and $P_5(u)/2$ to the Hamiltonian functional (4.A.3) with $\mu = 1/6$. Hence, all three functionals (4.A.5), (4.A.7) and (4.A.3) are conserved quantities of the equation (4.A.1) for any value of $\mu$.

### 4.A.2 Spectral truncation

As noted by McLachlan [81], the Hamiltonian structure of a PDE can often be retained in a finite dimensional truncation, by taking care to discretize the Poisson bracket and Hamiltonian separately. The Poisson bracket should be truncated such that remains skew-symmetric and, when nonlinear, satisfies the Jacobi identity. The Hamiltonian can be approximated by any consistent finite dimensional truncation. Majda & Timofeyev [76] present such a truncation for the BH equation, and show that it retains as first integrals approximations of (4.A.6) for $p = 1, 2, 3$. We recall their discretization here and note that it readily extends to the KdV equation.

Let $\mathcal{P}_N$ denote the standard $N$-mode Fourier projection operator, i.e.

$$f_N := \mathcal{P}_N f(x) = \sum_{|n|\leq N} \hat{f}_n e^{inx},$$

where

$$\hat{f}_n = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-inx}\,\mathrm{d}x$$

is the $n^{th}$ Fourier coefficient of the function $f(x)$. Since $f(x)$ is real we have

$$\hat{f}_{-n} = \hat{f}_n^*.$$

It can be directly verified that $\mathcal{P}_N$ is symmetric with respect to the $L^2$ inner product $(\cdot,\cdot)$ and commutes with the derivative operator $\frac{\partial}{\partial x}$. Consequently the composite operator $\frac{\partial}{\partial x}\mathcal{P}_N$ is skew-symmetric with respect to $(\cdot,\cdot)$ and a truncated Poisson bracket (4.A.4) may be defined by

$$\{\mathcal{F}_N, \mathcal{G}_N\} := -\int_0^{2\pi} \frac{\delta \mathcal{F}_N}{\delta u_N}\frac{\partial}{\partial x}\mathcal{P}_N\frac{\delta \mathcal{G}_N}{\delta u_N}\,\mathrm{d}x. \tag{4.A.8}$$

The Hamiltonian restricted to the truncated function $u_N$ is given by

$$H = \int_0^{2\pi}\left(\frac{1}{6}u_N^3 - \frac{\mu}{2}\left(\frac{\partial u_N}{\partial x}\right)^2\right)\mathrm{d}x.$$

Therefore the finite truncation follows from (4.A.2):

$$\frac{\partial u_N}{\partial t} = -\frac{\partial}{\partial x}\mathcal{P}_N\frac{\delta H}{\delta u_N},$$

where

$$\frac{\delta H}{\delta u_N} = \frac{1}{2}u_N^2 + \mu\frac{\partial^2 u_N}{\partial x^2}.$$

That is,

$$\frac{\partial u_N}{\partial t} + \frac{1}{2}\frac{\partial \mathcal{P}_N(u_N^2)}{\partial x} + \mu\frac{\partial^3 u_N}{\partial x^3} = 0. \tag{4.A.9}$$

In terms of Fourier coefficients this can be written

$$\frac{\mathrm{d}\hat{u}_n}{\mathrm{d}t} = -\frac{in}{2}\left(\sum_{|n-m|\le N} \hat{u}_{n-m}\hat{u}_m\right) + in^3\mu\hat{u}_n = -\frac{in}{2\pi}\frac{\partial H}{\partial \hat{u}_n^*}, \quad |n| \le N, \qquad (4.A.10)$$

and the Hamiltonian is

$$H = \frac{\pi}{3}\sum_{\substack{\ell+m+n=0\\ |\ell|,|m|,|n|\le N}} \hat{u}_\ell\hat{u}_m\hat{u}_n - \mu\pi\sum_{|\ell|\le N} \ell^2\hat{u}_\ell\hat{u}_\ell^*.$$

The Poisson bracket (4.A.8) possesses a Casimir invariant

$$M = \int_0^{2\pi} u_N\,\mathrm{d}x = \hat{u}_0,$$

the total momentum, which without loss of generality we assume to be zero.

Additionally the quadratic invariant

$$E = \frac{1}{2}\int_0^{2\pi} u_N^2\,\mathrm{d}x$$

is conserved since

$$\begin{aligned}\{E,H\} &= -\frac{1}{2}\int_0^{2\pi} u_N\frac{\partial \mathcal{P}_N(u_N^2)}{\partial x}\,\mathrm{d}x - \mu\int_0^{2\pi} u_N\frac{\partial^3 u_N}{\partial x^3}\,\mathrm{d}x\\ &= \frac{1}{2}\int_0^{2\pi} u_N^2\frac{\partial u_N}{\partial x}\,\mathrm{d}x + \mu\int_0^{2\pi}\frac{\partial u_N}{\partial x}\frac{\partial^2 u_N}{\partial x^2}\,\mathrm{d}x\\ &= \frac{1}{6}\int_0^{2\pi}\frac{\partial u_N^3}{\partial x}\,\mathrm{d}x + \frac{\mu}{2}\int_0^{2\pi}\frac{\partial}{\partial x}\left(\frac{\partial u_N}{\partial x}\right)^2\mathrm{d}x = 0,\end{aligned}$$

due to symmetry of $\mathcal{P}_N$ and its commutativity with $\frac{\partial}{\partial x}$. In terms of Fourier coefficients,

$$E = 2\pi\sum_{|n|\le N}\frac{1}{2}\hat{u}_n\hat{u}_n^* = \pi\hat{u}_0^2 + 2\pi\sum_{n=1}^{N}|\hat{u}_n|^2 = 2\pi\sum_{n=1}^{N}|\hat{u}_n|^2.$$

To solve (4.A.9) numerically, we evaluate the nonlinear terms in real space using a standard pseudospectral approach (see, e.g. [110]). Due to cubic terms in the Hamiltonian and the thermostat equation, anti-aliasing requires applying the FFTs on a grid of dimension $4(N+1)$, where $N$ is the number of Fourier modes retained in the truncation. All computations are done for fixed value of $N = 15$.

# Bibliography


[1] Abramov, R., Kovačič, G., and Majda, A.J. Hamiltonian structure and statistically relevant conserved quantities for the truncated Burgers-Hopf equation. *Communications on Pure and Applied Mathematics 56*, 1 (2003), 1–46.

[2] Abramov, R., and Majda, A.J. Statistically relevant conserved quantities for truncated quasi-geostrophic flow. *Proceedings of the National Academy of Science 100*, 7 (2003), 3841–3846.

[3] Arnold, V.I. *Mathematical Methods of Classical Mechanics*, second ed. Springer-Verlag, 1989.

[4] Bouchet, F., and Venaille, A. Statistical mechanics of geophysical flows. Emphasis on analytically solvable problems and geophysical applications. In *Peyresq lecture notes on nonlinear physics, volume 3* (in press), World Scientific.

[5] Bühler, O. *A Brief Introduction to Classical, Statistical, and Quantum Mechanics*, vol. 13 of *Courant Lecture Notes*. AMS Bookstore, Rhode Island, 2006.

[6] Bühler, O., and Holmes-Cerfon, M. Decay of an internal tide due to random topography in the ocean. *Journal of Fluid Mechanics 678* (2011), 271–293.

[7] Bussi, G., Donadio, D., and Parrinello, M. Canonical sampling through velocity rescaling. *Journal of Chemical Physics 126* (2007), 014101.

[8] Bussi, G., and Parrinello, M. Stochastic thermostats: comparison of local and global schemes. *Computer Physics Communications 179* (2008), 26–29.

[9] Caflisch, R.E. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica 7* (1998), 1–49.

[10] Carnevale, G.F., and Frederiksen, J.S. Nonlinear stability and statistical mechanics of flow over topography. *Journal of Fluid Mechanics 175* (1987), 157–181.

[11] Cassandro, M., Ciccotti, G., Rosato, V., and Ryckaert, J.P. Statistical mechanics of rigid systems: an atomic description. Unpublished note.

[12] Ciccotti, G., Kapral, R., and Vanden-Eijnden, E. Blue Moon sampling, vectorial reaction coordinates, and unbiased constrained dynamics. *Chemical Physics and Physical Chemistry 6*, 9 (2005), 1809–1814.

[13] Ciccotti, G., and Ryckaert, J.P. Molecular dynamics simulation of rigid molecules. *Computer Physics Reports 4*, 6 (1986), 346–392.

[14] da Silva, J.C.B., Magalhães, J., Gerkema, T., and Maas, L.R.M. Internal solitary waves in the Red Sea: an unfolding mystery. *Oceanography 25*, 2 (2012), 96–107.

[15] Darve, E. Méthodes multipôles rapides: résolution des équations de maxwell par formulations intégrales. Ph.D. Thesis.

[16] DelSole, T. A fundamental limitation of Markov models. *Journal of the Atmospheric Sciences 57* (2000), 2158–2168.

[17] den Otter, W.K., and Briels, W.J. The reactive flux method applied to complex isomerization reactions: Using the unstable normal mode as a reaction coordinate. *Journal of Chemical Physics 106*, 13 (1997), 5494.

[18] Dintrans, B., Rieutord, M., and Valdettaro, L. Gravito-inertial waves in a rotating stratified spherical shell. *Journal of Fluid Mechanics 398* (1999), 271–297.

[19] Drijfhout, S., and Maas, L.R.M. Impact of channel geometry and rotation on the trapping of internal tides. *Journal of Physical Oceanography 37* (2007), 2740–2763.

[20] Dubinkina, S., and Frank, J. Statistical mechanics of Arakawa's discretizations. *Journal of Computational Physics 227*, 2 (2007), 1286–1305.

[21] Dubinkina, S., and Frank, J. Statistical relevance of vorticity conservation in the Hamiltonian particle-mesh method. *Journal of Computational Physics 229*, 7 (2010), 2634–2648.

[22] Dubinkina, S., Frank, J., and Leimkuhler, B. Simplified modelling of a thermal bath, with application to a fluid vortex system. *SIAM Multiscale Modeling and Simulation 8* (2010), 1882–1901.

[23] Durran, D.R. *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*, vol. 32 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1999.

[24] E, W., and Mattingly, J.C. Ergodicity for the Navier-Stokes equation with degenerate random forcing: Finite-dimensional approximation. *Communications in Pure and Applied Mathematics 54* (2001), 1386–1402.

[25] E, W., Ren, W., and Vanden-Eijnden, E. String method for the study of rare events. *Physical Review B 66* (2002), 052301.

[26] Echeverri, P., Yokossi, T., Balmforth, N.J., and Peacock, T. Tidally generated internal-wave attractors between double ridges. *Journal of Fluid Mechanics 669* (2011), 354–374.

[27] Ellis, R.S., Haven, K., and Turkington, B. Nonequivalent statistical equilibrium ensembles and refined stability theorems for most probable flows. *Nonlinearity 15*, 2 (2002), 239–255.

[28] Fixman, M. Classical statistical mechanics of constraints: A theorem and application to polymers. *The Proceedings of the National Academy of Sciences USA 71*, 8 (1974), 3050–3053.

[29] Fixman, M. Simulation of polymer dynamics. II. Relaxation rates and dynamic viscosity. *Journal of Chemical Physics 69*, 4 (1978), 1527–1538.

[30] Frank, J., and Gottwald, G.A. The Langevin limit of the Nosé-Hoover-Langevin thermostat. *Journal of Statistical Physics 143*, 4 (2011), 715–724.

[31] Fricker, P., and Nepf, H. Bathymetry, stratification, and internal seiche structure. *Journal of Geophysical Research 105* (2000), 14,237–14,251.

[32] Gerkema, T., and van Haren, H. Absence of internal tidal beams due to non-uniform stratification. *Journal of Sea Research* (2012). DOI: 10.1016/j.seares.2012.03.008.

[33] Go, N., and Scheraga, H.A. Analysis of the contribution of internal vibrations to the statistical weights of equilibrium conformations of macromolecules. *Journal of Chemical Physics 51*, 11 (1969), 4751–4767.

[34] Go, N., and Scheraga, H.A. On the use of classical statistical mechanics in the treatment of polymer chain conformation. *Macromolecules 9*, 4 (1976), 535–542.

[35] Golub, G.H., and van Loan, C.F. *Matrix Computations*, second ed. The Johns Hopkins University Press, Baltimore and London, 1993.

[36] Grisouard, N., Staquet, C., and Pairaud, I. Numerical simulation of a two-dimensional internal wave attractor. *Journal of Fluid Mechanics 614* (2008), 1–14.

[37] Hairer, E., Lubich, C., and Wanner, G. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer, Berlin, 2006.

[38] Hairer, M. Convergence of Markov processes. Lecture notes, University of Warwick, 2010.

[39] Hairer, M., and Mattingly, J.C. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI* (2011), vol. 63 of *Progress in Probability*, pp. 109–117.

[40] Hartmann, C. Constraints in Molecular Simulation 2010, Zaragosa, slides `http://neptuno.unizar.es/events/constraints2010/files/Carsten_Hartmann.pdf`.

[41] Hazewinkel, J., Grisouard, N.G., and Dalziel, S.B. Comparison of laboratory and numerically observed scalar fields of an internal wave attractor. *European Journal of Mechanics- B/Fluids 30*, 1 (2011), 51–56.

[42] Hazewinkel, J., Tsimitri, C., Maas, L.R.M., and Dalziel, S.B. Observations on the robustness of internal wave attractors to perturbations. *Physics of Fluids 22* (2010), 107102.

[43] Hazewinkel, J., van Breevoort, P., Dalziel, S.B., and Maas, L.R.M. Observations on the wavenumber spectrum and evolution of an internal wave attractor. *Journal of Fluid Mechanics 598* (2008), 373–382.

[44] Higham, D.J. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review 43*, 3 (2001), 525–546.

[45] Holm, D.D., Marsden, J.E., and Ratiu, T.S. The Euler-Poincaré equations in geophysical fluid dynamics. In *Large-Scale Atmosphere-Ocean Dynamics II*, I. Roulstone and J. Norbury, Eds. Cambridge University Press, 2002, pp. 251–300.

[46] Hoover, W.G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A 31* (1985), 1695–1697.

[47] Iserles, A. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.

[48] Izaguirre, J.A., Catarello, D.P., Wozniak, J.M., and Skeel, R.D. Langevin stabilization of molecular dynamics. *Journal of Chemical Physics 114*, 4 (2001), 2090–2098.

[49] John, F. The Dirichlet problem for a hyperbolic equation. *American Journal of Mathematics 63* (1941), 141–154.

[50] Jones, A., and Leimkuhler, B. Adaptive stochastic methods for sampling driven molecular systems. *Journal of Chemical Physics 135*, 8 (2011), 084125.

[51] Khinchin, A.I. *Mathematical Foundations of Statistical Mechanics*. Dover, 1960.

[52] Kliemann, W. Recurrence and invariant measures for degenerate diffusions. *Annals of Probability 15*, 2 (1987), 690–707.

[53] Kloeden, P.E., and Platen, E. *Numerical Solution of Stochastic Differential Equations*, vol. 23 of *Applications of Mathematics*. Springer-Verlag, Berlin, 1992.

[54] Kopecz, S. Fractal internal wave patterns in a tilted square. Unpublished report, Kassel University, `http://www.nioz.nl/kopecz`, 2006.

[55] Kraichnan, R.H. Statistical dynamics of two-dimensional flow. *Journal of Fluid Mechanics 67* (1975), 155–175.

[56] Lam, F.-P.A., and Maas, L.R.M. Internal wave focusing revisited; a reanalysis and new theoretical links. *Fluid Dynamics Research 40*, 2 (2008), 95–122.

[57] LaZerte, B.D. The dominating higher order vertical modes of the internal seiche in a small lake. *Limnology and Oceanography 25*, S (1980), 846–854.

[58] Leimkuhler, B. Generalized Bulgac-Kusnezov methods for sampling of the Gibbs-Boltzmann measure. *Physical Review E 81* (2010), 026703.

[59] Leimkuhler, B. *Molecular Dynamics: An Introduction*. Springer, in press.

[60] Leimkuhler, B., Noorizadeh, E., and Penrose, O. Comparing the efficiencies of stochastic isothermal molecular dynamics methods. *Journal of Statistical Physics 143*, 5 (2011), 921–942.

[61] Leimkuhler, B., Noorizadeh, E., and Theil, F. A gentle stochastic thermostat for molecular dynamics. *Journal of Statistical Physics 135*, 2 (2009), 261–277.

[62] Leimkuhler, B., and Reich, S. Symplectic integration of constrained Hamiltonian systems. *Mathematics of Computation 63* (1994), 598–605.

[63] Leimkuhler, B., and Reich, S. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2004.

[64] Leith, C. Climate response and fluctuation dissipation. *Journal of the Atmospheric Sciences 32* (1975), 2022–2026.

[65] Lelièvre, T., Rousset, M., and Stoltz, G. Langevin dynamics with constraints and computation of free energy differences. *Mathematics of Computation* (2011). Accepted for publication.

[66] LeVeque, R.J. *Numerical Methods for Conservation Laws*. Birkhauser-Verlag, Basel, 1990.

[67] LeVeque, R.J. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. SIAM, 2007.

[68] Lighthill, J. Internal waves and related initial-value problems. *Dynamics of Atmospheres and Oceans 23* (1996), 3–17.

[69] Lynden-Bell, D. Statistical mechanics of violent relaxation in stellar systems. *Monthly Notices of the Royal Astronomical Society 136* (1967), 101–121.

[70] Maas, L.R.M. Wave attractors: linear yet nonlinear. *International Journal of Bifurcation and Chaos 15*, 9 (2005), 2757–2782.

[71] Maas, L.R.M. Exact analytic self-similar solution of a wave attractor field. *Physica D: Nonlinear Phenomena 238*, 5 (2009), 502–505.

[72] Maas, L.R.M., Benielli, D., Sommeria, J., and Lam, F.-P.A. Observation of an internal wave attractor in a confined, stably stratified fluid. *Nature 388* (Aug. 1997), 557–561.

[73] Maas, L.R.M., and Lam, F.-P.A. Geometric focusing of internal waves. *Journal of Fluid Mechanics 300* (1995), 1–41.

[74] Majda, A.J., Abramov, R., and Grote, M. *Information Theory and Stochastics for Multiscale Nonlinear Systems*, vol. 25 of *CRM Monograph Series*. American Mathematical Society, 2005.

[75] Majda, A.J., Gershgorin, B., and Yuan, Y. Low-frequency climate response and fluctuation-dissipation theorems: theory and practice. *Journal of the Atmospheric Sciences 67*, 4 (2009), 1186–1201.

[76] Majda, A.J., and Timofeyev, I. Remarkable statistical behavior for truncated Burgers-Hopf dynamics. *Proceedings of the National Academy of Sciences 97*, 23 (2000), 12413–12417 (electronic).

[77] Majda, A.J., and Timofeyev, I. Statistical mechanics for truncations of the Burgers-Hopf equation: a model for intrinsic stochastic behavior with scaling. *Milan Journal of Mathematics 70* (2002), 39–96.

[78] Majda, A.J., and Wang, X. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge University Press, Cambridge, 2006.

[79] Mattingly, J.C., Stuart, A.M., and Higham, D.J. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications 101*, 2 (2002), 185–232.

[80] McEwan, A.D., and Robinson, R.M. Parametric instability of internal gravity waves. *Journal of Fluid Mechanics 67*, 04 (1975), 667–687.

[81] McLachlan, R. Symplectic integration of Hamiltonian wave equations. *Numerische Mathematik 66* (1994), 465–492.

[82] Melchionna, S. Design of quasi-symplectic propagators for Langevin dynamics. *Journal of Chemical Physics 127* (2007), 044108.

[83] Miller, J. Statistical mechanics of Euler equations in two dimensions. *Physical Review Letters 65*, 17 (1991), 2137–2140.

[84] Miller, J., Weichman, P.B., and Cross, M.C. Statistical mechanics, Euler's equation, and Jupiter's Red Spot. *Physical Review A 45*, 4 (1992), 2328–2359.

[85] Minary, P., Martyna, G.J., and Tuckerman, M.E. Long time molecular dynamics for enhanced conformational sampling in biomolecular systems. *Physical Review Letters 93* (2004), 150201.

[86] Miura, R.M., Gardner, C.S., and Kruskal, M.D. Korteweg-de Vries Equation and Generalizations. II. Existence of Conservation Laws and Constants of Motion. *Journal of Mathematical Physics 9* (Aug. 1968), 1204–1209.

[87] Morrison, P.J. Hamiltonian description of the ideal fluid. *Reviews of Modern Physics 70*, 2 (1998), 467–521.

[88] Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics 52* (1984), 255–268.

[89] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *Journal of Chemical Physics 81* (1984), 511.

[90] Ogilvie, G.I. Wave attractors and the asymptotic dissipation rate of tidal disturbances. *Journal of Fluid Mechanics 543* (2005), 19–44.

[91] Olver, P.J. *Applications of Lie Groups to Differential Equations*, second ed., vol. 107 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1993.

[92] Onsager, L. Statistical hydrodynamics. *Nuovo Cimento (9) 6*, Supplemento 2, Convegno Internazionale di Meccanica Statistica (1949), 279–287.

[93] Ortega, J.M. *Numerical Analysis: A Second Course*. Classics in Applied Mathematics. SIAM, Philadelphia, 1990.

[94] Pavliotis, G.A., and Stuart, A.M. *Multiscale Methods: Averaging and Homogenization*. Texts in Applied Mathematics. Springer, New York, 2008.

[95] Penrose, O. *Foundations of Statistical Mechanics: A Deductive Treatment*. Dover, 2005.

[96] Reich, S. Smoothed dynamics of highly oscillatory Hamiltonian systems. *Physica D 89* (1995), 28–42.

[97] Rey-Bellet, L. Ergodic properties of Markov processes. In *Open Quantum Systems II*, vol. 1881 of *Lecture Notes in Mathematics*. Springer Berlin / Heidelberg, 2006, pp. 1–39.

[98] Rieutord, M., Georgeot, B., and Valdettaro, L. Wave attractors in rotating fluids: a paradigm for ill-posed Cauchy problems. *Physical Review Letters 85* (2000), 4277–4280.

[99] Robert, R. A maximum-entropy principle for two-dimensional perfect fluid dynamics. *Journal of Statistical Physics 65*, 3-4 (1991), 531–553.

[100] Robert, R., and Sommeria, J. Statistical equilibrium states for two-dimensional flows. *Journal of Fluid Mechanics 229* (1991), 291–310.

[101] Rogers, L.C.G., and Williams, D. *Diffusions, Markov Processes, and Martingales, Vol.2: Itô Calculus*. John Wiley & Sons, 1987.

[102] Salmon, R. *Lectures on Geophysical Fluid Dynamics*. Oxford University Press, 1998.

[103] Salmon, R., Holloway, G., and Hendershott, M.C. The equilibrium statistical mechanics of simple quasi-geostrophic models. *Journal of Fluid Mechanics 75* (1976), 691–703.

[104] Samoletov, A.A., Dettmann, C.P., and Chaplain, M.A.J. Thermostats for "slow" configurational modes. *Journal of Statistical Physics 128*, 6 (2007), 1321–1336.

[105] Sanz-Serna, J.M., and Calvo, M.P. *Numerical Hamiltonian Problems*, vol. 7 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1994.

[106] Swart, A., Sleijpen, G.L.G., Maas, L.R.M., and Brandts, J. Numerical solution of the two-dimensional Poincaré equation. *Journal of Computational and Applied Mathematics 200*, 1 (2007), 317–341.

[107] Swaters, G.E. *Introduction to Hamiltonian Fluid Dynamics and Stability Theory*, vol. 102 of *Monographs and Surveys in Pure and Applied Mathematics*. Chapman & Hall/CRC, 2000.

[108] Tang, W., and Peacock, T. Lagrangian coherent structures and internal wave attractors. *Chaos 20* (2010), 017508.

[109] Tilgner, A. Driven inertial oscillations in spherical shells. *Physical Review E 59* (1999), 1789–1794.

[110] Trefethen, L.N. *Spectral Methods in Matlab*. SIAM, 2000.

[111] van Kampen, N.G. Statistical mechanics of trimers. *Journal of Applied Sciences Research 37* (1981), 67–75.

[112] van Kampen, N.G., and Lodder, J.J. Constraints. *American Journal of Physics 52*, 5 (1984), 419–424.

[113] Whitham, G.B. *Linear and Nonlinear Waves*, second ed. Wiley, New York, 1999.

[114] Zeitlin, V. Finite-mode analogues of 2D ideal hydrodynamics: Coadjoint orbits and local canonical structure. *Physica D 49*, 3 (1991), 353–362.

# Summary

The research in this thesis is devoted to geometric numerical integration and thermostat methods for Hamiltonian ODEs and PDEs with applications in molecular dynamics and geophysical fluid dynamics.

We develop geometric methods for Hamiltonian wave equations arising in fluid dynamics. The separation of physical phenomena, such as energy conservation, advection, viscous diffusion and forcing, allows us to study these effects in isolation and identify the components responsible for a given phenomenon. In Chapter 2 we apply this philosophy to the model equations of internal gravity waves in a stratified fluid. We solve two initial value problems with several disturbances of the initial stream function field: for its free evolution and for its evolution under parametric excitation. We do this by developing a structure-preserving numerical method for internal gravity waves in a 2D stratified fluid domain. We recall the linearized, inviscid Euler-Boussinesq model, identify its Hamiltonian structure, and derive a staggered finite difference scheme that preserves this structure. For the discretized model, the initial condition can be projected onto normal modes whose dynamics is described by independent harmonic oscillators. This fact is used to explain the persistence of various classes of wave attractors in a freely evolving (i.e. unforced) flow. Under parametric forcing, the discrete dynamics can likewise be decoupled into Mathieu equations. The most unstable resonant modes dominate the solution, forming wave attractors.

In many cases Hamiltonian structure must be violated to simulate a given phenomenon. An example is given by constant temperature molecular dynamics, where it is mean kinetic energy, rather than total energy, that is conserved. In Chapter 3 we emphasise that a broad array of canonical sampling methods is available for molecular simulation based on stochastic-dynamical perturbation of Hamiltonian (Newtonian) dynamics, including Langevin dynamics, Stochastic Velocity Rescaling, and methods that combine Nosé-Hoover dynamics with stochastic perturbation. For temperature control we discuss several stochastic-dynamical thermostats in the setting of simulating Hamiltonian systems with holonomic constraints in contact with a thermal energy reservoir. The approaches described are easily implemented and facilitate the recovery of correct canonical averages with minimal disturbance of the underlying dynamics. For the purpose of illustrating our results, we examine the numerical application of these methods to a simple atomic chain, where a Fixman term is required to correct the thermodynamic ensemble.

In Chapter 4 thermal bath coupling mechanisms as utilized in molecular dy-

namics are extended to partial differential equation models. Working from a semi-discrete (Fourier mode) formulation for the Burgers-Hopf or KdV equation, we introduce auxiliary variables and stochastic perturbations in order to drive the system to sample a target ensemble which may be a Gibbs state or, more generally, any smooth distribution defined on a constraint manifold. We examine the ergodicity of approaches based on coupling the heat bath to the high wave numbers, with the goal of controlling the ensemble through the fast modes. We also examine different thermostat methods in the extent to which dynamical properties are corrupted in order to accurately compute the average of a desired observable with respect to the invariant distribution. The principal observation of this chapter is that convergence to the invariant distribution can be achieved by thermostatting just the highest wave number, while the evolution of the slowest modes is little affected by such a thermostat.

# Samenvatting

Het onderzoek in dit proefschrift is gewijd aan geometrische numerieke integratie en thermostaatmethoden voor Hamiltoniaanse gewone en partiële differentiaalvergelijkingen met toepassingen in de moleculaire dynamica en de geofysische stromingsleer.

We ontwikkelen geometrische methoden voor Hamiltoniaanse golfvergelijkingen die opduiken in de stromingsleer. De scheiding van fysische verschijnselen, zoals energiebehoud, advectie, viskeuze diffusie en aandrijving, staat ons toe deze effecten afzonderlijk te bestuderen en de componenten te identificeren die verantwoordelijk zijn voor een bepaald verschijnsel. In hoofstuk 2 passen we deze filosofie toe op de modelvergelijkingen voor interne zwaartekrachtgolven in een gestratificeerde vloeistof. We lossen twee beginwaardeproblemen op met diverse verstoringen van het initiële stroomfunctieveld: een met vrije evolutie en een met parametrische aandrijving. We doen dit door een structuurbehoudende numerieke methode voor interne zwaartekrachtgolven in een 2D gestratificeerd vloeistofdomein te ontwikkelen. We nemen het gelineariseerde viscositeitsvrije Euler-Boussinesq model in herbeschouwing, identificeren zijn Hamiltoniaanse structuur en leiden een versprongen ('staggered') eindig differentieschema af dat deze structuur behoudt. In het gediscretiseerde model kan de beginconditie geprojecteerd worden op normale modi wier dynamica beschreven wordt door harmonische oscillatoren. Dit gebruiken we om de persistentie van verscheidene klassen van golf-aantrekkers in een vrijelijk ontwikkelende (d.w.z. niet aangedreven) stroming te verklaren. In het geval van parametrische aandrijving kan de discrete dynamica evenzo worden ontkoppeld in Mathieu vergelijkingen. De instabielste resonante modi domineren de oplossing en vormen golf-aantrekkers.

In veel gevallen moet de Hamiltoniaanse structuur worden opgeheven om een bepaald verschijnsel te simuleren. Bijvoorbeeld, in de constante temperatuur moleculaire dynamica is het de gemiddelde kinetische energie en niet de totale energie die behouden blijft. In hoofdstuk 3 benadrukken we dat er een breed scala aan canonieke steekproefmethoden beschikbaar is voor moleculaire simulatie die gebaseerd zijn op stochastisch-dynamische verstoring van Hamiltoniaanse (Newtoniaanse) dynamica, waaronder Langevin dynamica, de zogenaamde Stochastic Velocity Rescaling, en methoden die Nosé-Hoover dynamica combineren met stochastische verstoring. Wat betreft temperatuurcontrole bediscussiëren we diverse stochastisch-dynamische thermostaten in de context van het simuleren van Hamiltoniaanse systemen met holonomische beperkingen binnen een thermische-energie reservoir. De beschreven benaderingen zijn gemakkelijk te implementeren en vergemakkelijken het terugvin-

den van de correcte canonieke gemiddelden met een minimale verstoring van de onderliggende dynamica. Met het doel onze resultaten te illustreren, bestuderen we de numerieke toepassing van deze methoden op een simpele atoomketting, waarbij een Fixman term nodig is om het thermodynamisch ensemble te corrigeren.

In hoofdstuk 4 worden thermischbadkoppelmechanismen, zoals gebruikt in de moleculaire dynamica, uitgebreid voor de toepassing in partiële differentiaalvergelijkingen. Werkende vanuit een semi-discrete (Fourier modus) formulering van de Burgers-Hopf of Korteweg-de Vries vergelijking, introduceren we hulpvariabelen en stochastische perturbaties zodat het systeem bij het nemen van steekproeven een bepaalde doelverdeling toont, wat een Gibbs toestand kan zijn of algemener elke gladde verdeling die gedefinieerd is op een beperkte variëteit. We bestuderen de ergodiciteit van aanpakken die gebaseerd zijn op koppeling van het warmtebad aan de hoge golfgetallen, met het doel het ensemble te beheersen via de snelle modi. Ook bestuderen we in hoeverre diverse thermostaatmethoden dynamische eigenschappen corrumperen om het gemiddelde van een gekozen waarneming, met betrekking tot de invariante verdeling, accuraat te berekenen. De belangrijkste observatie van dit hoofdstuk is dat de convergentie naar de invariante verdeling bereikt kan worden met een thermostaat die alleen het hoogste golfgetal direct aandrijft, terwijl de evolutie van de trage modi weinig beïnvloed wordt.

# Acknowledgements

I would like to express my gratitude to all those who gave me the possibility and encouragement to complete this thesis. I am truly indebted and thankful to my supervisor Jason Frank and professor Jan Verwer (1946-2011) for the opportunity, support, help and new vision to numerical analysis. This thesis would not have been possible without successful collaborations with Leo Maas and Ben Leimkuhler, to whom I owe my deepest gratitude. To Kees Oosterlee I am earnestly thankful for the additional work options which kept my PhD studies alive.

I am obliged to family, friends, relatives and colleagues. It is a great pleasure to thank everyone: My parents Jānis and Iveta, my sister Guna and her family, Kārlis and Andris, my grandparents Rasma, Halina (1930-2012), Elmārs (1925-2005) and Jurijs (1914-1996), my godmother Vita and her sister Inese with their families, my godfather Normunds with his brother Guntis and their families, my Latvian friends Ieva, Liene, Anna, Ilze, Evita, Matīss and Kristaps, my good friends and colleagues, Aram Markosyan, Shashi Jain, Yunus Hassen, Georgiana Caltais, Jesse Dorrestijn, Keith Myerscough, Christoph Koehn, Halldora Thorsdottir, Maarten Dijkema, Lacramioara Astefanoaei, Bram van Es, Wagner Fortes, Svetlana Dubinkina, Wander Wadman, Marjon Ruijter, Peter van Heijster, Anika Rinke, Iraes Rabbers, Anna Mozartova, Valeriu Savcenco, Liesbeth Vanherpe, Behnaz Changizi with Rory, Ivan Zapreev, Martina Chirilus-Bruckner, Severine Urdy, Maksat Ashyraliyev, Willem Haverkort, Linda Plantagie, Alexandra Silva, Shavarsh Nurijanyan, Henk Roose, Minnie Middelberg, Johan Schlepers, Coby van Vonderen, Bikkie Aldeias, Dubravka Tepsic, Daan Crommelin, Jeroen Hazewinkel, Barry Koren, Willem Hundsdorfer, Maciej Dobrzynski, Antonios Zagaris, Jens Rademacher, Ben Sommeijer (1949-2009), Margreet Nool, Nada Mitrovic, Hans Hidskes, Martine Anholt Gunzeln, Susanne van Dam, Irma van Lunenburg, Marlin van der Heijden, Karin van Gemert, Miriam Gravemaker...