**Academic year**
2024 - 2025

# Numerical Splitting Schemes for ODEs with a View Towards PDEs

**Bram Lens**

**Master of Science in Fundamental Mathematics**

Faculty of Science

Promotor
**Dr. Federico Zadra**

University of Antwerp

# Acknowledgements

Completing this thesis would not have been possible without the input of the people around me. I would like to start by thanking my promotor, dr. Zadra, for his continued support and encouragement: especially when MATLAB and I weren't exactly on speaking terms. I also want to show my deepest gratitude to:

- My parents: thank you for giving me the opportunity to pursue my interests;
- My sister, Anke: thank you for reading through this mathematical thesis, despite having been traumatised by Statistics I, II, III, and IV;
- Naomi, Andreas, Tom: thank you for your suggestions to improve my writing;
- Thomas: thank you for sharing your vast LaTex knowledge.

Finally, I would like to thank all the amazing people I got to know at the University of Antwerp. The last 5 years were amazing, and they have flown by in your presence.

**Abstract**

In this thesis, we perform a literary survey on the subject of splitting methods for the numerical integration of ordinary differential equations (ODEs). We also mention how splitting integrators designed for ODEs can be used to solve linear partial differential equations (PDEs) of first order. Splitting schemes reduce a given differential equation to several solvable subproblems. First, we present geometrical preliminary results that are needed later on in the thesis. Afterwards, we discuss several of their features and we focus on the subclass of geometrical integrators: e.g. when applied to Hamiltonian systems, they preserve the underlying symplectic structure. We implement several splitting integrators in MATLAB and make a comparison to classical integrators. We then move on to PDEs of first order and discuss how and when they can be solved analytically by means of the method of characteristics. Finally, we propose a class of splitting integrators for these PDEs that make use of their characteristics. These integrators have not garnered much attention yet in the literature. We focus in particular on the Liouville PDE applied to Hamiltonian systems and we provide an implementation in MATLAB.

**Samenvatting**

In deze thesis voeren we een literatuuronderzoek uit inzake het onderwerp van splijtingsmethoden voor de numerieke integratie van differentiaalvergelijkingen. Deze methoden reduceren een gegeven gewone of partiële differentiaalvergelijking (GDV resp. PDV) tot enkele oplosbare deelproblemen. We voorzien de lezer eerst van de benodigde meetkundige achtergrondkennis en gaan dan verder naar een bespreking van deze numerieke schema's voor GDVen. We bespreken enkele van hun belangrijkste kwaliteiten en focussen op de deelklasse van meetkundige integratieschema's: bv. wanneer ze worden toegepast op Hamiltoniaanse GDVen, dan behouden ze de onderliggende symplectische structuur van het probleem. We implementeren meerdere splijtingsmethoden en maken de vergelijking met klassieke integratoren. In het laatste hoofdstuk bespreken we voor lineaire PDVen de methode van der karakteristieken, waarmee deze vergelijkingen analytisch opgelost kunnen worden. Ten slotte introduceren we een klasse van integratoren voor PDVen die gebruik maken van de karakteristieken van de vergelijking. Deze integratieschema's hebben tot op heden nog niet veel aandacht gekregen in de literatuur. We vestigen onze focus specifiek op de Liouville PDV toegepast op Hamiltoniaanse systemen en we voorzien een implementatie in MATLAB.

# Contents

# Introduction

Historically, the notion of *divide et impera* (divide and rule) has described a political strategy employed to gain and maintain power over people by sowing division among them. A similar concept exists in numerical analysis [3], although in this case one is not interested in ruling human beings, but rather in conquering differential equations. Given a certain problem, albeit an ODE or a PDE that is generally hard to solve, the main idea is to split or divide it into two or more subproblems that allow for an easier or more efficient solution. After solving the respective subproblems, one then combines the obtained results into a solution to the initial problem. If this procedure creates sensible results, we can metaphorically say that we have *divided and conquered* the problem.

The idea of splitting in mathematics is not unique to the numerical integration of differential equations. For example, clustering algorithms are studied in data analysis [6]. These are algorithms that divide a given data set into several groups of similar observations. Consequently, a greater understanding of the data set can be created through understanding its clusters. A second example is the divide-and-conquer programming paradigm [11] in computer science. An example of a related problem is the sorting of an array of numbers: by dividing the array into smaller arrays, you can decrease the complexity of the algorithm and reduce the needed amount of computational time. A third and final example takes place in a completely different setting. Namely, a standard result in linear algebra states that any square, real, and symmetric matrix $\boldsymbol{A}$ admits a spectral decomposition $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\intercal$. By understanding $\boldsymbol{A}$ as a linear transformation, this *splitting* allows us to form greater insight into several of $\boldsymbol{A}$'s geometrical qualities, such as when $\boldsymbol{A}$ acts on a vector $v$ purely by scaling.

In Chapter 1, we provide the background knowledge that is needed throughout this thesis. In particular, results from differential geometry and the uniqueness and existence theory for ordinary differential equations (ODEs) are presented.

Chapter 2 offers a literary review on splitting methods for ODEs. We start by giving several examples of such integrators and then discuss a strategy to obtain higher order schemes. These higher order integrators can be obtained by appropriately composing lower order schemes with certain step sizes. We then move on to discussing their linear stability, and afterwards we focus on their place in the class of geometrical integrators: splitting integrators show a particularly good conservation of dynamical systems' integrals of motion. For example, in autonomous Hamiltonian systems, they show a near conservation of the total energy. To close this chapter, we consider some numerical simulations that showcase the behaviour of splitting schemes in comparison to the classical forward Euler integrator.

In Chapter 3, we focus on partial differential equations (PDEs) and how splitting methods can be constructed for their numerical integration. We mention finite

difference methods and operator splitting integrators, and we discuss the method of characteristics which can be used to analytically solve (quasi) linear PDEs of first order. We use this method to discuss a type of integrator that has not garnered much attention in the literature. These integrators make use of a PDE's characteristic curves. We make a comparison to the finite difference solution of the PDEs that we considered. In particular, we will consider the Liouville equation from statistical mechanics.

# Chapter 1

# Prerequisites

In this chapter, we introduce background information and notations that will be used throughout this thesis.

## 1.1 Differential geometry

### 1.1.1 Symplectic geometry

Symplectic geometry is essentially one of the languages of mechanics: the underlying phase spaces of Hamiltonian systems are modeled by symplectic manifolds. We will first introduce the concept of symplectic vector spaces and use this to define symplectic manifolds. Finally, we will discuss some key notions pertaining Hamiltonian systems. We will closely follow the strategy proposed in [9].

#### 1.1.1.1 Symplectic vector spaces

Consider a real vector space $V$ such that $\dim_{\mathbb{R}}(V) = d$.

**Definition 1.1.1**
Consider a mapping $\omega : V \times V \to \mathbb{R}$.

1. $\omega$ is called **bilinear** if it is linear in both of its arguments. Moreover, $\omega$ is called **anti-symmetric** if for any $u, v \in V : \omega(u, v) = -\omega(v, u)$.
2. An anti-symmetric bilinear map $\omega$ is called a **linear symplectic form** or **non-degenerate map** if the linear map $\tilde{\omega} : V \to V^* : v \mapsto \omega(v, \cdot)$ is bijective.
3. A vector space $V$ endowed with a linear symplectic form $\omega$ is called a **symplectic vector space**, and is denoted as a pair $(V, \omega)$.

Since we are considering finite-dimensional vector spaces, the non-degeneracy of $\omega$ is equivalent to

$$\ker(\omega) := \{u \in V \mid \forall v \in U : \omega(u, v) = 0\} = \{0\}$$

which is in turn equivalent to the statement

$$\forall v \in V \setminus \{0\} : \exists\, u \in V : \omega(v, u) \neq 0.$$

We can use linear symplectic forms to find a basis for $V$ that allows us to find a matrix representation of $\omega$.

**Theorem 1.1.2 (Darboux: linear case)**
Let $\omega : V \times V \to \mathbb{R}$ be a bilinear anti-symmetric map. There exists a basis $\{\alpha_1, \ldots, \alpha_k, e_1, \ldots, e_d, f_1, \ldots, f_d\}$ of $V$ satisfying the relations

1. $\omega(e_i, e_j) = \omega(f_i, f_j) = 0$ for any $i, j \in \{1, \ldots, d\}$;
2. $\omega(e_j, f_j) = \delta_{ij}$ for any $i, j \in \{1, \ldots, d\}$;
3. $\omega(\alpha_i, v) = 0$ for any $i \in \{1, \ldots, k\}$ and $v \in V$;

and $\omega$ can be expressed as follows with respect to this basis:

$$M_\omega = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \mathbb{I}_d \\ 0 & -\mathbb{I}_d & 0 \end{pmatrix}. \tag{1.1}$$

*Proof.* See [9, p. 3, Theorem 1.1]. □

If we let $(V, \omega)$ be a symplectic vector space of dimension $d = 2n$, then the theorem guarantees the existence of a so-called symplectic basis $\{e_1, \ldots, e_n, f_1, \ldots, f_n\}$ with $k = 0$ such that the properties $\omega(e_i, e_j) = \omega(f_i, f_j) = 0$ and $\omega(e_i, f_j) = \delta_{ij}$ hold for any indices $i, j \in \{1, \ldots, n\}$. In matrix notation, the linear symplectic form $\omega$ acts on vectors $u, v \in V$ as follows:

$$\omega(u, v) = u^\intercal \begin{pmatrix} 0 & \mathbb{I}_n \\ -\mathbb{I}_n & 0 \end{pmatrix} v.$$

where $\omega$'s representing matrix, usually denoted $\boldsymbol{J}_n$, is called the symplectic matrix. Note that here, the zero column and row in (1.1) no longer appear. That is because the theorem holds for general bilinear and anti-symmetric maps. Since symplectic linear forms are additionally non-degenerate, their representing matrices have full rank, hence $k = 0$ and $M_\omega$ is a matrix of dimension $2d \times 2d$.

**Proposition 1.1.3**
Symplectic vector spaces are even-dimensional.

*Proof.* Let $(V, \omega)$ be an $m$-dimensional symplectic vector space with basis $\mathcal{S}$ and let $\boldsymbol{\Omega}$ denote the matrix representing $\omega$ with respect to $\mathcal{S}$. Since $\omega$ is anti-symmetric, $\boldsymbol{\Omega}$ is too, and it follows that $\det(\boldsymbol{\Omega}) = \det(\boldsymbol{\Omega}^\intercal) = \det(-\boldsymbol{\Omega}) = (-1)^m \det(\boldsymbol{\Omega})$. Suppose $m$ is odd, then $\det(\boldsymbol{\Omega}) = 0$, which cannot be the case by $\omega$'s non-degeneracy. Therefore, $m > 0$ is even. □

**Example**
The simplest example of a symplectic vector space is $(\mathbb{R}^{2n}, \omega_{\text{st}})$, which is nothing more than the Euclidean space $\mathbb{R}^{2n}$ endowed with a *standard* linear symplectic form $\omega_{\text{st}}$. Here, $\omega_{\text{st}}$ is defined such that the vectors $e_i$ and $f_i$, with 1 at position $i$ resp. $2n + i$, form a symplectic basis for $\mathbb{R}^{2n}$.

We now introduce a group of mappings between symplectic vector spaces that preserve the symplectic structure.

**Definition 1.1.4**
Let $(V, \omega)$ and $(\tilde{V}, \tilde{\omega})$ be symplectic vector spaces. We call an isomorphism $\varphi : V \to \tilde{V}$ such that $\varphi^*\tilde{\omega} = \omega$ a (linear) **symplectomorphism**. If such a symplectomorphism exists, $(V, \omega)$ and $(\tilde{V}, \tilde{\omega})$ are called **symplectomorphic** and we write $V \cong_s W$.

The pullback condition assures that $\tilde{\omega}$ can be turned into a symplectic form on $V$ through the symplectomorphism $\varphi$. In the linear case, this condition states

$$\varphi^*\tilde{\omega} = \omega \Leftrightarrow \forall v_1, v_2 \in V : \varphi^*\tilde{\omega}(v_1, v_2) = \tilde{\omega}(\varphi(v_1), \varphi(v_2)) = \omega(v_1, v_2).$$

Since the map $\varphi$ in this definition is in particular linear, it can be represented by a matrix $\boldsymbol{A}$. If we choose $\omega = \tilde{\omega} = \omega_{\text{st}}$ on $\mathbb{R}^{2n}$, it follows that

$$\varphi^*\tilde{\omega}(v_1, v_2) = \tilde{\omega}(\boldsymbol{A}v_1, \boldsymbol{A}v_2) = (\boldsymbol{A}v_1)^\intercal \boldsymbol{J}_n(\boldsymbol{A}v_2) = v_1^\intercal \boldsymbol{A}^\intercal \boldsymbol{J}_n \boldsymbol{A} v_2$$

and since $\omega(v_1, v_2) = v_1^\intercal \boldsymbol{J}_n v_2$, the equality $\boldsymbol{A}^\intercal \boldsymbol{J}_n \boldsymbol{A} = \boldsymbol{J}_n$ must hold. This provides an alternative, equivalent definition of linear symplectomorphisms.

#### 1.1.1.2 Symplectic manifolds

We will endow manifolds with an additional *symplectic* structure. To do so, we will use the linear notions discussed in the previous section.

**Definition 1.1.5**
Let $\omega$ be a 2-form on a manifold $M$, and let $\omega$ be a 2-form on $M$ such that for any $m \in M$ :

1. The map $\omega_m : \mathrm{T}_m M \times \mathrm{T}_m M \to \mathbb{R}$ is anti-symmetric and bilinear;
2. The mapping $m \mapsto \omega_m$ is smooth.

Such a 2-form $\omega$ is called **symplectic** if it is also closed (i.e. $d\omega = 0$), and if $\omega_m$ is a linear symplectic form for any $m \in M$. A smooth manifold $M$ along with such a symplectic form $\omega$ is called a **symplectic manifold**, denoted by the pair $(M, \omega)$.

In short, a 2-form $\omega$ on $M$ is symplectic if it is smooth, closed and non-degenerate. If we have an $2n$-dimensional symplectic manifold $(M, \omega)$, then $\omega$ can be used to construct the *Liouville volume form* $\frac{\omega^n}{n!}$ on $M$ which is non-zero as a result of $\omega$'s non-degeneracy. As a result, any symplectic manifold is orientable.

**Proposition 1.1.6**
Symplectic manifolds $(M^m, \omega)$ are even-dimensional.

*Proof.* This proof is essentially the same as the linear case. If we choose $x \in M$ and let $\Omega_x \in \mathbb{R}^{m \times m}$ be $\omega$'s representing matrix w.r.t. a chosen basis of $\mathrm{T}_x M$, it follows that $\det(\Omega_x) = (-1)^m \det(\Omega_x)$. This leads to a contradiction with $\omega$'s non-degeneracy if $m$ is odd, so $m$ has to be even. □

**Example**
The 2-sphere $\mathbb{S}^2$ is a symplectic manifold when endowed with the symplectic form

$\omega_m(u,v) = m \cdot (u \times v)$ for $m \in \mathbb{S}^2$ and $u, v \in T_m\mathcal{S}^2$.

#### 1.1.1.3 The canonical symplectic form

Suppose $Q$ is an $n$-dimensional manifold and consider its cotangent bundle

$$T^*Q := \{(q, \alpha_q) \mid q \in Q, \alpha_q \in T_q^*Q\}$$

where $\alpha_q : T_qQ \to \mathbb{R}$ is a linear map. We denote the canonical projection $T^*Q \to Q$ by $\pi$.

**Definition 1.1.7**
The **tautological 1-form** $\vartheta$ on $T^*Q$, also known as the **Liouville 1-form** on $T^*Q$, is defined by

$$\vartheta_{(q,\alpha_q)}(v) := \alpha_q(T_{(q,\alpha_q)}\pi(v))$$

for $(q, \alpha_q) \in T^*Q$ and $v \in T_{(q,\alpha_q)}T^*Q$.

Using this 1-form, we will prove that the cotangent bundle of any smooth $n$-dimensional manifold can be equipped with the *canonical* symplectic form $\omega = d\vartheta$.

**Lemma 1.1.8**
Let $(U, \varphi)$ be a chart of $Q$ with local coordinates $q := (q^1, \dots, q^n)$ such that $(T^*U, T^*\varphi)$ is a chart of the cotangent bundle, where

$$T^*\varphi : T^*U \to \varphi(U) \times \mathbb{R}^n : (q, \alpha_q = \sum_i p_i dq^i \mid_q) \mapsto (q^1, \dots, q^n, p_1, \dots, p_n).$$

Then $\vartheta$ and $\omega$ can locally be written as $\vartheta = \sum_i p_i dq^i$ and $\omega = \sum_i dp_i \wedge dq^i$.

*Proof.* Consider a chart $(T^*U, T^*\varphi)$ of $T^*Q$ with coordinates $(q^1, \dots, q^n, p_1, \dots, p_n)$. Take $(q, \alpha_q) \in T^*Q$ such that $\alpha_q = \sum_i p_i dq^i$ by definition and take $v \in T_{(q,\alpha_q)}T^*Q$. It follows that there exists some $\xi := (\xi_1, \dots, \xi_n), \eta := (\eta_1, \dots, \eta_n) \in \mathbb{R}^n$ such that

$$v = \sum_i \left( \xi_i \left.\frac{\partial}{\partial q^i}\right|_{(q,\alpha_q)} + \eta_i \left.\frac{\partial}{\partial p_i}\right|_{(q,\alpha_q)} \right)$$

and it also follows that there exists some smooth curve $\gamma : ]-\varepsilon, \varepsilon[ \to T^*Q$ defined by $t \mapsto (\gamma_1(t), \gamma_2(t))$ such that $\gamma(0) = (q, \alpha_q)$ and $\dot\gamma(0) = v = (\xi, \eta)$. Since the local representation of $\pi$ is given by $(q^1, \dots, q^n, p_1, \dots, p_n) \mapsto (q^1, \dots, q^n)$ , it follows that

$$T_{(q,\alpha_q)}\pi(v) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (\pi(\gamma(t)) = \dot\gamma_1(0) = \xi = \sum_j \xi_j \left.\frac{\partial}{\partial q^j}\right|_q$$

and the tautological 1-form can be rewritten as

$$\begin{aligned}
\vartheta_{(q,\alpha_q)}(v) &= \left( \sum_i p_i dq^i \mid_q \right) \left( \sum_j \xi^j \left.\frac{\partial}{\partial q^j}\right|_q \right) \\
&= \left( \sum_i p_i dq^i \mid_q \right) \left( \sum_j \left( \xi^j \left.\frac{\partial}{\partial q^j}\right|_q + \eta_j \left.\frac{\partial}{\partial p_j}\right|_{(q,\alpha_q)} \right) \right) \\
&= \left( \sum_i p_i dq^i \mid_q \right) (v)
\end{aligned}$$

from which it follows that $\vartheta = \sum_i p_i dq^i$ and

$$\begin{aligned}\omega &= d\vartheta \\ &= \sum_i (dp_i \wedge dq^i + p_i \wedge d^2 q^i) \\ &= \sum_i dp_i \wedge dq^i\end{aligned}$$

which ends the proof. □

**Lemma 1.1.9**
$\omega$ as defined in the previous lemma, is a symplectic form on $T^*M$.

*Proof.* Since $\omega$ is exact, it is also closed. It is clearly smooth and non-degenerate, since it can be rewritten as $\omega = \sum_i dp_i \wedge dq^i$: the component functions are all smooth, and the matrix representing $\omega$ is of the form

$$\begin{pmatrix} 0 & \mathbb{I}_n \\ -\mathbb{I}_n & 0 \end{pmatrix}$$

which proves the statement. □

It follows from thest lemmas that any smooth $n$-dimensional manifold's cotangent bundle is a symplectic manifold. In other words, the cotangent bundle's symplectic structure arises naturally. It also serves as a foundation to Hamiltonian mechanics, since the symplectic structure gives rise to Hamilton's equations.

#### 1.1.1.4 Symplectomorphisms and Darboux' theorem

Symplectomorphisms preserve the symplectic structure of symplectic manifolds.

**Definition 1.1.10**
Let $(M_1, \omega_1)$ and $(M_2, \omega_2)$ be symplectic manifolds of dimension $2n$ and let $\varphi$ be a diffeomorphism $M_1 \to M_2$. Then $\varphi$ is called a **symplectomorphism** if $\varphi^*\omega_2 = \omega_1$.

For $p \in M$, the pullback condition in this definition reads as

$$\forall u, v \in \mathrm{T}_m M_1 : (\varphi^*\omega_2)_m(u, v) = (\omega_2)_{\varphi(m)}(d\varphi_m(u), d\varphi_m(v)) = \omega_1(u, v).$$

Since symplectomorphisms preserve symplectic forms, they preserve in particular the volume form on $(M, \omega)$. Suppose $\varphi : (M, \omega) \to (M, \omega)$ is a symplectomorphism on the $2n$-dimensional symplectic manifold $(M, \omega)$. Then

$$\varphi^*\left(\frac{\omega^n}{n!}\right) = \frac{1}{n!}\varphi^*\omega^n = \frac{1}{n!}\varphi^*(\omega \wedge \ldots \wedge \omega) = \frac{1}{n!}\varphi^*(\omega) \wedge \ldots \wedge \varphi^*(\omega) = \omega^n$$

by the properties of pullbacks.

**Proposition 1.1.11**
The composition of two symplectomorphisms is a symplectomorphism.

*Proof.* Let $\varphi_1 : (M_1, \omega_1) \to (M_2, \omega_2)$ and $\varphi_2 : (M_2, \omega_2) \to (M_3, \omega_2)$ be symplectomorphisms. Then it follows that

$$(\varphi_2 \circ \varphi_1)^* \omega_3 = \varphi_1^*(\varphi_2^* \, \omega_3) = \varphi_1^* \omega_2 = \omega_1$$

since $(f \circ g)^* = g^* \circ f^*$ for smooth maps $f, g \in \mathcal{C}^\infty(M)$. This proves the claim. □

Symplectomorphisms allow us to form a classification of symplectic manifolds.

**Theorem 1.1.12 (Darboux)**
Let $(M, \omega)$ be a $2n$-dimensional symplectic manifold and let $m \in M$. Then there exists a so-called Darboux-chart $(\varphi, U)$ with smooth coordinates $x_1, y_1, \ldots, x_n, y_n$ such that on $U$, $\omega$ can locally be written as

$$\omega = \sum_{i=1}^{n} dx_i \wedge dy_i.$$

*Proof.* A proof can be found in [18], for example. □

### 1.1.2 Hamiltonian mechanics

One of the most important formulations of classical mechanics is given by Hamiltonian mechanics, which emerged in the 19th century as an equivalent reformulation of Lagrangian mechanics [8]. Although both describe a system by its kinetic and potential energy $T$ and $V$, they do so in different manners: in Lagrangian mechanics, one considers the *Lagrangian function* given by $L = T - V$ (kinetic energy minus potential energy), whereas in Hamiltonian mechanics one considers the *Hamiltonian function* $H = T + V$ which usually denotes the total energy of a system. $H$ can in turn be used to define a vector field whose integral curves correspond to the system's equation of motion, allowing for a mathematical description of the dynamics. Moreover, what sets Hamiltonian mechanics apart from Lagrangian mechanics is the prevalence of symplectic geometry in its formulation: Hamiltonian mechanics is set on symplectic manifolds.

**Definition 1.1.13**
Let $(M, \omega)$ be a symplectic manifold and let $H : M \to \mathbb{R}$ be a smooth function on $M$. The unique vector field $X^H$ satisfying the equation

$$\iota_{X^H} \omega = -dH \text{ i.e. } \omega(X^H, \cdot) = -dH(\cdot)$$

is called the **Hamiltonian vector field** induced by $H$. The function $H$ is called the **Hamiltonian function** of the dynamical system. The triple $(M, \omega, H)$ is called a **Hamiltonian system**.

The integral curves of the Hamiltonian vector field $X^H$ are the solutions to the system's equations of motions, which are given by $\dot{x} = X^H(x)$. They are called *Hamilton's equations*.

**Remark 1.1.14**
If we consider the basic setting $(\mathbb{R}^{2n}, \omega_{\text{st}} = -\sum_{i=1}^{n} dq_i \wedge dp_i)$, we retrieve the

familiar definition of Hamiltonian vector fields. For some smooth function $H : \mathbb{R}^{2n} \to \mathbb{R}$, its resulting Hamiltonian vector field is defined by

$$X^H(q,p) := \begin{pmatrix} \partial_p H \\ -\partial_q H \end{pmatrix} = \partial_p H \frac{\partial}{\partial q} - \partial_q H \frac{\partial}{\partial p}.$$

An important feature of Hamiltonian vector fields is that their flows are symplectic. This property will be of great importance in the study of splitting integrators.

**Proposition 1.1.15**
The flow $\phi_t$ of a Hamiltonian vector field $X^H$ on a symplectic manifold $(M, \omega)$ is symplectic, i.e. $\phi_t^* \omega = \omega$.

*Proof.* Recall that the *Lie derivative* of the symplectic form $\omega$ with respect to the vector field $X^H$ is defined as

$$\mathcal{L}_{X^H}\omega := \lim_{h \to 0} \frac{\phi_h^* \omega(m) - \omega(m)}{h}$$

and that the celebrated *Cartan formula* is the identity $\mathcal{L}_{X^H}\omega = i_{X^H} d\omega + d i_{X^H}\omega$. From these definitions, it follows that

$$\phi_t^* \mathcal{L}_{X^H}\omega(m) = \lim_{h \to 0} \frac{\phi_t^*(\phi_h^*\omega(m)) - \phi_t^*\omega(m)}{h} = \lim_{h \to 0} \frac{\phi_{t+h}^*\omega(m) - \varphi_t^*(m)}{h} = \frac{\mathrm{d}}{\mathrm{d}t}\phi_t^*\omega$$

by definition of the derivative. Since $\omega$ is closed and by definition of Hamiltonian vector fields, Cartan's formula can be rewritten as $\mathcal{L}_{X^H}\omega = 0 + d(-dH) = -d^2 H = 0$. This implies that $\phi_t^* \mathcal{L}_{X^H}\omega(m) = 0$, which in turn implies $\phi_t^*\omega = \phi_0^*\omega = Id^*\omega = \omega$. □

A vector field $X$ on a smooth symplectic manifold $(M, \omega)$ is called *symplectic* if $\mathcal{L}_X \omega = 0$, i.e. if its flow preserves the symplectic form $\omega$. By the proposition above, any Hamiltonian vector field is symplectic. However, the converse implication is not true.

**Example ([9], p. 106)**
Consider the 2-torus $(\mathbb{T}^2, d\vartheta_1 \wedge d\vartheta_2)$. Then the vector fields

$$X_1 = \frac{\partial}{\partial \vartheta_1} \text{ and } X_2 = \frac{\partial}{\partial \vartheta_2}$$

are symplectic, but not Hamiltonian.

#### 1.1.2.1 Poisson brackets and conserved quantities

The symplectic structure of the phase space $M$ of a Hamiltonian system $(M, \omega, H)$ naturally equips $M$ with a bracket structure, giving the space $\mathcal{C}^\infty(M)$ of all smooth functions $f : M \to \mathbb{R}$ the structure of a Poisson algebra under the operation of composition.

**Definition 1.1.16**
Let $(M, \omega)$ be a $2n$-dimensional symplectic manifold. Then the *Poisson bracket*

of smooth functions $f$ and $g$ on $M$ is defined as the map

$$\{\cdot,\cdot\} : \mathcal{C}^\infty(M) \times \mathcal{C}^\infty(M) \to \mathcal{C}^\infty(M) : (f,g) \mapsto \{f,g\} := -\omega(X^f, X^g).$$

In canonical coordinates, the Poisson bracket can be written as

$$\{f,g\} = \sum_{i=1}^{n} \left( \frac{\partial f}{\partial q_i}\frac{\partial g}{\partial p_i} - \frac{\partial f}{\partial p_i}\frac{\partial g}{\partial q_i} \right).$$

Moreover, the Poisson bracket is related to the Lie (commutator) bracket by the identity $[X^f, X^g] = X^{-\{f,g\}}$: i.e. the map $H \mapsto X - H$ is a Lie algebra antihomomorphism from the Lie algebra of smooth functions endowed with the Poisson bracket to the Lie algebra of vector fields with the Lie bracket.

**Definition 1.1.17**
Let $(M,\omega)$ be a symplectic manifold and consider $H, f \in \mathcal{C}^\infty(M)$ to be smooth functions. The function $f$ is called an **integral** of $H$ if it is Poisson commutative with $H$, i.e. if $\{f,H\} = 0$.

Note that by the definition of Poisson brackets and Hamiltonian vector fields, the condition $\{f,H\}$ is equivalent to

$$\{f,H\} = -\omega(X^f, X^H) = -X_H(f).$$

If $H$ is the Hamiltonian function determining a Hamiltonian system, then the condition $\{f,H\} = 0$ can be interpreted to mean that $f$ remains constant along the flow of $X^H$. Moreover, any solution to the equations of motion lives in one and the same level set of $H$. This implies that Hamiltonian solutions are energy conserving.

**Example**
Suppose the Hamiltonian $H = H(q,p)$ is autonomous, i.e. it does not explicitly depend on time $t$. Then the Hamiltonian itself is an integral, meaning that the total energy of the system is constant along each trajectory.

This can be validated by computing the total derivative of $H$

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\partial H}{\partial t} + \{H,H\} = 0 + 0 = 0$$

since $H$ is assumed to be independent of time ($\partial_t H = 0$) and $\{\cdot,\cdot\}$ is anti-symmetric.

#### 1.1.2.2 Liouville's theorem

Liouville's theorem states that Hamiltonian flows preserve volumes in the phase space if $H = H(q,p)$ is autonomous. To prove this, we will follow [1, section 16]. We start by proving a more general proposition from which the theorem immediately follows. To this aim, we consider a system of ODEs $\dot{x} = X(x)$ with $x = (x_1, \ldots, x_m)$ for some $m \in \mathbb{N}$ and $X$ a complete vector field. We denote by $\phi_t$ the flow of $X$, which can be expanded as

$$\phi_t(x) = x + tf(x) + \mathcal{O}(t^2) \text{ as } t \downarrow 0 \tag{1.2}$$

by Taylor's theorem. We also consider a region $U_0$ in the space where the variables $x$ live, with volume $V_0 = \mathrm{Vol}(U_0)$ and we set $V(t) = \phi_t(V_0)$. We want to prove that $V(t) = V_0$. We will first prove the following lemma.

**Lemma 1.1.18**
Let $A = (a_{ij})_{i,j=1}^m$ be a square matrix and let $X$ be the vector field defined earlier.

1. Let $\varepsilon > 0$ be a constant. Then $\det(\boldsymbol{I} + \varepsilon \boldsymbol{A}) = 1 + t \cdot \mathrm{Trace}(\boldsymbol{A}) + \mathcal{O}(t^2)$.
2. $\frac{\mathrm{d}V}{\mathrm{d}t}\big|_{t=0} = \int_{U_0} \mathrm{div}(f)dx$ where $dx = dx_1 \dots dx_m$ and $\mathrm{div}(X)$ denotes the divergence of $X$.

*Proof.* See [1, p. 69-70, Lemmas 1 and 2]. □

Using this lemma, we can now prove the following general result.

**Proposition 1.1.19**
If $\mathrm{div}(X) = 0$, then $\phi_t$ preserves volumes: $V(t) = V_0$.

*Proof.* The second item in Lemma 1.1.18 can be rewritten as

$$\frac{\mathrm{d}V}{\mathrm{d}t}\bigg|_{t=t_0} = \int\limits_{U_{t_0}} \mathrm{div}(X)dx$$

for arbitrary $t_0 \in R$. Since $X$ is assumed to be divergence free, the expression above is equal to 0. Since this holds for any value of $t_0$, the function $V$ must be constant, hence $V(t) = V_0$. □

Liouville's theorem now immediately follows.

**Theorem 1.1.20 (Liouville's theorem)**
Hamiltonian flows preserve volumes in phase space.

*Proof.* By Proposition 1.1.19, it only rests to show that Hamiltonian vector fields are divergence-free. By routine calculations, we find that

$$\mathrm{div}(X_H) = \frac{\partial}{\partial q}\frac{\partial H}{\partial p} + \frac{\partial}{\partial p}\left(-\frac{\partial H}{\partial q}\right) = \frac{\partial^2 H}{\partial q \partial p} - \frac{\partial^2 H}{\partial p \partial q} = 0$$

which is 0 as a result of Schwarz' theorem. □

Liouville's theorem states that, although a region $U_0$ may change shape under the action of a Hamiltonian flow, its volume remains the same. This result also has an important consequence in statistical mechanics [28, p. 27-29]: the Hamiltonian flow is of class $L^1$ with respect to the volume measure induced by the symplectic form $\omega$. If we consider how a cloud of initial points evolves throughout phase space under the action of the Hamiltonian flow $\phi_t$, then its density must remain constant. In other words, if $\rho(q, p, t)$ denotes a probability density function (expressing the probability of the system being found in an infinitesimal volume of the point $(q, p)$ at a time $t$) on the phase space, then it must remain constant along trajectories of the system. As described in [13], this is summarized in the *Liouville equation*

$$\frac{\partial \rho}{\partial t} = -\sum_{j=1}^{n}\left(\dot{q}_i \frac{\partial \rho}{\partial q_i} + \dot{p}_i \frac{\partial \rho}{\partial p_i}\right) = -\{H, \rho\} \tag{1.3}$$

where $q := (q_1, \dots, q_n)$ and $p := (p_1, \dots, p_n)$. This is a linear, first order PDE. We will study it further in a later chapter.

### 1.1.3 The exponential map

We will first recall some definitions from differential geometry.

**Definition 1.1.21**
Let $M$, $N$ be manifolds and let $f : M \to N$ be a smooth map.

1. Pick $m \in M$, $v_m \in T_mM$, and let $g$ be a smooth function defined in a neighbourhood of $f(m)$. Then $T_mf(v_m)(g) := v_m(g \circ f)$ defines a map $T_mM \to T_{f(m)}N$ called the **tangent map**.

2. Let $X$ be a vector field on $M$ and $I \subseteq \mathbb{R}$ an open interval containing 0. An **integral curve** of $X$ is a curve $\gamma : I \to M$ such that $\gamma(0) = m$ and $T_t\gamma\left(\frac{d}{dt}\big|_t\right) = X(\gamma(t))$ for any $t \in I$.

3. The **flow of a vector field** $X$ on $M$ is the map $\mathcal{D}_X \to M$ defined by $\phi_t(m) := \phi(t, m)$ where $\mathcal{D}_X := \{(t, m) \mid m \in \mathcal{D}_t\}$ and $\mathcal{D}_t$ is defined as the set $\{m \in M \mid t \in I_m\}$ where $I_m$ is an open interval whose bounds depend on $m$.

4. A vector field $X$ on $M$ is **complete** if $\mathcal{D}_t = M$ for any $t \in \mathbb{R}$.

5. A vector field $X$ on a Lie group $G$ is called **left (resp. right) invariant** if $L_g^*X = X$ (resp. $R_g^*X = X$), where $L_g$ and $R_g$ denote left (resp. right) multiplication with some element $g \in G$. The set of left invariant vector fields on $G$ is denoted $\mathcal{X}^L(G)$ (resp. $\mathcal{X}^R(G)$).

If a vector field $X$ on a manifold $M$ is complete (i.e. its flow is defined at all times $t \in \mathbb{R}$) then the flow of $X$ satisfies the identities $\varphi_0 = \mathrm{Id}_m$ and $\varphi_{s+t}(m) = (\varphi_s \circ \varphi_t)(m)$ for any $m \in M$ and $s, t \in \mathbb{R}$. Using these definitions, we define the exponential map as follows.

**Definition 1.1.22**
Let $G$ be a Lie group and let $\mathfrak{g} \cong T_eG$ be its Lie algebra. The **exponential map** is defined as $\exp : \mathfrak{g} \to G : X \mapsto \gamma_X(1)$ where $\gamma_X : \mathbb{R} \to G$ is the maximal integral curve of $X$ through $e \in G$.

The well-definedness of the exponential map can be shown through the completeness of any $X \in \mathcal{X}^L(G)$, see [18]. This reference shows that $\mathcal{X}^L(G) \cong \mathfrak{g}$ as vector spaces, so any $X \in \mathfrak{g}$ can be seen as a left invariant vector field on $G$ (theorem 8.37). Moreover, it is shown (theorem 20.5) that for any $t \in R$ and $X \in \mathcal{X}^L(G)$: $\exp(tX) = \gamma_X(t)$, i.e. the exponential map can be used to express the flow of $X$ at any time $t \in \mathbb{R}$, since $X$ is complete. In other words, another way to express $X$ is by solving a system of ODEs in each chart of the underlying manifold, which is the Lie group $G$ in this case.

**Proposition 1.1.23 (Properties of the exponential map)**
Consider a Lie group G and its Lie algebra $\mathfrak{g} \cong T_eG$. Let $X \in \mathfrak{g}$.

1. The exponential map as defined above is smooth.

2. Consider $s, t \in \mathbb{R}$. Then $\exp(sX)\exp(tX) = \exp((s + t)X)$.

3. For any $n \in \mathbb{Z}$, it holds that $(\exp(X))^{-n} = \exp(-nX)$. In particular $(\exp(X))^{-1} = \exp(-X)$.

4. Identifying $T_0\mathfrak{g}$ with $\mathfrak{g}$, the tangent map $T_0\exp$ at 0 is the identity on $\mathfrak{g}$.

5. The map exp is a diffeomorphism of a neighbourhood of $0 \in \mathfrak{g}$ onto a neighbourhood of $e \in G$.

6. The flow $\phi$ at a time $t \in \mathbb{R}$ of $X \in \mathcal{X}^L(G)$ is given by left multiplication with $\exp(tX)$, i.e. $\phi_t = L_{\exp(tX)}$.

*Proof.* A proof is given in [18, Proposition 20.8]. □

The exponential map can be used in the description of a vector field's flow: writing out the result in this proposition, we find that the flow of $X$ is given by

$$\phi_t(p) = \exp(tX)p$$

for $t \in \mathbb{R}$ and $p \in G$.
Recall that the classic exponential map $\exp : \mathbb{R} \to \mathbb{R}$ satisfies the well-known identity $\exp(x + y) = \exp(x)\exp(y)$ for any $x, y \in \mathbb{R}$. However, the exponential map as introduced in Definition 1.1.22 does not.

**Proposition 1.1.24 (Baker-Campbell-Hausdorff (BCH))**
Let $G$ be a Lie group and $\mathfrak{g} \cong T_eG$ its Lie algebra. Consider $X, Y \in \mathfrak{g}$ in a sufficiently small neighbourhood $U \subseteq \mathfrak{g}$ of 0. Then $\exp(X)\exp(Y) = \exp(Z)$, where $Z$ is given by the following sum of commutators

$$\begin{aligned} Z = X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}([X, [X, Y]] + [Y, [Y, X]]) \\ - \frac{1}{24}[Y, [X, [X, Y]]] + \dots . \end{aligned}$$

*Proof.* See [23]. □

In particular, if we consider $t \in \mathbb{R}$ and $tX, tY \in \mathcal{X}^L(G)$, then

$$Z = tX + tY + \frac{t^2}{2}[X, Y] + \mathcal{O}(t^3).$$

In the matrix setting, the exponential map coincides with the matrix exponential, defined as follows for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.

**Definition 1.1.25**
Consider a square matrix $\boldsymbol{A} \in \mathbb{K}^{n\times n}$. The **matrix exponential** of $\boldsymbol{A}$ is defined by the real or complex $n \times n$ matrix

$$e^{\boldsymbol{A}} = \sum_{j=0}^{\infty} \frac{\boldsymbol{A}^j}{j!}.$$

This instance of the exponential map is discussed in detail in [32].

## 1.2 Ordinary differential equations

In this section, we will consider ordinary differential equations, or ODEs in short. First, we will recall some results on the existence and uniqueness of solutions to initial value problems. Then, we will present some methods to solve these initial value problems: we will discuss the class of Runge-Kutta integrators and the matrix exponential function.

### 1.2.1 Existence and uniqueness

The existence and unicity of solutions to ordinary differential equations is the subject of the following important theorem.

**Theorem 1.2.1 (Unique solution)**
Consider a function $f : [\alpha, \beta] \times \mathbb{R}^n \to \mathbb{R}^n$ such that

*(i)* $f$ is continous;

*(ii)* $f$ satisfies the Lipschitz condition in its $x$-variable, i.e.

$$\exists \vartheta > 0 : \forall x_1, x_2 \in \mathbb{R}^n : \|f(t, x_1) - f(t, x_2)\| \leq \vartheta \|x_1 - x_2\| \text{ for } \alpha \leq t \leq \beta$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^n$ and $\vartheta$ is known as $f$'s Lipschitz constant with respect to its $x$-variable;

then the ODE $\dot{u}(t) = f(t, u(t))$ with initial value $u(\alpha) = u_0 \in \mathbb{R}^n$ has a unique solution on the time interval $t \in [\alpha, \beta]$.

To prove this theorem, we will use Banach's fixed point theorem. To this aim, a contraction is defined as a map $\psi : X \to X$ on a complete metric space $(X, \|\cdot\|)$ satisfying the condition

$$\exists\, 0 < L < 1 : \forall x_1, x_2 \in X : \|\varphi(x_1) - \varphi(x_2)\| \leq L \|x_1 - x_2\|$$

which is also known as a Lipschitz condition with Lipschitz constant $L$. In [33, p. 27], it is shown that a function $\varphi(t, x)$ is Lipschitz with respect to its $x$-variable if $f$ is continuously differentiable with respect to $x$, and if $\partial_x f$ is bounded.

**Theorem 1.2.2 (Banach's fixed point theorem)**
Consider a complete metric space $(X, \|\cdot\|)$ and a contraction $\varphi : X \to X$. Then $\varphi$ has a unique fixed point, i.e. $\exists!\ \chi \in X : \varphi(\chi) = \chi$.

*Proof.* See [16]. □

Banach's fixed point theorem allows us to prove the theorem.

*Proof of the unique existence theorem.* Consider the Banach space $(X, \|\cdot\|)$ of continuous functions $[\alpha, \beta] \to \mathbb{R}^n$ with pointwise addition and scalar multiplication. Let $u \in X$ and $t \in [\alpha, \beta]$. The result immediately follows from an application of

Banach's fixed point theorem to the functional

$$\varphi(u)(t) := u_0 + \int_\alpha^t f(s, u(s))ds.$$

□

**Remark 1.2.3 (Counterexamples to the theorem)**
What happens when the conditions of the unique existence theorem are not met?

1. Consider the initial value problem $\dot{x}(t) = \sqrt[3]{x}$ and $x(0) = 0$. An elementary computation shows that
$$x(t) = \pm\frac{1}{3}\sqrt{6t^3}$$
are two possible solutions to the initial value problem, violating the unicity result in Theorem 1.2.1. This is because $f(t, x) = \sqrt[3]{x}$ is not Lipschitz in $x$, since its derivative $\partial_x f = \frac{1}{3}x^{-\frac{2}{3}}$ is unbounded.

2. Consider the initial value problem $\dot{x}(t) = \frac{1}{t}$ and $x(0) = 0$. Clearly, the function $f(t, x) = \frac{1}{t}$ is discontinuous in $t = 0$. If we try to solve the problem, we find that
$$x(t) = \ln|t| + C \text{ for some } C \in \mathbb{R}$$
is a solution to the problem. However, since $\ln 0$ is undefined, this cannot be a solution to the problem. The problem has no solutions at all, violating the existential result from the theorem.

The theorem only speaks of ODEs of the form $\dot{u}(t) = f(t, u(t))$. Since any explicit $n$-th order ODE can be expressed in this form (i.e. a system of $n$ first order ODEs), the theorem's result is not too restrictive.

**Example**
We consider two $n$-th order ODEs and rewrite them as a system of $n$ first order ODEs. To do this, we will make use of the identifications $x_n := x^{(n)}$.

1. Consider $\ddot{x} = x$. Then this ODE is equivalent to
$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = x_1 \end{cases}$$
which is a system of $n = 2$ first order ODEs. If we define $\vec{x}(t) := (x_1(t), x_2(t))^{\mathrm{T}}$, then we can rewrite this system as
$$\dot{\vec{x}}(t) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \vec{x}(t)$$
such that $f(t, \vec{x}(t)) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \cdot \vec{x}(t)$.

2. Consider $\dddot{x} - \cos(\ddot{x})e^x - \dot{x} = e^{42t}$. Then this ODE is equivalent to

$$\begin{cases} \dot{x}_0 = x_1 \\ \dot{x}_1 = x_2 \\ \dot{x}_2 = \cos(x_2)e^{x_0} + x_1 + e^{42t} \end{cases}$$

which is a system of $n = 3$ first order ODEs. In a similar manner as before, the right-hand side of the equations in this system give us an expression for the function $f(t, x(t))$.

### 1.2.2 Solutions using exponentials

This section is based on [21] and [32]. The exponential map can be used to compute the solutions to both linear and non-linear ODEs.

#### 1.2.2.1 Linear systems

**Theorem 1.2.4**
Let $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and consider the homogeneous system of ODEs $\dot{x} = f(t, x(t)) = \boldsymbol{A}x$ with initial value condition $x(0) = x_0 \in \mathbb{K}^n$. Suppose $f$ is continuous. Then $x(t) = e^{t\boldsymbol{A}} \cdot x_0$ is a solution of the initial value problem.

*Proof.* This result is implied by the basic properties of the matrix exponential. □

**Theorem 1.2.5**
The matrix exponential can also be used to solve an inhomogeneous linear system of ODEs $\dot{x} = f(t, x) = \boldsymbol{A}x + g(t)$ with initial value condition $x(0) = x_0$ where $\boldsymbol{A} \in \mathbb{K}^{n\times n}$ for $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, $f$ is continuous and $g(t)$ is a continuous vector valued function, i.e. $g(t) \in \mathbb{K}^n$. The solution is given by

$$x(t) = x_h(t) + x_p(t)$$

where $x_h(t)$ denotes the homogeneous solution (see Theorem 1.2.4) and $x_p(t)$ denotes the particular solution, given by

$$x_p(t) = e^{t\boldsymbol{A}} \int_0^t e^{-s\boldsymbol{A}} g(s) ds.$$

*Proof.* To prove this, one would use the integrating factor $\mu(t) := e^{-t\boldsymbol{A}}$ to transform the system and find a solution describe its solution. A proof is given in [21]. □

By combining Theorem 1.2.4 and Theorem 1.2.5, it now becomes easy to find the solution to an inhomogeneous linear system of ODEs.

**Example**
Consider the inhomogeneous linear system $\dot{x}(t) = \boldsymbol{A}x(t) + g(t)$ where

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad g(t) = \begin{pmatrix} e^t \\ e^t \end{pmatrix}, \quad x(0) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

The matrix exponential $e^{t\boldsymbol{A}}$ is given by

$$e^{t\boldsymbol{A}} = \begin{pmatrix} \cosh(t) & \sinh(t) \\ \sinh(t) & \cosh(t) \end{pmatrix}$$

for $t \geq 0$. By Theorem 1.2.4, we find that

$$x_H(t) = e^{t\boldsymbol{A}} \cdot x(0) = \begin{pmatrix} e^{-t} \\ -e^{-t} \end{pmatrix}$$

is a homogeneous solution to our problem. We can now use Theorem 1.2.5 to find the particular solution

$$x_P(t) = \frac{1}{2} \begin{pmatrix} e^t + e^{-t} & e^t - e^{-t} \\ e^t - e^{-t} & e^t + e^{-t} \end{pmatrix} \int_0^t \begin{pmatrix} \cosh(s) & \sinh(s) \\ -\sinh(s) & \cosh(s) \end{pmatrix} \begin{pmatrix} e^s \\ e^s \end{pmatrix} ds = \begin{pmatrix} te^t \\ te^t \end{pmatrix}$$

which results in the total solution

$$x(t) = x_H(t) + x_P(t) = \begin{pmatrix} te^t + e^{-t} \\ te^t - e^{-t} \end{pmatrix}.$$

#### 1.2.2.2 Non-linear systems

We now consider non-linear ODEs of the form $\dot{x} = X(x)$ where $X : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous. Recall that the *Lie derivative* of a smooth function $h : \mathbb{R}^n \to \mathbb{R}$ at $p \in \mathbb{R}^n$ with respect to the vector field $X$ is given by

$$\mathcal{L}_X h(p) := \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} \left( \left( h \circ \Phi_t^X \right)(p) \right) = \lim_{t \to 0} \frac{\left( h \circ \Phi_t^X \right)(p) - h(p)}{t} = X(p) \cdot \nabla h(p).$$

The flow of $\dot{x} = X(x)$ then satisfies the identity

$$h\left( \Phi_t^X(p) \right) = \left( \sum_{j=0}^{+\infty} \frac{t^j}{j!} \mathcal{L}_X^j h \right)(p) = \left( e^{t\mathcal{L}_X} h \right)(p)$$

involving an operator exponential, as mentioned in [3, pg. 6]. If we consider $h$ to be the identity map, we recover the flow of $X$. Since the exponential, see Definition 1.1.22, maps a vector field onto its maximal integral curve through the identity, we can also write down the formal solution to the ODE $\dot{x} = X(x)$ with $x(0) = x_0$ as

$$x(t) = \exp(tX) x_0$$

instead of considering the vector field's Lie derivative $\mathcal{L}_X$. The operator $\exp(tX)$ then denotes the flow of the vector field $X$ at a time $t$.
As an example, we consider the ODE $\dot{x} = X(x) = x^2$ defined on $|t| < 1$ such that $x(0) = x_0$. By induction, we can easily prove that $\mathcal{L}_X^n h(x) = n! \cdot x^{n+1}$ for any $n \in \mathbb{N}$. It immediately follows that

$$\left( e^{t\mathcal{L}_X} h \right)(x_0) = \sum_{j=0}^{+\infty} \frac{t^j}{j!} j! x_0^{j+1} = x_0 \sum_{j=0}^{+\infty} (x_0 t)^j = \frac{x_0}{1 - x_0 t}$$

which is indeed the solution to the ODE.

# Chapter 2

# Splitting methods for ODEs

In this chapter, we review splitting integrators for ODEs. We present strategies for increasing their order of accuracy and we analyse their linear stability. We then perform a backward error analysis to determine how well these integrators preserve structural properties of ODEs, such as energy and other first integrals. Finally, we provide a numerical implementation of several splitting integrators and we compare their performance to the forward Euler method.

## 2.1 Splitting integrators

In this section, the problem we face is an autonomous ODE of the form

$$\dot{y} = \frac{\mathrm{d}y}{\mathrm{d}t} = f(y(t)) \text{ with } y(T) = y_0 \in \mathbb{R}^D \tag{2.1}$$

where $f : \mathbb{R}^D \to \mathbb{R}^D$ is a function, $D \in \mathbb{N}^+$, and $T \in I \subseteq \mathbb{R}$ for some open interval $I$ containing $T$. We will closely follow the exposition in [3]. We assume that $f$ can be *split* as $f = f_1 + \ldots + \ldots f_m$ for some integer $m > 1$, where the $f_i : \mathbb{R}^D \to \mathbb{R}^D$ are functions such that each subproblem

$$\dot{y} = f_i(y(t)) \text{ with } y(0) = y_0 \tag{2.2}$$

admits an easier or more efficient solution than (2.1). The combination of these subproblems' solutions gives rise to numerical schemes approximating the solution to (2.1). These integrators are called splitting integrators. In general, there are three steps [22, pg. 8] to numerically solve an ODE using a splitting integrator:

1. Choosing a splitting $f = f_1 + \ldots + f_m$;
2. Solving the subproblems, albeit exact or numerically;
3. Combining the subproblems' solutions into an approximation for the initial problem.

By construction, splitting methods form a natural choice of integrator for systems whose vector field $f$ consists of several different physical components [3]. For example, systems in Hamiltonian mechanics can often be split into a kinetic component $T(p)$ and a potential component $V(q)$, or systems expressing fluid dynamics can be split into an advection and a convection component. By building numerical schemes

on the level of the subproblems, these components' properties can be preserved better than is the case for classical integrators. Moreover, different splittings of $f$ give rise to different integrators, which can each have different qualitative properties. Let us now move on to two first examples of splitting integrators for (2.1).

**Example (Lie-Trotter, Strang splitting)**
Consider the ODE (2.1), suppose $f$ can be split as $f = f_1 + f_2$, and denote the flow of each subproblem (2.2) by $\varphi_h^i$. Then the **Lie-Trotter** approximations to the problem are given by

$$y_i = (\varphi_h^2 \circ \varphi_h^1)(y_{i-1}) \tag{2.3}$$

for $i \geq 1$ where $y_i = y_0 \in \mathbb{R}^D$ for $i = 1$ and $h > 0$ is the step size. Similarly, the Strang approximations to the problem are given by

$$y_i = \left(\varphi_{\frac{h}{2}}^1 \circ \varphi_h^2 \circ \varphi_{\frac{h}{2}}^1\right)(y_{i-1}) \tag{2.4}$$

The Lie-Trotter integrator [3, p. 3] is equivalent to first flowing along the solution of the first subproblem for a step of size $h$, and then flowing along the second subproblem's solution for a step of size $h$. The Strang integrator [3, p. 4] can be interpreted in a similar way, but with steps of size $\frac{h}{2}$ being used for the flow of the second subproblem. Moreover, each integrator can be seen as a map that sends the previous approximation to the next one.

**Definition 2.1.1**
If a numerical integrator computes its approximations as $Y_{n+1} = g(Y_n)$ for some function $g$, then the map $Y_n \mapsto Y_{n+1}$ is called the **numerical flow** if the integrator.

Splitting methods are not only useful in the case of splittings $f = f_1 + \ldots + f_m$ such that each subproblem has an exact solution. If a subproblem does not have an exact solution, say $f^j$, then we can include the numerical flow of some integrator applied to $\dot{y} = f_j(y(t))$ in the compositions (2.3) or (2.4) to define a new integrator. This introduces an approximation error on top of the error due to the splitting.
Splitting methods can be expressed in terms of compositions of exponential maps. The ODE (2.1) is nothing more than the system locally determining the integral curves of the vector field $f$, and we can consider the Lie derivative $\mathcal{L}_f$. Consequently, the exponential $\exp(h\mathcal{L}_f)$ can be used to express the exact solution to (2.1). If we apply this reasoning to the splitting methods discussed above, we can rewrite them as compositions

$$y_i = \exp(h\mathcal{L}_{f_1}) \exp(h\mathcal{L}_{f_2}) y_{i-1} \text{ resp. } y_i = \exp\left(\frac{h}{2}\mathcal{L}_{f_1}\right) \exp\left(h\mathcal{L}_{f_2}\right) \exp\left(\frac{h}{2}\mathcal{L}_{f_1}\right) y_{i-1}$$

and the same can be done for splitting methods in general.
Splitting methods have a multitude of desirable features that make them a decent choice of integrator in many settings. For example, they perform well when each subproblem is assumed to have a common property that is preserved by composition. Since splitting methods are constructed as compositions of the subproblems' flows, the integrator itself will also conserve these properties [12] by construction. One such property that we will study throughout this chapter is symplecticity in Hamiltonian

systems, see Proposition 1.1.11. Other examples of properties that are conserved by splitting methods are the symmetry and reversibility of systems, which will be discussed later. In summary, splitting methods are well-adapted to preserve several structural and geometrical features of dynamical systems, which is why they are also known as a type of *geometrical* integrator.
There are several advantages in regards to the implementation of splitting methods. They are usually explicit, typically do not have high storage requirements, and the implementation of higher order schemes is quite straightforwad [3, p. 23]. There are also some disadvantages to splitting schemes. For example, when constructing splitting schemes of order $> 3$, negative time step coefficients are necessarily involved, which makes for a bad integration scheme for e.g. reaction-diffusion equations and for increasing order, the stability of a splitting integrator usually decreases [3, p. 24]. In the literature, while the construction of higher order integrators is often discussed (see for example [34]), the stability of these integrators is frequently ignored.

### 2.1.1 Non-autonomous systems

So far, we have only discussed splitting methods for autonomous ODEs (2.1). In this section, we will discuss how splitting methods can be implemented for non-autonomous ODEs, that is, ODEs with an explicit time-dependence

$$\dot{y}(t) = f(t, y(t)) \text{ with } y(0) = y_0 \tag{2.5}$$

where $f : \mathbb{R}^D \to \mathbb{R}^D$ can be split as $f = f_1 + \dots f_m$. The discussion in this section is based on [3, section 3.5]. Splitting methods can be applied to these non-autonomous systems by converting them into autonomous systems. This can be done by considering $t$ to be an additional variable of the function $f$. In other words, if we let $y = (y_1, \dots, y_D)$ and set $y_{D+1} = t$, then we can equivalently rewrite (2.5) as follows

$$\dot{z} = \tilde{f}(z) \text{ with } z = (y_{D+1}, y_1, \dots, y_D) = (y_{D+1}, y) \tag{2.6}$$

where $\tilde{f} = (f_1, \dots, f_{m+1})$ is given by

$$\begin{pmatrix} \dot{y}_{D+1} \\ \dot{y} \end{pmatrix} = f(y_{D+1}, y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ f_1(y_{D+1}, y) \end{pmatrix} + \dots + \begin{pmatrix} 0 \\ f_m(y_{D+1}, y) \end{pmatrix}$$

with

$$\tilde{f}_i(z) = \begin{pmatrix} 0 \\ f_i(z) \end{pmatrix} \text{ for } i \in \{1, \dots, m\}$$

$$\tilde{f}_{m+1}(z) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

which means that $y_{D+1} = t$ is only moved by the flow of the subproblem $\tilde{f}_{m+1}$ and remains constant with respect to the other flows. As (2.5) has now been transformed into an autonomous sytem (2.6), the previously discussed splitting methods are applicable.

In the case $m = 2$, one can also consider $t$ as an additional variable twice by setting $y_{D+1} = t$ and $y_{D+2} = t$. This results in the autonomous system

$$\begin{pmatrix} \dot{y}_{D+1} \\ \dot{y}_{D+2} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ f_1(y_{D+1}, y) \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ f_2(y_{D+2}, y) \end{pmatrix}$$

where $\tilde{f}_1$ is the first term and $\tilde{f}_2$ is the second term on the right-hand side of the equation. Moreover, $y_{D+1}$ is only moved by the flow of $\tilde{f}_2$ and $y_{D+2}$ is only moved by the flow of $\tilde{f}_1$. We mention this, because further on in this chapter we will consider autonomous Hamiltonian systems induced by separable Hamiltonian functions $H(q, p) = T(p) + V(q)$ leading to a Hamiltonian vector field that can be split as $X_H = X_T + X_V$. The strategy proposed above offers a method to deal with non-autonomous Hamiltonian systems. However, these will not be discussed in the sequel, because autonomous Hamiltonian systems possess some properties that nicely demonstrate the nice features of splitting schemes.

## 2.2 Higher order integrators

There are several ways to increase the precision of splitting integrators. A first strategy consists of looking for conditions on the step sizes to obtain a desired order. This can be done by comparing Taylor expansions. After all, an integrator is classically said to be of some order $n$ if it coincides with the exact solution's Taylor expansion up to the term of order $h^n$ [35]. We will only demonstrate this method for orders $n \in \{1, 2\}$, since the same procedure can be carried out for higher order conditions. A second strategy consists of using the BCH-formula. We specifically focus on the recursive construction of Yoshida's even order integrators, which uses the BCH-formula. Since this second strategy starts from a second order integrator, we focus our study of the direct method on the cases $n \in \{1, 2\}$. Finally, we briefly discuss a third strategy to construct more efficient and *economical* integrators of orders $n = 6$ and $n = 8$. This section is based on Yoshida's paper [34].

### 2.2.1 Compositions with identical step sizes

We first consider ODEs of the form (2.1) whose corresponding vector fields can be split as $f = f_1 + \ldots + f_m$ and we consider integrators that move along each subproblem's flow with the same step size $h > 0$. In the case $m = 2$

$$\dot{y} = f(y) = f_1(y) + f_2(y)$$

we find that the resulting subproblems' flows can be expanded as

$$\varphi_h^{f_1}(y) = y + hf_1(y) + \frac{h^2}{2} f_1'(y) f_1(y) + \mathcal{O}(h^3) =: y^*$$
$$\varphi_h^{f_2}(y) = y + hf_2(y) + \frac{h^2}{2} f_2'(y) f_2(y) + \mathcal{O}(h^3)$$

where $h \downarrow 0$. Here, we use Landau's *big O* symbol to mean that the omitted terms in the expansions vanish at the same rate as $h^3$ when $h \downarrow 0$, see [14]. Moreover, we

introduce the notation $y^*$ for the first expansion to improve legibility in the following equations. From the equations above, it follows that

$$\left(\varphi_h^{f_2} \circ \varphi_h^{f_1}\right)(y) = \varphi_h^{f_2}(y^*) = y^* h f_2(y^*) + \frac{h^2}{2} f_2'(y^*) f_2(y^*) + \mathcal{O}(h^3).$$

If we now compute the factors $f_2(y^*)$ and $f_2'(y^*)$ as follows

$$\begin{aligned} f_2(y^*) &= f_2(y) + h f_2'(y) f_1(y) + \mathcal{O}(h^2) \\ f_2'(y^*) &= f_2'(y) + \mathcal{O}(h) \end{aligned}$$

then we can rewrite the composition $\varphi_h^{f_2} \circ \varphi_h^{f_1}$ as

$$\left(\varphi_h^{f_2} \circ \varphi_h^{f_1}\right)(y) = y + h f_1(y) + h f_2(y) + \frac{h^2}{2} f_1'(y) f_1(y) + h^2 f_2'(y) f_1(y) + \frac{h^2}{2} f_2'(y) f_2(y) + \mathcal{O}(h^3).$$

Since the problem's exact flow can be expanded as

$$\varphi_h^f(y) = y + h f(y) + \frac{h^2}{2}(f_1'(y) f_1(y) + f_1'(y) f_2(y) + f_2'(y) f_1(y) + f_2'(y) f_2(y)) + \mathcal{O}(h^3)$$

it follows that the local error is given by

$$\left(\varphi_h^{f_2} \circ \varphi_h^{f_1}\right)(y) - \varphi_h^f(y) = \frac{h^2}{2}(f_2'(y) f_1(y) - f_1'(y) f_2(y)) + \mathcal{O}(h^3).$$

**Proposition 2.2.1**
Consider an ODE of the form (2.1), suppose $f$ can be split into $f = f_1 + f_2$ and consider the splitting integrator $\Phi_h = \varphi_h^{f_2} \circ \varphi_h^{f_1}$. Then $\Phi_h$'s approximations are at least of first order.

*Proof.* See the computations above. □

We say that the approximations are *at least* of first order, since in some cases the term

$$f_2'(y) f_1(y) - f_1'(y) f_2(y)$$

mentioned higher can be 0. If it is zero, then the approximations are at least of second order. Higher order terms can be computed similarly to the proof for this proposition and are subject to this same remark. From this proposition, we can now inductively show the following result for the case where $m > 2$.

**Proposition 2.2.2**
Consider an ODE of the form (2.1), suppose $f$ can be split into $f = f_1 + \ldots + f_m$ and consider the splitting integrator $\Phi_h = \varphi_h^{f_m} \circ \ldots \circ \varphi_h^{f_1}$. Then $\Phi_h$'s approximations are at least of first order.

*Proof.* By induction, we assume that the claim holds for some $m \in \mathbb{N}$ and we will prove that it must also be valid for $m+1$. By defining $\psi_h^k := \varphi_h^{f_k} \circ \ldots \circ \varphi_h^{f_1}$, we find that our prospective first order integrator can be written as $\psi_h^{m+1} = \varphi_h^{f_{m+1}} \circ \psi_h^m$. By the induction hypothesis, $\psi_h^m$ satisfies

$$\psi_h^m(y) = y + h \sum_{i=1}^{m} f_i(y) + \mathcal{O}(h^2)$$

and by definition as well as Taylor's theorem, it now follows that

$$\begin{aligned}\psi_h^{f_{m+1}}(y) &= \varphi_h^{f_{m+1}}\left(y + h\sum_{i=1}^{m} f_i(y) + \mathcal{O}(h^2)\right) \\ &= y + h\sum_{i=1}^{m} f_i(y) + \mathcal{O}(h^2) + hf_{m+1}\left(y + h\sum_{i=1}^{m} f_i(y) + \mathcal{O}(h^2)\right) \\ &= y + h\sum_{i=1}^{m} f_i(y) + hf_{m+1}(y) + \mathcal{O}(h^2) \\ &= y + hf(y) + \mathcal{O}(h^2)\end{aligned}$$

which proves the claim. □

By these two propositions, it follows that splitting schemes of the form $\varphi_h^{f_{i_1}} \circ \ldots \circ \varphi_h^{f_{i_m}}$ are always of first order if the same step size is used in each step. So in order to construct higher order integrators, we need to consider other types of compositions. As we will see in a next section, symmetrical compositions using varying step sizes are well-suited for the construction of higher order integrators..

### 2.2.2 Direct method

In this section, let $U$ and $V$ be complete vector fields (in the sense of definition 1.1.21) that are at least $n+1$ times continuously differentiable. We can assume that an $n$-th order splitting integrator is the product of exponentials (more precisely, lower order Lie-Trotter integrators), i.e. the integrator is given by

$$\mathrm{Int}_n(\tau) = \prod_{i=1}^{k} e^{c_i\tau U} e^{d_i\tau V} \tag{2.7}$$

for some integer $k \geq 1$ such that the exact solution satisfies the relation

$$e^{\tau(U+V)} = \mathrm{Int}_n(\tau) + \mathcal{O}(\tau^{n+1}). \tag{2.8}$$

as $\tau \downarrow 0$. By the BCH formula, (2.8) is equivalent to

$$\mathrm{Int}_n(\tau) = e^{\tau(U+V)+\mathcal{O}(\tau^{n+1})}.$$

We will now derive conditions on the pairs of real numbers $\{(c_i, d_i) \mid 1 \leq i, j \leq k\}$ in (2.7) such that (2.8) holds. Let us first consider the case $n = 1$. Since

$$e^{\tau(U+V)} = I + \tau(U+V) + \mathcal{O}(\tau^2) \tag{2.9}$$

it follows that

$$\begin{aligned}\mathrm{Int}_1(\tau) &= \prod_{i=1}^{k}(I + c_i\tau U)(I + d_i\tau V) + \mathcal{O}(\tau^2) \\ &= \prod_{i=1}^{k}(I + c_i\tau U + d_i\tau V) + \mathcal{O}(\tau^2) \\ &= \mathcal{P}_k + \mathcal{O}(\tau^2)\end{aligned}$$

where $\mathcal{P}_k := \prod_i (I + c_i\tau U + d_i\tau V)$. We can make the following observations.

- If $k = 1$, then $\mathcal{P}_1 = I + c_1\tau U + d_1\tau V$. A comparison of these coefficients to the exact solution's Taylor expansion (2.9) results in $c_1 = d_1 = 1$. This results in
$$\text{Int}_1(\tau) = e^{\tau U}e^{\tau V}$$
which is the Lie-Trotter integrator.
- If $k = 2$, then $\mathcal{P}_2 = I + \tau(c_1 + c_2)U + \tau(d_1 + d_2)V + \mathcal{O}(\tau^2)$ and a comparison of coefficients to (2.9) results in $c_1 + c_2 = 1$ and $d_1 + d_2 = 1$.

This observed pattern is the subject of the next proposition.

**Proposition 2.2.3 (Order conditions for $n = 1$)**
Any solution to the system of algebraic equations
$$\begin{cases} c_1 + c_2 + \ldots + c_k = 1 \\ d_1 + d_2 + \ldots + d_k = 1 \end{cases}$$
is a solution to (2.8). In other words, this choice of coefficients results in a first order integrator.

*Proof.* We follow the strategy proposed in [29, p. 30-33]. We start by considering the Taylor expansions around the origin of both the exact solution $\exp(\tau(U + V))$ and the prospective integrator $\text{Int}(\tau)$ as defined by (2.7). This results in the following identities
$$\begin{aligned} e^{\tau(U+V)} &= \sum_{i=0}^{n} \frac{\tau^i}{i!}(U + V)^i + \mathcal{O}(\tau^{n+1}) \\ \text{Int}(\tau) &= \sum_{i=0}^{n} \frac{\tau^i}{i!} \frac{\mathrm{d}^i}{\mathrm{d}\tau^i}\bigg|_{\tau=0} (\text{Int}(\tau)) + \mathcal{O}(\tau^{n+1}) \end{aligned}$$
as $\tau \downarrow 0$ where in both cases, the first term ($i = 0$) is equal to Id, the identity map. We do not yet fill in the value for $n = 1$ to make it clear how to extend this strategy to higher orders. To lighten notation, we will define a "truncated" version of the integrator by $\text{Int}_x^y(\tau)$, where the lower bound of $\text{Int}(\tau)$'s Taylor series is set to $x$ and the upper bound is set to $y$. The integer values for $0 < x, y \leq k$ must satisfy $y \geq x$. We can now compute the derivative of $\text{Int}(\tau) = \text{Int}_1^k(\tau)$ at $\tau = 0$:
$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau}\bigg|_{\tau=0} (\text{Int}(\tau)) &= \frac{\mathrm{d}}{\mathrm{d}\tau}\bigg|_{\tau=0} \left(e^{c_1\tau U}e^{d_1\tau V}e^{c_2\tau U}e^{d_2\tau V} \cdot \ldots \cdot e^{c_k\tau U}e^{d_k\tau V}\right) \\ &= c_1 U \cdot \text{Int}_1^k(\tau) + d_1 e^{c_1\tau U}Ve^{d_1\tau V} \cdot \text{Int}_2^k(\tau) \\ &\quad + \sum_{j=1}^{k-1} c_{j+1} \cdot \text{Int}_1^j(\tau) \cdot U \cdot \text{Int}_{j+1}^k(\tau) \\ &\quad + \sum_{j=1}^{k-1} d_{j+1} \cdot \text{Int}_1^j(\tau) \cdot e^{c_{j+1}\tau U}Ve^{d_{j+1}\tau V} \cdot \text{Int}_{j+2}^k(\tau)\bigg|_{\tau=0} \\ &= \left(\sum_{j=1}^{k} c_j\right) U + \left(\sum_{j=1}^{k} d_j\right) V \end{aligned}$$

since $\mathrm{Int}_x^y(0) = \mathrm{Id}$. Comparing these coefficients to the first order coefficients of the exact solution's Taylor expansion, we arrive at the result. □

Note that the proposition's system of linear equations is underdetermined when $k \geq 2$, resulting in infinitely many solutions. One only gets a unique solution if $k = 1$. We will now consider the case $n = 2$. The exact solution can be rewritten as

$$\begin{aligned} e^{\tau(U+V)} &= I + \tau(U+V) + \frac{\tau^2}{2}(U+V)^2 + \mathcal{O}(\tau^3) \\ &= I + \tau(U+V) + \frac{\tau^2}{2}(U^2+V^2) + \frac{\tau^2}{2}(UV+VU) + \mathcal{O}(\tau^3) \end{aligned} \tag{2.10}$$

as $\tau \downarrow 0$ from which it follows that the integrator satisfies the relation

$$\begin{aligned} \mathrm{Int}_2(\tau) &= \prod_{i=1}^{k} \left( I + c_i\tau U + d_i\tau V + \frac{c_i^2}{2}\tau^2U^2 + \frac{d_i^2}{2}\tau^2V^2 + c_id_i\tau^2UV \right) + \mathcal{O}(\tau^3) \\ &= \mathcal{P}_k + \mathcal{O}(\tau^3) \end{aligned}$$

where $\mathcal{P}_k := \prod_i \left( I + c_i\tau U + d_i\tau V + \frac{c_i^2}{2}\tau^2U^2 + \frac{d_i^2}{2}\tau^2V^2 + c_id_i\tau^2UV \right)$.

- If $k = 1$, then $\mathcal{P}_1 = I + c_1\tau U + d_1\tau V + \frac{c_1^2}{2}\tau^2U^2 + \frac{d_1^2}{2}\tau^2V^2 + c_1d_1\tau^2UV$ and a comparison of coefficients to (2.10) results in $c_1 = 1$, $d_1 = 1$, $c_1^2 = 1$, $d_1^2 = 1$, and $c_1d_1 = 1$.
- If $k = 2$, then

  $$\begin{aligned} \mathcal{P}_2 = I &+ \tau(c_1+c_2)U + \tau(d_1+d_2)V + \frac{\tau^2}{2}(c_1+c_2)^2U^2 + \frac{\tau^2}{2}(d_1+d_2)^2V^2 \\ &+ \tau^2(c_1(d_1+d_2) + c_2d_2)UV + \tau^2c_2d_1VU + \mathcal{O}(\tau^3) \end{aligned}$$

  and comparison of coefficients to (2.10) results in the system

  $$\begin{cases} c_1 + c_2 = 1 \\ d_1 + d_2 = 1 \\ c_1(d_1+d_2) + c_2d_2 = \frac{1}{2} \\ c_2d_1 = \frac{1}{2} \end{cases}$$

  where the first two equations immediately imply $(c_1+c_2)^2 = (d_1+d_2)^2 = 1$.

These findings indicate another pattern.

**Proposition 2.2.4 (Order conditions for $n = 2$)**
Any solution to the system of algebraic equations

$$\begin{cases} \sum_{i=1}^{k} c_i = \sum_{i=1}^{k} d_i = 1 \\ \sum_{i=1}^{k} c_i \sum_{j=i}^{k} d_j = \sum_{i=1}^{k-1} c_{i+1} \sum_{j=1}^{i} d_j = \frac{1}{2} \end{cases}$$

is a solution to (2.8). In other words, this choice of coefficients results in a first order integrator.

*Proof.* We follow the strategy proposed in [29, p. 30-33]. This proof is similar to the proof for Proposition 2.2.3. We denote by $\underline{\text{Int}}(\tau)$ the derivative with respect to $\tau$. Continuing our calculations from the last proof, we compute

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\bigg|_{\tau=0}(\text{Int}(\tau)) = c_1U\cdot\underline{\text{Int}}_1^k(0) + d_1V\cdot\underline{\text{Int}}_2^k(0) + (c_1d_1U + d_1^2V^2)\cdot\text{Int}_2^k(0)\\ + \sum_{j=1}^{k-1} d_{j+1}(\underline{\text{Int}}_1^j(0) + c_{j+1}\text{Int}_1^j(0)U)V\\ + \sum_{j=1}^{k-1} d_{j+1}V(d_{j+1}V + \underline{\text{Int}}_{j+2}^k(0))\\ + \sum_{j=1}^{k-1} c_{j+1}\underline{\text{Int}}_1^j(0)U + \sum_{j=1}^{k-1} c_{j+1}U\cdot\underline{\text{Int}}_{j+1}^k(0)\end{aligned}$$

where after re-ordering and re-indexing the summations, we arrive at

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}\bigg|_{\tau=0}(\text{Int}(\tau)) = \left(\sum_{i=1}^k c_i\right)^2U^2 + \left(\sum_{i=1}^k d_i\right)^2V^2\\ + 2\left(\sum_{i=1}^k c_i\sum_{j=i}^k d_j\right)UV + 2\left(\sum_{i=1}^{k-1} c_{i+1}\sum_{j=1}^i d_j\right)VU.\end{aligned}$$

Comparing these coefficients to the exact solution's Taylor series as $\tau\downarrow 0$

$$\begin{aligned}e^{\tau(U+V)} &= \text{Id} + \tau(U+V) + \frac{\tau^2}{2}(U+V)^2 + \mathcal{O}(\tau^3)\\ &= \text{Id} + \tau(U+V) + \frac{\tau^2}{2}(U^2 + UV + VU + V^2) + \mathcal{O}(\tau^3)\end{aligned}$$

leads to the result. □

The proofs of Propositions 2.2.3 and 2.2.4 prove that for increasing order $n$, the computations become increasingly more complex. It is theoretically possible to extend this approach to orders $n > 2$, but in practice this quickly becomes infeasible. Finally, we note that in the description of this direct method, we made the assumption that the prospective $n$-th order integrator (2.7) was the product of $k$ lower order Lie-Trotter integrators. This does not necessarily have to be the case: we can also assume that it is the product of $k$ lower order Strang integrators $S(\tau) = e^{\frac{\tau}{2}U}e^{\tau V}e^{\frac{\tau}{2}U}$, for example. This would result in different order conditions, see [3, chapter 2].

### 2.2.3 Using the BCH-formula

The BCH-formula can be used to recursively construct symplectic integrators of even order. The techniques that are used can also directly be used to find odd order conditions: the method consists of a repeated application of the BCH-formula. To demonstrate this, we will revisit the order conditions for the case $n = 1$.

**Proposition 2.2.5 (Reworked Proposition 2.2.3)**
Any solution to the system of algebraic equations

$$\begin{cases} c_1 + c_2 + \ldots + c_k = 1 \\ d_1 + d_2 + \ldots + d_k = 1 \end{cases}$$

is a solution to (2.8). In other words, this choice of coefficients results in a first order integrator.

*Proof.* To lighten notations, we will write $\mathrm{Int}^k(\tau) = \prod_{i=1}^{k} e^{c_i \tau U} e^{d_i \tau V}$. Then (2.7) becomes

$$\begin{aligned}
\mathrm{Int}_1(\tau) &= \mathrm{Int}^{k-1}(\tau) e^{c_k \tau U} e^{d_k \tau V} \\
&= \mathrm{Int}^{k-1}(\tau) \exp(\tau c_k U + \tau d_k V + \mathcal{O}(\tau^2)) \\
&= \mathrm{Int}^{k-2}(\tau) e^{\tau c_{k-1} U} e^{\tau d_{k-1} V} \exp(\tau c_k U + \tau d_k V + \mathcal{O}(\tau^2)) \\
&= \mathrm{Int}^{k-2}(\tau) \exp(\tau c_{k-1} U + \tau d_{k-1} V + \mathcal{O}(\tau^2)) \exp(\tau c_k U + \tau d_k V + \mathcal{O}(\tau^2)) \\
&= \mathrm{Int}^{k-2}(\tau) \exp(\tau (c_{k-1} + c_k) U + \tau (d_{k-1} + d_k) V + \mathcal{O}(\tau^2)) \\
&= \ldots \\
&= \exp\left( \tau U \sum_{i=1}^{k} c_i + \tau V \sum_{i=1}^{k} d_i + \mathcal{O}(\tau^2) \right)
\end{aligned}$$

which, after comparing the coefficients in the exponent, implies the result. □

The BCH-formula can also be applied in the context of Hamiltonian systems with Hamiltonian functions of the form $H(q,p) = T(p) + V(q)$. We will use it to recursively construct symplectic integrators of order $2n+2$ by symmetrically composing integrators of order $2n$. The remainder of this section is based on [34].

**Definition 2.2.6**
Let $\mathrm{Int}_n(\tau)$ be an integrator of the form $\prod_{i=1}^{k-1} e^{\tau c_i X_V} e^{\tau c_{i+1} X_T}$.

1. The integrator is said to be a **symmetric composition** if for any $j \in \{1, 2, \ldots, k\}$ such that $c_j \neq 0$, it holds that $c_j = c_{k-j+1}$.
2. The integrator is called **reversible in time** if $\mathrm{Int}_n(\tau)\mathrm{Int}_n(-\tau) = \mathrm{Id}$, or equivalently $\mathrm{Int}_n(-\tau)\mathrm{Int}_n(\tau) = \mathrm{Id}$. Equivalently, the integrator is called **symmetric** if $\mathrm{Int}_n(\tau) = \mathrm{Int}_n(-\tau)^{-1}$.

The construction starts from a second order integrator

$$\mathcal{Y}_2(\tau) := e^{\frac{\tau}{2} X_V} e^{\tau X_T} e^{\frac{\tau}{2} X_V}, \tag{2.11}$$

hence it is of the form (2.7) with $c_1 = c_2 = \frac{1}{2}$, $d_1 = 1$ and $d_2 = 0$. By Proposition 2.2.4 it follows that it is indeed a second order integrator. An application of the BCH-formula shows that it can be written in the form

$$\mathcal{Y}_2(\tau) = \exp(\tau \alpha_1 + \tau^3 \alpha_3 + \tau^5 \alpha^5 + \ldots)$$

where $\alpha_1 = X_V + X_T$, $\alpha_3 = \frac{1}{12}[X_T, X_T, X_V] - \frac{1}{24}[X_V, X_V, X_T]$, $\alpha_5 = \frac{7}{5760}[X_V, X_V, X_V, X_V, X_T] + \ldots$ and so on. These higher order commutators are defined as follows.

$$[X_T, X_T, X_V] := [X_T, [X_T, X_V]]$$
$$[X_V, X_V, X_V, X_V, X_T] := [X_V, [X_V, [X_V, [X_V, X_T]]]]$$

Moreover, $\mathcal{Y}_2(\tau)$ is clearly symmetric (in both senses) and reversible in time.

**Lemma 2.2.7**
Suppose $\mathrm{Int}_n(\tau)$ is an integrator of the form 2.2.4 that is reversible in time. Then its expansion $\mathrm{Int}_n(\tau) = \exp(\tau\beta_1 + \tau^2\beta_2 + \tau^3\beta_3 + \ldots)$ is such that $\beta_{2i} = 0$ for any $i \in \mathbb{N}$.

*Proof.* Since the integrator is reversible in time, we know that $\mathrm{Int}_n(\tau)\mathrm{Int}_n(-\tau) = \mathrm{Id}$. Applying the BCH-formula to this identity as $\tau \downarrow 0$ results in

$$\mathrm{Int}_n(\tau) \cdot \mathrm{Int}_n(-\tau) = \mathrm{Id} = e^0 = \exp(\tau^2\beta_2 + \mathcal{O}(\tau^3))$$

which means that $\beta_2 = 0$. Another application of the BCH-formula as $\tau \downarrow 0$ leads to

$$\mathrm{Int}_n(\tau) \cdot \mathrm{Int}_n(-\tau) = \mathrm{Id} = \exp(0) = \exp(\tau^2\beta_4 + \mathcal{O}(\tau^5))$$

which again means that $\beta_4 = 0$. Repeated applications of the BCH-formula show that all even order coefficients are zero. □

Any symplectic integrator that is reversible in time is necessarily of even order. It is used in the recursive construction of higher order symplectic integrators, as we will demonstrate below. First, we will consider the case of a fourth order integrator.

**Theorem 2.2.8 (Fourth order integrator)**
There exist constants $x_0, x_1 \in \mathbb{R}$ such that the symmetric composition

$$\mathcal{Y}_4(\tau) := \mathcal{Y}_2(x_0\tau)\mathcal{Y}_2(x_1\tau)\mathcal{Y}_2(x_0\tau)$$

is a fourth order integrator, i.e. $e^{\tau(U+V)} = \mathcal{Y}_4(\tau) + \mathcal{O}(\tau^5)$ as $\tau \downarrow 0$. More precisely

$$x_0 = \frac{1}{2 - \sqrt[3]{2}} \text{ and } x_1 = -\frac{\sqrt[3]{2}}{2 - \sqrt[3]{2}}.$$

*Proof.* As mentioned earlier, the second order integrators $\mathcal{Y}_2(x_0\tau)$ and $\mathcal{Y}_2(x_1\tau)$ are symmetric and reversible in time. By Lemma 2.2.7, their expansions are of the forms

$$\mathcal{Y}_2(x_j\tau) = \exp(\alpha_1 x_j \tau + \alpha_3 \tau^3 x_j^3 + \alpha_5 \tau^5 x_j^5 + \mathcal{O}(\tau^6))$$

where the $\alpha_i$ still denote the aforementioned commutators of $U$ and $V$. We will now apply the BCH-formula to the symmetric composition of these second order integrators. This results in

$$\mathcal{Y}_4(\tau) = \exp((2x_0 + x_1)\alpha_1\tau + (2x_0^3 + x_1^3)\alpha_3\tau^3 + \mathcal{O}(\tau^5)) = \exp(\alpha_1\tau + \mathcal{O}(\tau^5))$$

where the second equality must old in order for $\mathcal{Y}_4$ to be of fourth order. A comparison of coefficients in the exponents leads to the system of algebraic equations

$$\begin{cases} 2x_0 + x_1 = 1 \\ 2x_0^3 + x_1^3 = 0 \end{cases}$$

which has the unique real solution

$$x_0 = \frac{1}{2-\sqrt[3]{2}} \text{ and } x_1 = -\frac{\sqrt[3]{2}}{2-\sqrt[3]{2}},$$

finishing the proof. □

We can use this result to formulate an answer to the initial problem (2.7) of finding coefficients $(c_i, d_i)$. If we fully write out the triple composition of second order integrators in the definition of $\mathcal{Y}_4(\tau)$, we find

$$\mathcal{Y}_4(\tau) = e^{\frac{x_0}{2}\tau X_V} e^{x_0\tau X_T} e^{\frac{x_0+x_1}{2}\tau X_V} e^{x_1\tau X_T} e^{\frac{x_0+x_1}{2}\tau X_V} e^{x_0\tau X_T} e^{\frac{x_0}{2}\tau X_V}$$

which means that $c_1 = c_4 = \frac{x_0}{2}$, $d_1 = d_3 = x_0$, $c_2 = c_3 = \frac{x_0+x_1}{2}$, and $d_4 = 0$. Clearly, the integrator is still symmetric by Definition 2.2.6. We can now analogously construct a sixth order integrator by using a symmetric composition of $\mathcal{Y}_4$ instead of $\mathcal{Y}_2$. That is, the sixth order integrator is of the form

$$\mathcal{Y}_6(\tau) = \mathcal{Y}_4(y_0\tau)\mathcal{Y}_4(y_1\tau)\mathcal{Y}_4(y_0\tau) \tag{2.12}$$

and using the BCH-formula, it can be shown that

$$y_0 = \frac{1}{2-\sqrt[5]{2}} \text{ and } y_1 = -\frac{\sqrt[5]{2}}{2-\sqrt[5]{2}}.$$

By fully writing out (2.12), we find that $c_1 = c_{10} = \frac{x_0y_0}{2}$, $c_2 = c_3 = c_8 = c_9 = \frac{x_0+x_1}{2}y_0$, $c_5 = c_6 = \frac{x_0+x_1}{2}y_1$, $c_4 = c_7 = x_0\frac{y_0+y_1}{2}$, $d_1 = d_3 = d_7 = d_9 = x_0y_0$, $d_2 = d_8 = x_1y_0$, $d_4 = d_6 = x_0y_1$, $d_5 = x_1y_1$, and $d_{10} = 0$.

**Theorem 2.2.9 (Yoshida's even order integrators)**
Suppose $\mathcal{Y}_2(\tau) = e^{\frac{\tau}{2}X_V} e^{\tau X_T} e^{\frac{\tau}{2}X_V}$. Then for any integer $n \geq 1$, there exist some constants $z_0(n), z_1(n) \in \mathbb{R}$ such that the recursion

$$\mathcal{Y}_{2n+2}(\tau) = \mathcal{Y}_{2n}(z_0(n)\tau)\mathcal{Y}_{2n}(z_1(n)\tau)\mathcal{Y}_{2n}(z_0(n)\tau)$$

defines integrators of order $2n+2$, i.e. $e^{\tau(X_V+X_T)} = \mathcal{Y}_{2n+2}(\tau) + \mathcal{O}(\tau^{2n+3})$ as $\tau \downarrow 0$. More precisely

$$z_0(n) = \frac{1}{2-\sqrt[2n+1]{2}} \text{ and } z_1(n) = -\frac{\sqrt[2n+1]{2}}{2-\sqrt[2n+1]{2}}.$$

*Proof.* This proof is completely analogous to the proof of Theorem 2.2.8. □

Theorem 2.2.9 tells us that starting from $\mathcal{Y}_2(\tau)$, we can construct an integrator of arbitrary even order $n > 2$. However, this is still impractical to implement for orders higher than 4. If we take a look at the fourth order integrator $\mathcal{Y}_4(\tau)$, we can see that it consists of a triple composition of the second order integrators $\mathcal{Y}_2(\tau)$, which is nothing more than a composition of 7 exponentials, hence it needs $k = 4$ steps in (2.7). For higher order integrators, the same analysis shows that the even order integrators $\mathcal{Y}_{2m}(\tau)$ require the second order integrator a total amount of $3^{n-1}$ times such that a total amount of $k = 3^{n-1} + 1$ steps are needed. Clearly, the amount of

steps increase exponentially as the order grows larger. This means that in practice, Yoshida's integrators are very computationally expensive for higher orders. To avoid this, Yoshida introduced the following strategy in [34] to arrive at 6th and 8th order integrators that require less steps, and therefore less computations. His main idea is to consider a symmetrical integrator of the form with unknown constants $w_0, w_1, \ldots, w_m$

$$\mathcal{Y}_m(\tau) = \mathcal{Y}_2(w_m\tau)\mathcal{Y}_2(w_{m-1}\tau)\cdot\ldots\cdot\mathcal{Y}_2(w_0\tau)\cdot\ldots\cdot\mathcal{Y}_2(w_{m-1}\tau)\mathcal{Y}_2(w_m\tau)$$

which can once again be rewritten as a single exponential

$$\mathcal{Y}_m(\tau) = \exp(\tau A_{1,m}\alpha_1 + \tau^3 A_{3,m}\alpha_3 + \tau^5(A_{5,m}\alpha_5 + B_{5,m}\beta_5) + \ldots)$$

which can be used to recursively define the higher order integrator

$$\mathcal{Y}_{m+1}(\tau) = \mathcal{Y}_2(w_{m+1})\mathcal{Y}_m(\tau)\mathcal{Y}_2(w_{m+1}).$$

An application of the BCH-formula and comparing the coefficients in the exponent to arrive at order conditions, we arrive at recursive formulas for the coefficients $A_{i,m+1}$, $B_{i,m+1}$ and so on appearing in the newly obtained exponent. This results in a system algebraic equations, which has to be solved to find the coefficients $w_i$. Since the resulting system is not analytically solvable in general, numerical methods must be used and the coefficients cannot be given exactly. This method reduces the steps needed in the implementation of Yoshida's 6th and 8th order integrators. For example, only 8 resp. 16 steps are now needed, whereas before $3^2 + 1 = 10$ resp. $3^3 + 1 = 28$ steps were needed. For further details, see [34].

## 2.3 Linear stability analysis

This section is based on [2]. We will study the linear stability of splitting methods of the form (2.7).

**Definition 2.3.1 ([2], p. 2)**
Consider a system (2.1) that has a bounded exact solution on some interval $I$. Then an integrator is called **stable** if its numerical solution over the interval $I$ does not tend to infinity.

The largest possible interval on which the integrator is stable in the sense of the definition above, is called the stability interval of the integrator. A more precise definition of this interval will be provided later. From now on, we focus on the *linear stability* of splitting integrators. That is, we will apply them to a model linear system to base our findings upon. The reasoning behind this is as follows: if a given splitting method with step size $h > 0$ already shows bad (i.e. unbounded) behaviour on this "toy" example, then it cannot be expected to show good (i.e. bounded) behaviour for more complicated systems [3, p. 61]. The toy model that we will use is nothing more than the harmonic oscillator

$$\ddot{y} + \lambda^2 y = 0 \text{ with } \lambda > 0$$

which can be rewritten as a system of two first order ODEs as folllows

$$\begin{pmatrix}\dot{q}\\ \dot{p}\end{pmatrix} = \begin{pmatrix}0 & \omega\\ -\omega & 0\end{pmatrix}\begin{pmatrix}q\\ p\end{pmatrix} = \left[\begin{pmatrix}0 & \omega\\ 0 & 0\end{pmatrix} + \begin{pmatrix}0 & 0\\ -\omega & 0\end{pmatrix}\right]\begin{pmatrix}q\\ p\end{pmatrix} \tag{2.13}$$

where we denote the left splitting matrix by $\mathbf{\Omega}_1$ and the right one by $\mathbf{\Omega}_2$. Its bounded exact solution is given by

$$\begin{pmatrix} q(t) \\ p(t) \end{pmatrix} = O(\lambda t) \begin{pmatrix} q_0 \\ p_0 \end{pmatrix} \text{ with } O(\omega t) := \begin{pmatrix} \cos(\lambda t) & \sin(\lambda t) \\ -\sin(\lambda t) & \cos(\lambda t) \end{pmatrix}. \tag{2.14}$$

By applying an integrator (2.7) to (2.13), we obtain the so-called *stability matrix* of the splitting integrator, from which we can extract the corresponding *stability function*. As an example, consider the Lie-Trotter integrator $\varphi_h^{\mathrm{LT}} = \varphi_h^2 \circ \varphi_h^1$ where $\varphi_h^1$ resp. $\varphi_h^2$ denotes the flow of (2.13) with respect to $\mathbf{\Lambda}_1$ resp. $\mathbf{\Lambda}_2$. By computing these compositions, we find

$$\varphi_h^{\mathrm{LT}}(q_0, p_0) = K(x) \begin{pmatrix} q_0 \\ p_0 \end{pmatrix} \text{ with } \begin{pmatrix} 1 - x^2 & x \\ -x & 1 \end{pmatrix}$$

where $x := \lambda t$ and we set $p_h^{\mathrm{LT}}(x) = \frac{1}{2}(2 - x^2)$.

**Definition 2.3.2**
The **stability matrix** of an integrator (2.7) is the matrix

$$K(x) = e^{hc_k \mathbf{\Lambda}_1} e^{hd_k \mathbf{\Lambda}_2} \dots e^{hc_2 \mathbf{\Lambda}_1} e^{hd_2 \mathbf{\Lambda}_2} e^{hc_1 \mathbf{\Lambda}_1} e^{hd_1 \mathbf{\Lambda}_2}$$

and the **stability function** $p(x)$ of (2.7) is defined as half the trace of $K(x)$, i.e. $p(x) = \frac{1}{2}\mathrm{Tr}(K(x))$.

Since $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are both nilpotent matrices of order 2, their matrix exponentials are equal to $\boldsymbol{I} + \mathbf{\Lambda}_1$ and $\boldsymbol{I} + \mathbf{\Lambda}_2$. As a result, $K(x)$ is the product of $k$ upper triangular matrices $e^{hc_j \mathbf{\Lambda}_1}$ and $k$ lower triangular matrices $e^{hd_j \mathbf{\Lambda}_2}$. Furthermore, it can be inductively proven that if we write the stability matrix as

$$K(x) = \begin{pmatrix} K_1(x) & K_2(x) \\ K_3(x) & K_4(x) \end{pmatrix}$$

then $K_1(x)$ and $K_4(x)$ are even polynomials in $x = \omega t$ (in the sense that $K_{1,4}(-x) = K_{1,4}(x)$, and $K_2(x)$ and $K_3(x)$ are odd polynomials in $x$ ($K_{2,3}(-x) = -K_{2,3}(x)$). Moreover, $\det K(x) = 1$, $K_1(0) = K_4(0) = 1$ and $K_2'(0) = K_3'(0) = 0$ since we are only considering consistent integrators.
From the definition of stability, it follows [2, p. 7] that an integrator is stable when $K(x)^n$ can be bounded independently of the power $n \geq 1$ and if the method is stable, then $|\, p(x) \,| \leq 1$. Unfortunately, the converse is not true, as shown by the following example.

$$K(x) = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \Rightarrow K(x)^n = \begin{pmatrix} 1 & nx \\ 0 & 1 \end{pmatrix}$$

We need further conditions to be satisfied in order for a splitting integrator to be stable. We present the following result that offers a characterisation of linear stability for splitting methods.

**Proposition 2.3.3 (Blanes et al. [2], p. 8, proposition 2.2)**
Suppose $K(x)$ is a $2 \times 2$ matrix with determinant $\det K(x) = 1$ for any $x$ and

set $p(x) = \frac{1}{2}\mathrm{Tr}(K(x))$. Then the matrix $K(x)$ is stable if and only if one of the following conditions is satisfied

1. $K(x)$ is diagonalisable and its eigenvalues have modulus 1;
2. The stability polynomial satisfies $\mid p(x) \mid \leq 1$ and $K(x)$ is similar to the matrix

$$M(x) := \begin{pmatrix} \cos(\arccos(p(x))) & \sin(\arccos(p(x))) \\ -\sin(\arccos(p(x))) & \cos(\arccos(p(x))) \end{pmatrix}. \tag{2.15}$$

*Proof.* Suppose $K(x)$ is stable and let $\lambda_1(x)$ and $\lambda_2(x)$ be its eigenvalues. By assumption, it then follows that $\det K(x) = \lambda_1(x)\lambda_2(x) = 1$ from which it follows that $|\lambda_1(x)| \cdot |\lambda_2(x)| = 1$. By the positivity of the modulus function, the only possibility is now $|\lambda_1(x)| = |\lambda_2(x)| = 1$. Furthermore, $K(x)$ is diagonalisable unless $\lambda_1(x) = \lambda_2(x) = \pm 1$, in which case $K(x)$ would cause linear growth and would no longer be stable. So, if $K(x)$ is stable, then it is diagonalisable and its eigenvalues have modulus 1, from which it follows that

$$|p(x)| = \frac{1}{2}|\mathrm{Tr}(K(x))| = \frac{1}{2}|\lambda_1(x) + \lambda_2(x)| \leq \frac{1}{2}(|\lambda_1(x)| + |\lambda_2(x)|) = 1.$$

Since the eigenvalues of $K(x)$ satisfy $\lambda - \mathrm{Tr}(K(x))\lambda + 1 = 0$, it follows that

$$\lambda_{1,2}(x) = p(x) \pm i\sqrt{1 - p(x)^2}$$

such that both eigenvalues lie on the complex unit circle, so they can be written as $e^{i\vartheta} = \cos\vartheta + i\sin\vartheta$, it follows that $p(x) = \cos\vartheta(x)$ and $\vartheta(x) = \arccos(p(x))$. It now immediately follows that $K(x)$ is similar to (2.15). Conversely, the second condition immediately implies the stability of $K(x)$. □

We can now give the definition of a splitting integrator's stability interval.

**Definition 2.3.4**
Let $K(x)$ be a matrix as described in Proposition 2.15. If $x^* > 0$ is the largest number such that $K(x)$ is stable for any $x = h\lambda \in \mathcal{I}_{\mathrm{stab}} := [-x^*, x^*]$ , then $x^*$ is called the **stability threshold** of $K(x)$ and is called the **stability interval** of $K(x)$.

As an example, we found that the stability polynomial of the Lie-Trotter integrator is $p_h^{\mathrm{LT}}(x) = \frac{1}{2}(2 - x^2)$. Hence its stability interval is given by $[-2, 2]$ since

$$|p_h^{\mathrm{LT}}(x)| \leq 1 \Leftrightarrow |2 - x^2| \leq 2 \Leftrightarrow x \in [-2, 2]$$

which means that Lie-Trotter is only (linearly) stable for step sizes $h \in ]0, 2]$.

## 2.4 Geometric numerical integration

In classical numerical integration, the main objective is usually to approximate the solution to a differential equation

$$\dot{y}(t) = f(t, y(t)) \text{ with } y(0) = y_0$$

as accurately as possible. In applying a numerical integrator to this differential equation, one often studies its local error (the error after a single iteration), its global error (the accumulated error after multiple iterations), and the integrator's stability and consistency to get a sense of accuracy for the integrator. Although this approach has proven to be fruitful in delivering reliable numerical schemes, it does not take into account the underlying geometry that the differential equation (2.4) may have. For example, as listed in [3], Hamiltonian systems have an underlying symplectic structure, some systems in quantum mechanics have an underlying unitary structure, other mechanical systems have the additional properties of time reversibility or volume preservation, and under certain conditions dissipative systems can be encoded in a contact geometric framework (see [4]). The search for and study of integrators that conserve these geometrical properties is referred to as geometric numerical integration. In this section, we will consider splitting schemes as an important class of geometrical integrators and we will study their properties as well as their errors.

### 2.4.1 Motivation

This section is based on [35]. We consider the harmonic oscillator, whose Hamiltonian is given by $H(q,p) = \frac{1}{2}(q^2+p^2)$ for $q,p \in \mathbb{R}$ resulting in the equations of motion

$$\begin{cases} \dot{q} = p \\ \dot{p} = -q \end{cases} \tag{2.16}$$

with initial values $(q_0, p_0)$ and with the exact solution (2.14) for $\lambda = 1$. We can now apply the forward (explicit) Euler integrator to the system (2.16). Its numerical flow is given by the map

$$\Phi_h^{\mathrm{FE}} : \begin{pmatrix} q_n \\ p_n \end{pmatrix} \mapsto \begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \begin{pmatrix} q_n + hp_n \\ p_n - hq_n \end{pmatrix} \tag{2.17}$$

where $h > 0$ is the chosen step size. If we compute the energy associated to each iteration of this integrator, we find that

$$H(q_{n+1}, p_{n+1}) = \frac{1}{2}((q_n + hp_n)^2 + (p_n - hq_n)^2) = (1+h^2)H(q_n, p_n)$$

which means that in each iteration of the forward Euler scheme, the energy is multiplied by a strictly positive factor $1+h^2$. Since $H$ is autonomous, we know that the total energy (i.e. the Hamiltonain itself) is a conserved quantity. Therefore in each iteration the total energy should be equal to

$$H(q_0, p_0) = \frac{1}{2}(q_0^2 + p_0^2).$$

This immediately shows why the forward Euler integrator is not suitable for this system: it does not conserve $H$, it leads to an *indefinite increase of energy* [35].
Ideally, an integrator applied to a Hamiltonian system preserves both the total energy $H$ and the underlying symplectic structure, but this has been proven to be impossible for non-integrable Hamiltonian systems. An integrator can either preserve the total energy *or* the symplectic structure, but it cannot do both [36,

section 3]. As a compromise, one has to choose one of these properties that the integrators have to conserve. Integrators that preserve the symplectic structure (i.e. the symplectic form $\omega$) are called symplectic integrators.

**Definition 2.4.1**
A **symplectic integrator** is a numerical integrator Int that preserves the symplectic structure of a dynamical system. The numerical flow of a symplectic integrator is a symplectic map, i.e. $\text{Int}^*\omega = \omega$ where $\omega$ denotes the symplectic form being considered.

The discussion at the start of this section indicated that the forward Euler integrator is not symplectic.

**Proposition 2.4.2**
Forward Euler is not a symplectic integrator.

*Proof.* For the numerical flow $\Phi_h^{\text{FE}}(q,p) = (q+hp, p-hq)$ to be symplectic, it needs to satisfy the identity $(\Phi_h^{\text{FE}})^*(dq \wedge dp) = dq \wedge dp$. We compute

$$\begin{aligned}(\Phi_h^{\text{FE}})^*(dq \wedge dp) &= (\Phi_h^{\text{FE}})^*(dq) \wedge (\Phi_h^{\text{FE}})^*(dp) \\ &= d(q \circ \Phi_h^{\text{FE}}) \wedge d(p \circ \Phi_h^{\text{FE}}) \\ &= d(q+hp) \wedge d(p-hq) \\ &= dq \wedge dp - hdq \wedge dq + hdp \wedge dp + h^2 dq \wedge dp \\ &= (1+h^2) dq \wedge dp\end{aligned}$$

from which the result immediately follows, since the step size $h$ is always strictly positive. □

Yoshida's integrators as constructed in Theorem 2.2.9 are symplectic.

**Proposition 2.4.3**
The integrators $\mathcal{Y}_{2n+2}$ are symplectic.

*Proof.* By Propositions 1.1.11 and 1.1.15, the composition of symplectomorphisms is a symplectomorphism and the flow of a (complete) Hamiltonian vector field is a symplectic map at any time $t \in \mathbb{R}$. As a result, the second order integrator (2.11) is symplectic, because it is the product (composition) of Hamiltonian flows. It then follows by induction that for any $n \in \mathbb{N}$, the integrator $\mathcal{Y}_{2n+2}(\tau)$ is symplectic. □

Splitting integrators form a class of geometric integrators. This is because splitting schemes are compositions of maps that preserve some geometric property, such as symplecticity (see Proposition 1.1.11) in the Hamiltonian setting. Our focus will be on symplectic integrators applied to systems with a "separable" Hamiltonian.

**Definition 2.4.4**
A Hamiltonian function is called **kinetic-potential separable** or **T-V separable** if it can be written in the form $H(q,p) = T(p) + V(q)$. If this is the case, we will call $H$ **separable** for short.

In the next section, we will discuss examples of symplectic integrators for separable Hamiltonians.

### 2.4.2 Examples of symplectic integrators

In section 2.2.3, we already discussed a class of symplectic splitting methods, namely Yoshida's even order integrators. We will now consider two more examples of symplectic integrators, which we will later on use in some numerical simulations. If we consider a separable Hamiltonian $H(q,p) = T(p) + V(q)$, then the corresponding Hamiltonian system is of the form

$$\begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \frac{\partial T}{\partial p} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ -\frac{\partial V}{\partial q} \end{pmatrix} \text{ with } (q(0), p(0)) = (q_0, p_0) \tag{2.18}$$

where each subsystem is Hamiltonian in its own right, with flows

$$\Phi_h^T(q_0,p_0) := \begin{pmatrix} q_0 + h\frac{\partial T}{\partial p}(p_0) \\ p_0 \end{pmatrix} \text{ and } \Phi_h^V(q_0,p_0) := \begin{pmatrix} q_0 \\ p_0 - h\frac{\partial V}{\partial q}(q_0). \end{pmatrix}$$

The construction of splitting methods for the system (2.18) now consists of choosing a composition of these two flows to approximate the exact flow $\Phi_h^H$ of the system. Choosing the compositions $\Phi_h^V \circ \Phi_h^T$ and $\Phi_{\frac{h}{2}}^V \circ \Phi_h^T \circ \Phi_{\frac{h}{2}}^V$, we can compute that these integrators are given by the respective maps down below.

$$(\Phi_h^V \circ \Phi_h^T)(q_0,p_0) = \begin{pmatrix} q_0 + h\frac{\partial T}{\partial p}(p_0) \\ p_0 - h\frac{\partial V}{\partial q}(q_0) \end{pmatrix}$$

$$(\Phi_{\frac{h}{2}}^V \circ \Phi_h^T \circ \Phi_{\frac{h}{2}}^V)(q_0,p_0) = \begin{pmatrix} q_0 + h\frac{\partial T}{\partial p}(p_0) \\ p_0 - \frac{h}{2}\frac{\partial V}{\partial q}(q_0) - \frac{h}{2}\frac{\partial V}{\partial q}\left(q_0 + h\frac{\partial T}{\partial p}(p_0)\right) \end{pmatrix}$$

These integrators are called symplectic Euler with VT-splitting resp. Stormer-Verlet with VTV-splitting. If we choose a different composition, for example $\Phi_h^T \circ \Phi_h^V$ and $\Phi_h^T \circ \Phi_h^V \circ \Phi_h^T$, we would say that the integrators use TV- resp. TVT-splitting. They are all symplectic integrators, since they are compositions of Hamiltonian flows.

**Proposition 2.4.5**
The symplectic Euler integrator is of first order and the Stormer-Verlet integrator is of second order.

*Proof.* Note that the symplectic Euler VT integrator is of the form (2.7) with $c_1 = 1$ and $d_1 = 1$ and Stormer-Verlet VTV has coefficients $c_1 = c_2 = \frac{1}{2}$, $d_1 = 1$ and $d_2 = 0$. By Proposition 2.2.3, symplectic Euler VT is of first order and by Proposition 2.2.4, Stormer-Verlet VTV is of second order. The same result holds for symplectic Euler TV and Stormer-Verlet TVT. □

Splitting integrators are not the only class of symplectic integrators. In fact, there are various examples of symplectic integrators that are not splitting schemes. As an example, we consider the following result on Runge-Kutta integrators.

**Theorem 2.4.6 (Hairer, Lubich, Wanner [12], Theorem 4.3)**
An $m$-stage Runge-Kutta integrator is symplectic if its coefficients (A.1) satisfy the identity

$$b_s a_{st} + b_t a_{ts} = b_s b_t$$

for any indices $s, t \in \{1, \dots, m\}$.

If we now consider the well-known implicit midpoint rule, whose Runge-Kutta matrix is given by

$$M_1 = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$$

then we find that

$$b_1 a_{11} + b_1 a_{11} = 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 1 = b_1 b_1$$

which implies the integrator's symplecticity. Although there exist several other classes of symplectic integrators, we will not discuss them, since our focus is on splitting methods.

### 2.4.3 Backward error analysis

Backward error analyses are useful in the study of geometric integrators, since they offer results on the conserving properties of geometrical integrators and their long term behaviour. Instead of studying the local and global errors, a modified differential equation is studied that is in some sense *close* to the original problem and that is additionally equal to the integrator's numerical flow. These equations can then be used to explain the near-conservation of energy exhibited by symplectic integrators. Choosing backward error analyses when considering a problem with some interesting geometrical structure intuitively makes sense, as for these problems we are not primarily interested in an integrator's pointwise errors, but rather in how the integrator captures the geometrical context of the problem at hand. This section is heavily based on [12].

#### 2.4.3.1 Modified differential equations

We consider ODEs of the form (2.1). We denote by $\varphi_t$ the ODE's flow after a time $t$, and we denote by $\Phi_h$ the numerical flow $y_{n+1} = \Phi_h(y_n)$ of some integrator with step size $h > 0$ applied to the ODE. A backward error analysis applied to (2.1) consists of looking for a *modified differential equation* of the integrator, which is of the form

$$\dot{z} = f(z) + h f_2(z) + h^2 f_3(z) + \dots \tag{2.19}$$

such that it coincides with the ODE's numerical solution: $y_n = \Phi_h(y_{n-1}) = z(nh)$. Since this series generally does not converge, it needs to truncated, which introduces errors in the results. A next step in the backward error analysis is then to study the difference of the vector fields $f$ and $f_h$. We will now determine the coefficients $f_j(z)$ in the modified equation (2.19). To this aim, we fix $t \in \mathbb{R}$ and compute the Taylor

expansion of $z$

$$\begin{aligned} z(t+h) &= z(t) + h\frac{\mathrm{d}z}{\mathrm{d}t} + \frac{h^2}{2!}\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} + \frac{h^3}{3!}\frac{\mathrm{d}^3 z}{\mathrm{d}t^3} + \dots \\ &= z(t) + h(f(z) + hf_2(z) + h^2 f_3(z) + \dots) \\ &\quad + \frac{h^2}{2}(f'(z) + hf_2'(z) + h^2 f_3'(z) + \dots)\frac{\mathrm{d}z}{\mathrm{d}t} \\ &\quad + \frac{h^3}{6}(f''(z) + hf_2''(z) + h^2 f_3''(z) + \dots)\frac{\mathrm{d}z}{\mathrm{d}t} \\ &\quad + \frac{h^3}{6}(f'(z) + hf_2'(z) + h^2 f_3'(z) + \dots)\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} + \dots \end{aligned} \tag{2.20}$$

and we assume the integrator can be expanded as follows

$$\Phi_h(y) = y + hd_1(y) + h^2 d_2(y) + h^3 d_3(y) + \dots \tag{2.21}$$

for some dummy variable $y$. For the integrator to be *consistent*, the first coefficient must be $d_1(y) = f(y)$, since that is when the integrator will locally agree with the exact solution's Taylor expansion (i.e. it is at least of order 1). By comparing the coefficients of $h^2$ and $h^3$ in (2.20) and (2.21), we find

$$\begin{aligned} f_2(z) &= d_2(z) - \frac{1}{2}f'(z)f(z) \\ f_3(z) &= d_3(z) - \frac{1}{2}(f_2'(z)f(z) + f'(z)f_2(z)) - \frac{1}{6}(f''(z)f(z)^2 + f'(z)^2 f(z)) \end{aligned}$$

and the higher order coefficients $f_j(z)$ for $j > 3$ can be computed in a similar fashion. As an example of how modified equations can be computed, we provide the following example.

**Example**

We consider the uncoupled harmonic oscillator with Hamiltonian function $H(q,p) = \frac{1}{2}(q^2 + p^2)$. The equations of motion are given by the following system of ODEs.

$$\begin{cases} \dot{q} = p \\ \dot{p} = -q \end{cases}$$

We will compute the modified differential equation of the explicit Euler integrator.

For the forward Euler integrator, we have $\Phi_h(z) = z + hf(z)$, so $d_2(z) = 0$. To compute $f_2(z)$ we perform the following calculations

$$\begin{aligned} f_2(z) &= -\frac{1}{2}f'(z)f(z) \\ &= -\frac{1}{2}\begin{pmatrix} \partial_q(p) & \partial_p(p) \\ \partial_q(-q) & \partial_p(-q) \end{pmatrix}\begin{pmatrix} p \\ -q \end{pmatrix} \\ &= -\frac{1}{2}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} p \\ -q \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} q \\ p \end{pmatrix} \end{aligned}$$

and we find that the modified equation is given by

$$\begin{cases} \dot{q} = p + \frac{h}{2}q + \mathcal{O}(h^2) \\ \dot{p} = -q + \frac{h}{2}p + \mathcal{O}(h^2) \end{cases}$$

Our hope is that the modified equation will reflect the initial problem's properties. Therefore we will consider some properties of the initial system (2.1) and study how they translate to the modified equation. We start with the order of the integrator.

**Theorem 2.4.7 (Hairer-Lubich-Wanner [12], p. 340, Theorem 1.2)**
Consider an $n$-th order integrator with numerical flow $\Phi_h$. By definition, this means that

$$\Phi_h(z) - \varphi_h(z) = h^{n+1}\delta_{n+1}(z) + \mathcal{O}(h^{n+2})$$

where $h^{n+1}\delta_{n+1}(z)$ denotes the leading term of the local truncation error. Then the modified equation (2.19) satisfies the Cauchy problem

$$\dot{z} = f(z) + h^n f_{n+1}(z) + h^{n+1} f_{p+1}(z) + \dots \text{ with } z(0) = y_0$$

where $f_{n+1}(z) = \delta_{n+1}(z)$.

*Proof.* In the beginning of this section, we constructed the coefficients $f_j(z)$ for $j \leq 3$ appearing in the modified equation. From this construction, it immediately follows that

$$\forall j \in \{2, 3, \dots, n\} : f_j(z) = 0 \iff \Phi_h(z) - \varphi_h(z) = \mathcal{O}(h^{n+1})$$

which implies the theorem. □

This result tells us precisely which powers of the step size appear in the modified equation and as a result, it tells us more about how close it is to the original equation. Moreover, the integrator $\Phi_h$ can be seen as the exact flow (with step size $h$) of the modified equation, up to order $\mathcal{O}(h^{n+1})$.
We now move on to another property. Intuitively, if we consider two systems $\mathcal{S}_1$ and $\mathcal{S}_2$ with initial positions and velocities $(x_0, v_0)$ resp. $(x_0, -v_0)$, then we would expect both trajectories to be identical, except for the direction of traversal. Systems that obey this intuitive law are called *reversible*, defined mathematically as follows.

**Definition 2.4.8**
We define the notions of reversibility for ODEs as well as maps.

1. An ODE $\dot{y} = f(y)$ is called **$\eta$-reversible** if $\eta \circ f = -f \circ \eta$ for some bijective linear transformation $\eta$ on the phase space of the differential equation.

2. A map $\Phi$ is called **$\eta$-reversible** if $\eta \circ \Phi = \Phi^{-1} \circ \eta$.

For the exact flow $\varphi_t$ of an $\eta$-reversible ODE $\dot{y} = f(y)$, it can be shown [12, p. 144] that $\eta \circ \varphi_t = \varphi_t^{-1} \circ \eta$. This implies that the exact flow of an $\eta$-reversible system is $\eta$-reversible.

**Theorem 2.4.9 ([12], p. 343, Theorem 2.3)**
Consider an $\eta$-reversible ODE $\dot{y} = f(y)$ and an integrator with $\eta$-reversible numerical flow $\Phi_h$. Then every truncation of the modified differential equation is

an $\eta$-reversible ODE.

*Proof.* We will prove this result by induction. Since in any case, the first coefficient $f_1(z) = f(z)$, the base case follows by the reversibility of $\dot{y} = f(y)$. By induction, we can assume that there exists some $k \in \mathbb{N}$ such that

$$\forall i \in \{1, \dots, k\} : \eta \circ f_i = -f_i \circ \eta$$

i.e. each $f_i$ is reversible for any index $1 \leq i \leq k$. Our goal is now to prove that this also holds for $f_{k+1}$. By the induction hypothesis, the modified equation

$$\dot{z} = f(z) + hf_2(z) + \ldots + h^{k-1} f_k(z) + \mathcal{O}(h^k)$$

is $\eta$-reversible, and as a result so is its flow $\varphi_h^k$, meaning $\eta \circ \varphi_h^k = (\varphi_h^k)^{-1} \circ \eta$. By theorem 2.4.7, it follows that

$$\Phi_h(z) = \varphi_h^k(z) + h^{k+1} f_{k+1}(z) + \mathcal{O}(h^{k+2})$$

and by the defining group property of flow maps, we have

$$(\Phi_h)^{-1}(z) = (\varphi_h^k)^{-1}(z) - h^{k+1} f_{k+1}(z) + \mathcal{O}(h^{k+2}).$$

Since by assumption both $\Phi_h$ and $\varphi_h^k$ are $\eta$-reversible maps, it follows that

$$\begin{aligned}(\eta \circ \Phi_h)(y) &= (\eta \circ \varphi_h^k)(y) + h^{k+1}(\eta \circ f_{k+1})(y) + \mathcal{O}(h^{k+2}) \\ &= ((\Phi_h)^{-1} \circ \eta)(y) \\ &= ((\varphi_h^k)^{-1} \circ \eta)(y) - h^{k+1}(f_{k+1} \circ \eta)(y) + \mathcal{O}(h^{k+2})\end{aligned}$$

which implies that $\eta \circ f_{k+1} = -f_{k+1} \circ \eta$ since $\varphi_h^k(y) = y + \mathcal{O}(h)$. □

Finally, we consider how the symmetry of an integrator in the sense of Definition 2.2.6 is reflected in its modified equation. It can be shown that only even powers of the step size $h$ appear in the modified differential equation.

**Theorem 2.4.10 ([12], Theorem 2.2, p. 342)**
Consider a symmetric method with numerical flow $\Phi_h$. For any $1 \leq j \in \mathbb{N}$, its corresponding coefficient $f_{2j}(z) = 0$.

*Proof.* Recall that in order for $\Phi_h$ to be symmetric, it must satisfy the identity $\Phi_h = \Phi_{-h}^{-1}$. This implies in particular that the modified equations of both these integrators have to be identical. If we consider the modified equation of $\Phi_h$ to be

$$\dot{z} = f_h(z) = f(z) + hf_2(z) + h^2 f_3(z) + h^3 f_4(z) + \ldots$$

then the modified equation of $\Phi_{-h}^{-1}$ has to satisfy

$$\dot{z} = f_{-h}(z) = f(z) - hf_2(z) + h^2 f_3(z) - h^3 f_4(z) + \ldots$$

and an equation $f_h(z)$ and $f_{-h}(z)$ results in

$$f_{2j}(z) = -f_{2j}(z) \iff f_{2j}(z) = 0$$

for any $1 \leq j \in \mathbb{N}$. □

Note that since Yoshida's symplectic integrators from the previous section are symmetric and reversible, these theorems can be used to examine their modified equations.

#### 2.4.3.2 Modified Hamiltonians

We will now shift our focus to modified equations of symplectic integrators applied to Hamiltonian systems. Our main objective is to prove that these modified equations are Hamiltonian themselves. If this is the case, we call them modified Hamiltonians for brevity. We will consider Hamiltonian systems induced by smooth Hamiltonian functions. These systems can be written in the form

$$\dot{y} = \boldsymbol{J} \cdot \nabla H(y) \tag{2.22}$$

where $\nabla$ denotes the gradient operator and $\boldsymbol{J}$ is the canonical matrix

$$\boldsymbol{J} = \begin{pmatrix} 0 & \boldsymbol{I} \\ -\boldsymbol{I} & 0 \end{pmatrix}.$$

We start by providing a *local* result. It makes use of the following lemma.

**Lemma 2.4.11 ([12], Lemma 2.7, p. 186)**
Consider an open subset $U \subset \mathbb{R}^n$ and let $f : U \to \mathbb{R}^n$ be a continuously differentiable function on $U$ with symmetrical Jacobian matrix $\mathrm{Jac}_f(x) = (\partial_{x_k} f_i(x))_{i,k=1}^n$ for any $x \in U$. Then for any $x_0 \in U$, there exists an open neighbourhood $V \subset U$ and a function $H : V \to \mathbb{R}$ of $x_0$ in $V$ such that $f(x) = \nabla H(x)$ on $V$.

*Proof.* Without loss of generality, we can take $x_0 = 0_n := (0, \ldots, 0) \in \mathbb{R}^n$. We consider $\mathcal{B}$ to be a ball around $x_0$ contained in $U$ and define a function $H : U \to \mathbb{R}$ by

$$x \mapsto H(x) := \int_0^1 x \cdot f(tx) dt + C$$

where $\cdot$ denotes the Euclidean scalar product and $C \in \mathbb{R}$ is some arbitrary constant. By the symmetry of $f$'s Jacobian matrix and by the fundamental theorem of calculus, it follows that

$$\frac{\partial H}{\partial x_i}(x) = \int_0^1 \left( f_i(tx) + tx \cdot \frac{\partial f}{\partial x_i}(tx) \right) dt = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}(f_i(tx)) dt = f_i(y)$$

since the integrand is continuously differentiable by assumption. It now follows that $f(x) = \nabla H(x)$, which proves the claim. □

We can now prove the following theorem.

**Theorem 2.4.12 ([12], Theorem 3.1, p. 343)**
Consider a Hamiltonian system (2.22) induced by some smooth Hamiltonian function $H : \mathbb{R}^{2n} \to \mathbb{R}$ and a symplectic integrator $\Phi_h(y)$. Then the modified differential equation is Hamiltonian in its own right, i.e. for $j \geq 2$ there exist smooth functions $H_j : \mathbb{R}^{2n} \to \mathbb{R}$ such that $f_j(y) = \boldsymbol{J} \cdot \nabla H_j(y)$.

*Proof.* We will use induction to prove this theorem. By assumption, it holds that $f_1(y) = f(y) = \boldsymbol{J} \cdot \nabla H(y)$ is Hamiltonian, proving the base case $r = 1$. Assume now

that there exists some $k > 1$ such that for all $i \in \{1, \ldots, k\} : f_i(y) = \boldsymbol{J} \cdot \nabla H_i(y)$. We need to show that there exists a Hamiltonian function $H_{k+1}$ such that $f_{k+1}(y) = \boldsymbol{J} \cdot \nabla H_{k+1}(y)$. To this aim, consider the truncated modified equation

$$\dot{z} = f(z) + hf_2(z) + \ldots + h^{k-1} f_k(z)$$

whose exact and numerical flows satisfy the identities

$$\begin{aligned} \Phi_h^k(y) &= \varphi_h^k(y) + h^{k+1} f_{k+1}(y) + \mathcal{O}(h^{k+2}) \\ \left(\Phi_h^k\right)'(y) &= (\varphi_h^k)'(y) + h^{k+1} f'_{k+1}(y) + \mathcal{O}(h^{r+2}) \\ (\varphi_h^k)'(y) &= \boldsymbol{I} + \mathcal{O}(h) \end{aligned}$$

where $\varphi_h^k$ and $\Phi_h^k$ are symplectic maps by assumption and $\boldsymbol{I}$ denotes the identity matrix. By definition of symplecticity, it then follows that

$$\begin{aligned} \boldsymbol{J} = (\Phi_h^k)'(y)^\intercal \boldsymbol{J} (\Phi_h^k)'(y) &= \boldsymbol{J} + h^{k+1} f'_{k+1}(y)^\intercal \boldsymbol{J} + h^{k+1} \boldsymbol{J} f'_{k+1}(y) + \mathcal{O}(h^{k+2}) \\ &= \boldsymbol{J} + h^{k+1} (f'_{k+1}(y)^\intercal + \boldsymbol{J} f'_{k+1}(y)) + \mathcal{O}(h^{k+2}) \end{aligned}$$

which implies

$$f'_{k+1}(y)^\intercal \boldsymbol{J} + \boldsymbol{J} f'_{k+1}(y) = 0.$$

Since $\boldsymbol{J}$ is antisymmetric, it follows from the identity above that the matrix $\boldsymbol{J} f'_{k+1}(y)$ is symmetric. Using these results, the existence of a Hamiltonian function $H_{k+1}$ such that $f_{k+1}(y) = \boldsymbol{J} \cdot \nabla H_{k+1}(y)$ follows from Lemma 2.4.11. □

Next, we will prove a global variant of this theorem. We first prove the following proposition, which is a result about so-called *generating functions* for Hamiltonian systems. Essentially, these are functions that capture information about the dynamics of Hamiltonian systems. Additionally, our numerical integrator $\Phi_h : (q, p) \mapsto (Q, P)$ can locally be written in terms of such a generating function $S(Q, p, h)$ as follows [12, p. 197, Lemma 5.3].

$$\begin{cases} q = Q + \frac{\partial S}{\partial p}(Q, p, h) \\ P = p + \frac{\partial S}{\partial Q}(Q, p, h) \end{cases}$$

where $S$ can be seen as a solution of the following Cauchy problem

$$\frac{\partial S}{\partial t}(Q, p, t) = H\left(Q, p + \frac{\partial S}{\partial S}(Q, p, t)\right) \text{ with } S(Q, p, 0) = 0 \tag{2.23}$$

by [12, p. 201, Lemma 5.7]. In this thesis we are not expanding on the theory of generating functions. We will prove the next proposition and then move on to the result about global modified Hamiltonians.

**Proposition 2.4.13 (Hairer-Lubich-Wanner [12], p. 196, Theorem 5.1)**
A map $\varphi : (q, p) \mapsto (Q, P)$ is symplectic if and only if there locally exists a function $S(q, p)$ such that

$$\sum_i (Q_i dP_i - q_i dp_i) = dS. \tag{2.24}$$

*Proof.* We write the Jacobian of $\varphi$ as a block matrix

$$\mathbf{Jac}(\varphi) = \begin{pmatrix} \partial_q Q & \partial_p Q \\ \partial_q P & \partial_p P \end{pmatrix}.$$

$\varphi$'s symplecticity is equivalent to the symplecticity of its Jacobian matrix [12, p. 183, Definition 2.2], meaning

$$\mathbf{Jac}(\varphi)^\intercal \boldsymbol{J} \mathbf{Jac}(\varphi) = \boldsymbol{J}$$

which implies the following (equivalent) conditions for $\varphi$'s symplecticity.

$$\begin{aligned} \partial_p Q \cdot \partial_q P &= \partial_q P \cdot \partial_p Q \\ \partial_p Q \cdot \partial_p P &= \partial_p P \cdot \partial_p Q \\ \partial_p Q \cdot \partial_p P &= I + \partial_q P \cdot \partial_p Q \\ \partial_p P \cdot \partial_q Q &= I + \partial_p Q \cdot \partial_q P \end{aligned} \tag{2.25}$$

If we plug the total derivative $dP = P_q dq + P_p dp$ into (2.24), we find that

$$dS = \begin{pmatrix} \partial_q P^\intercal Q \\ \partial_p P^\intercal Q \end{pmatrix}^\intercal \begin{pmatrix} dq \\ dp \end{pmatrix} =: C \begin{pmatrix} dq \\ dp \end{pmatrix}.$$

It can now be shown that the Jacobian matrix of $C$ is symmetric as a result of the symplecticity conditions (2.25). The statement then follows from Lemma 2.4.11 and by rewriting the formula using summations. □

The global result below tells us that the modified Hamiltonian is in fact defined on the same domain as the generating function $S$ of the integrator, which is what grants it the status of being a *global* result.

**Theorem 2.4.14 (Hairer-Lubich-Wanner [12], p. 345, Theorem 3.2)**
Consider an integrator with numerical flow $\Phi_h$ and generating function

$$S(Q, p, h) = hS_1(Q, p) + h^2 S_2(Q, p) + h^3 S_3(Q, p) + \dots \tag{2.26}$$

such that each $S_j$ is a smooth function of $Q$ and $p$ defined on some open set $U$. Then the resulting modified differential equation is Hamiltonian, and it is of the form

$$\tilde{H}(q, p) = H(q, p) + hH_2(q, p) + h^2 H_3(q, p) + \dots \tag{2.27}$$

where the functions $H_j(q, p)$ are smooth and defined on all of $U$.

*Proof.* The Hamiltonian system corresponding to $\tilde{H}$ has the exact solution $(Q, P) = (\tilde{q}(t), \tilde{p}(t))$ such that

$$\begin{cases} q = Q + \frac{\partial \tilde{S}}{\partial p}(Q, p, t) \\ P = p + \frac{\partial \tilde{S}}{\partial Q}(Q, p, t) \end{cases}$$

where the generating function $\tilde{S}$ solves the Cauchy problem (2.23). Our goal is to find the functions $H_j(q, p)$ such that $\tilde{S}$ coincides with $S$ when $t = h$. First, we will expand $\tilde{S}$ as a series

$$\tilde{S}(Q, p, t) = t\tilde{S}_1(Q, p, h) + t^2 \tilde{S}_2(Q, p, h) + t^3 \tilde{S}_3(Q, p, h) + \dots$$

and plug it into (2.23). Comparing coefficients of like powers of $t$, we find that

$$\begin{aligned}
\tilde{S}_1(q,p,h) &= \tilde{H}(q,p) \\
\tilde{S}_2(q,p,h) &= \frac{1}{2}\left(\frac{\partial \tilde{H}}{\partial q}\frac{\partial \tilde{S}_1}{\partial P}\right)(q,p,h) \\
\tilde{S}_3(q,p,h) &= \frac{1}{3}\left(\frac{\partial \tilde{H}}{\partial q}\frac{\partial \tilde{S}_2}{\partial P}\right)(q,p,h) + \frac{1}{6}\left(\frac{\partial^2 \tilde{H}}{\partial q^2}\left(\frac{\partial \tilde{S}_1}{\partial P}, \frac{\partial \tilde{S}_1}{\partial P}\right)\right)(q,p,h) \\
&\vdots
\end{aligned} \tag{2.28}$$

and so on for the other coefficients. We can now again write each $\tilde{S}_j$ as a series

$$\tilde{S}_j(q,p,h) = \tilde{S}_j^1(q,p) + h\tilde{S}_j^2(q,p) + h^2\tilde{S}_j^3(q,p) + \dots$$

and we plug it, along with the expression (2.27), into (2.28). By comparing like powers of the step size $h$, we find

$$\tilde{S}_j^1(q,p) = H_j(q,p)$$

and for indices $k > 1$ it follows that $\tilde{S}_j^k(q,p)$ is a function dependent on derivatives of $H_l$ where $l < j$. Since we want $\tilde{S}(q,p,h) = S(q,p,h)$ it follows that

$$\begin{aligned}
S_1(q,p) &= \tilde{S}_1^1(q,p) \\
S_2(q,p) &= \tilde{S}_2^1(q,p) + \tilde{S}_1^2(q,p) \\
&\vdots
\end{aligned}$$

which shows that $S_j(q,p)$ is the sum of the coefficient $H_j(q,p)$ and some function dependent of derivatives of $H_s(q,p)$ with $s < j$. Given a generating function $S(Q,p,h)$, we can construct the functions $H_j(q,p)$ from these relations which are defined on the same domain as the coefficients $S_j$. This proves the claim. □

#### 2.4.3.3 Modified equations of splitting methods

Suppose we consider a Hamiltonian system induced by a separable Hamiltonian function $H(q,p) = T(p) + V(q)$. If we denote by $\Phi_h$ the numerical flow of the symplectic Euler (VT) integrator, then it follows from the BCH-formula that

$$\Phi_h = \exp\left(h(X_T + X_V) + \frac{h^2}{2}[X_T, X_V] + \frac{h^3}{12}\left([X_T,[X_T,\mathcal{L}_V]] + [X_V,[X_V,X_T]]\right) + \dots\right) =: e^{X_{\tilde{H}}}$$

which implies that the symplectic Euler (VT) integrator is the formal solution to the modified equation $\dot{z} = \left(\tilde{X}^H \circ \mathrm{Id}\right)(z)$, where $\tilde{X}_H = X_{\tilde{H}}$. Note that by Theorem 2.4.14 this notation as a Hamiltonian vector field is justified. Using the identity $[X^f, X^g] = X^{-\{f,g\}}$ for Hamiltonian vector fields, we find that

$$\Phi_h = \exp\left(hH + \frac{h^2}{2}\{V,T\} + \frac{h^3}{12}(\{V,\{T,T\}\} + \{T,\{V,V\}\}) + \dots\right)$$

which means that we have proven the following result.

**Proposition 2.4.15 (Yoshida [35], p. 36, eq. 50)**
Consider a smooth, separable Hamiltonian $H(q,p) = T(p) + V(q)$ and let $\Phi_h$ be the numerical flow of the symplectic Euler integrator using VT-splitting. Then this integrator is the exact solution of the Hamiltonian system induced by the modified Hamiltonian

$$\tilde{H}(q,p) = H + hH_2(q,p) + h^2 H_3(q,p) + \dots$$

where the coefficients $H_j(q,p)$ can be found through repeated Poisson brackets as follows.

$$\begin{aligned} H_2 &= \{V,T\} \\ H_3 &= \{V,\{T,T\}\} + \{T,\{V,V\}\} \\ &\vdots \end{aligned}$$

Switching the roles of $T$ and $V$, this result can also be proven for symplectic Euler using TV-splitting. Analogously, a similar result holds for Stormer-Verlet using VTV (or TVT) splitting. Since it is a symmetric integrator, only even powers of $h$ appear in the modified Hamiltonian. For example [35, p. 36, eq. (52)], the coefficient $H_2$ is given by

$$H_2(q,p) = \frac{1}{24}\left(\{V,\{T,V\}\} - 2\{T,\{V,T\}\}\right).$$

Finally, we present the following result [35, p. 36, eq. (53)].

**Corollary 2.4.16**
If $\Phi_h$ denotes the flow of an $n$-th order integrator applied to a separable Hamiltonian system, then its modified Hamiltonian satisfies the identity

$$\tilde{H} = H + h^n H_n + \mathcal{O}(h^{n+1}).$$

*Proof.* This follows by applying Theorem 2.4.7 to a Hamiltonian system. □

#### 2.4.3.4 The long-term conservation of energy

In [12, sections IX.7.1-IX.7.3] the local errors of symplectic integrators and the convergence of modified equations are studied by making analyticity assumptions [12, p. 360] on the differential equation $\dot{y} = f(y) = \boldsymbol{J}\nabla H(y)$ and on the numerical method. It is proven (see [12, Theorem 7.5]) that there exists an upper bound

$$\|f_j(z)\| \leq \ln(2)\eta M \left(\frac{\eta M j}{R}\right)^{j-1} \tag{2.29}$$

for the modified equation's coefficients $f_j = \boldsymbol{J}\nabla H$ in the complex ball $\|z - y_0\| \leq \frac{R}{2}$. Moreover, it is proven (see [12, Theorem 7.6]) that there exists an optimal choice of truncation index $N = N(h)$ such that

$$\|\Phi_h(y_0) - \tilde{\varphi}_h^N(y_0)\| \leq h\gamma M e^{-\frac{h_0}{h}} \tag{2.30}$$

provides an upper bound for the difference between the numerical method $\Phi_h$ and the exact flow $\tilde{\varphi}_h^N$ of the truncated modified equation.

Recall that by Theorems 2.4.12 and 2.4.14 the truncated modified equation of a Hamiltonian system $\dot{y} = \boldsymbol{J}\nabla H(y)$ is given by

$$\tilde{H}(z) = H(z) + h^r H_{r+1}(z) + \ldots + h^{N-1} H_N(z) \tag{2.31}$$

which is Hamiltonian in its own right.

**Theorem 2.4.17**
Consider a symplectic integrator with numerical flow $\Phi_h$ and step size $h > 0$, and a Hamiltonian system induced by a Hamiltonian function $H : U \to \mathbb{R}$ where $U \subset \mathbb{R}^{2n}$ is an open subset. Suppose the analyticity assumptions [12, p. 360] are satisfied. If the numerical solution obtained through $\Phi_h$ stays at all times in a compact set $K \subset U$, then there exist $h_0$ and $N = N(h)$ such as in [12, p. 365, theorem 7.6] such that over exponentially long time intervals $nh \leq e^{\frac{h_0}{2h}}$, we have

$$\tilde{H}(y_n) = \tilde{H}(y_0) + \mathcal{O}\left(e^{-\frac{h_0}{2h}}\right) \text{ and } H(y_n) = H(y_0) + \mathcal{O}(h^r)$$

where $r$ denotes the order of the integrator.

*Proof.* We denote by $\tilde{\varphi}_h^N$ the flow of (2.31). Note that for this modified Hamiltonian the identity $\tilde{H}(\tilde{\varphi}_\tau^N(y_n)) = \tilde{H}(y_n)$ must hold at any time $\tau$, since energy is conserved along Hamiltonian trajectories. By (2.29) and the analyticity assumptions, it follows that there exists some global, $h$-independent Lipschitz condition on $\tilde{H}$

$$\|\tilde{H}(y_{n+1}) - \tilde{H}(\tilde{\varphi}_h^N(y_n))\| \leq L\|y_{n+1} - \tilde{\varphi}_h^N(y_n)\|$$

and from (2.30) it follows that

$$L\|y_{n+1} - \tilde{\varphi}_h^N(y_n)\| \leq hL\gamma M e^{-\frac{h_0}{h}} = Che^{-\frac{h_0}{h}}$$

where $C = L\gamma M$. This implies that

$$\tilde{H}(y_{n+1}) - \tilde{H}(\tilde{\varphi}_h^N(y_n)) = \mathcal{O}\left(he^{-\frac{h_0}{h}}\right).$$

By rewriting the difference $\tilde{H}(y_n) - \tilde{H}(y_0)$ as a telescopic sum and using the identity $\tilde{H}(\tilde{\varphi}_\tau^N(y_n)) = \tilde{H}(y_n)$, we find

$$\tilde{H}(y_n) - \tilde{H}(y_0) = \sum_{i=1}^{n} \left(\tilde{H}(y_i) - \tilde{H}(y_{i-1})\right) = \sum_{i=1}^{n} \left(\tilde{H}(y_i) - \tilde{H}(\tilde{\varphi}_\tau^N(y_{i-1}))\right)$$

from which it follows that

$$\tilde{H}(y_n) - \tilde{H}(y_0) = \sum_{i=1}^{n} \mathcal{O}\left(he^{-\frac{h_0}{h}}\right) = n\mathcal{O}\left(he^{-\frac{h_0}{h}}\right) = \mathcal{O}\left(nhe^{-\frac{h_0}{h}}\right)$$

and since we assumed $nh \leq e^{\frac{h_0}{2h}}$ we have $\mathcal{O}\left(nhe^{-\frac{h_0}{h}}\right) = \mathcal{O}\left(e^{-\frac{h_0}{2h}}\right)$, implying

$$\tilde{H}(y_n) = \tilde{H}(y_0) + \mathcal{O}\left(e^{-\frac{h_0}{2h}}\right).$$

For the second statement about $H$, see [12, p. 367]. □

Symplectic integrators of order $r$ conserve the modified Hamiltonian *almost exactly* over exponentially long time intervals, with an exponentially small error. They do not conserve the Hamiltonian, but they make sure that the error from its initial value is of order $r$.
In section 2.4.1, we discussed how the explicit Euler integrator causes an *indefinite increase in energy* when applied to the uncoupled harmonic oscillator. If we denote by $y_n = (q_n, p_n)$ forward Euler's approximations and if we repeat the computations in the proof above, it can be shown [12, p. 368] that

$$H(y_n) = H(y_0) + \mathcal{O}(th) \tag{2.32}$$

where $t = nh$. This means that the error in energy grows linearly.

#### 2.4.3.5 Other conserved quantities

Some Hamiltonian systems possess additional conserved quantities besides the total energy $H$ (in the autonomous case).

**Definition 2.4.18 (McLachlan-Quispel, [22], p. 16)**
A function $I(x(t))$ is called a **first integral** or **conserved quantity** of a Hamiltonian system if

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sum_{i=1} \frac{\partial I}{\partial x_i} \frac{\mathrm{d}x^i}{\mathrm{d}t} = 0$$

where $x(t)$ denotes a solution to the corresponding Hamiltonian equations.

As an example, we consider the Kepler problem. It is a Hamiltonian system induced by the autonomous and separable Hamiltonian function

$$H(q_1, q_2, p_1, p_2) = \frac{p_1^2 + p_2^2}{2m} - \frac{GmM}{\sqrt{q_1^2 + q_2^2}}$$

and it has angular momentum $L(q_1, q_2, p_1, p_2) = q_1 p_2 - q_2 p_1$ as a first integral. To see this, we use the Hamiltonian equations induced by $H$ and compute

$$\begin{aligned}
\frac{\mathrm{d}L}{\mathrm{d}t}(q_1, q_2, p_1, p_2) &= \dot{q}_1 p_2 + q_1 \dot{p}_2 - \dot{q}_2 p_1 - q_2 \dot{p}_1 \\
&= p_1 p_2 - q_1 \frac{q_2}{\sqrt{q_1^2 + q_2^2}^3} - p_1 p_2 + q_2 \frac{q_1}{\sqrt{q_1^2 + q_2^2}^3} \\
&= 0.
\end{aligned}$$

Conserved quantities and their behaviour with symplectic integrators are studied in further detail in [12, chapter IV] and [22]. The main takeaway is that symplectic integrators can be constructed to preserve certain conserved quantities. As an example, consider the following proposition.

**Proposition 2.4.19 (Blanes et al. [3], p. 67)**
Let $\Phi_h$ be a symplectic integrator of order $r$ and apply it to a given Hamiltonian system with action-angle coordinates $(q, p)$. Suppose $I(q, p)$ is a first integral

that solely depends on $(q, p)$, then $I$ satisfies

$$I(q_n, p_n) = I(q_0, p_0) + \mathcal{O}(h^r)$$

where $q_n, p_n$ are $\Phi_h$'s approximations.

In other words, the first integral is conserved exactly up to the order of $\mathcal{O}(h^r)$.

## 2.5 Numerical examples

To illustrate the behaviour and qualitative superiority of symplectic integrators, we will unleash several integration schemes on three dynamical systems: the uncoupled harmonic oscillator, the mathematical pendulum, the Kepler problem, and the three body problem. We consider the forward (explicit) Euler integrator as an example of a classical integrator. For the symplectic integrators, we consider symplectic Euler (VT), Stormer-Verlet (VTV), and Yoshida's 4th order integrator [34]. For further details on the classical (Runge-Kutta) integrators, see Appendix A and for more information about the MATLAB code that was used, see Appendix B.

### 2.5.1 The harmonic oscillator

We consider the two-dimensional harmonic oscillator, which describes the movement of a mass hooked up to a spring. We consider an object of unit mass $m = 1$ [kg] and a spring with unit spring constant $k = 1$ [N/m] (Newton per meter). Its separable Hamiltonian is given by the function $H : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$H(q, p) := \frac{q^2}{2} + \frac{p^2}{2} = V(q) + T(p) \tag{2.33}$$

where the variables $q$ and $p$ express respectively the distance from the object to its point of equilibrium and its velocity. This results in the Hamiltonian system

$$\begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \partial_p H \\ -\partial_q H \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} p \\ -q \end{pmatrix} \tag{2.34}$$

with initial values $(q_0, p_0) = (q(0), p(0)) \in \mathbb{R}^2$, where $\partial_q$ and $\partial_p$ respectively denote the partial derivative operators with respect to $q$ and $p$. A routine computation verifies that the exact solution to (2.34) is given by

$$\begin{cases} q(t) = q_0 \cos(t) + p_0 \sin(t) \\ p(t) = p_0 \cos(t) - q_0 \sin(t) \end{cases} \tag{2.35}$$

which defines a circle with radius $\sqrt{q_0^2 + p_0^2}$ for initial values $(q_0, p_0) \in \mathbb{R}^2$, i.e. the level sets of (2.33) are circles. We now consider the initial values $(q_0, p_0) = (2, 1.5)$, resulting in the energy

$$H_0 := H(q_0, p_0) = \frac{2^2}{2} + \frac{1.5^2}{2} = 3.125$$

which remains constant along the system's exact solution. In our numerical integrations, we will use the step size $h = 0.01$ and we set the maximum amount of iterations to be calculated to $N = 2000$.

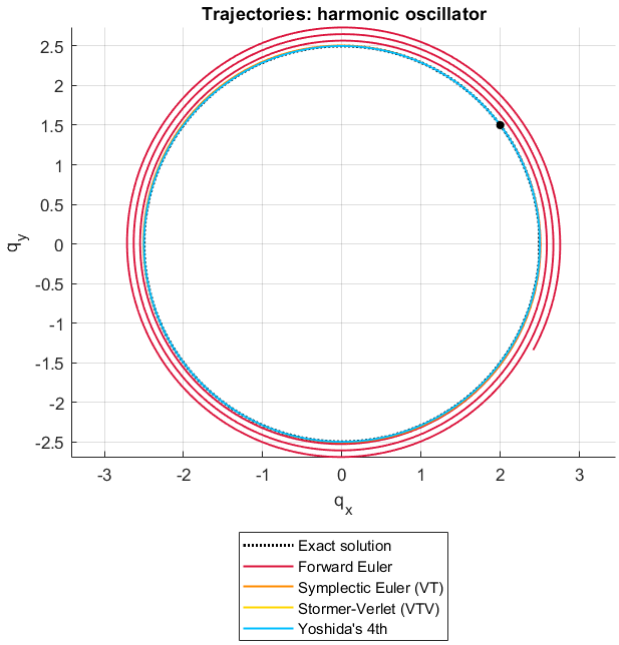


Figure 2.1: The trajectories of the numerical schemes (colored), along with the exact solution (dashed circle). The step size $h = 0.01$ was used in the numerical schemes and a total amount of $N = 2000$ iterations were calculated.

Upon taking a closer look at the calculated trajectories in Figure 2.1, we can see that it does not take long for the classical approximations to start diverging from the exact solution. Contrarily, the symplectic integrators' approximations do not stray far from the exact solution. The latter is the more ideal behaviour of numerical integrators applied to Hamiltonian systems - we know that the level sets of $H$ are circles, so the closer a scheme's approximations are to lying on a circle, the qualitatively better the scheme is for this specific problem. Now, let us take a look at the energy associated to each scheme.

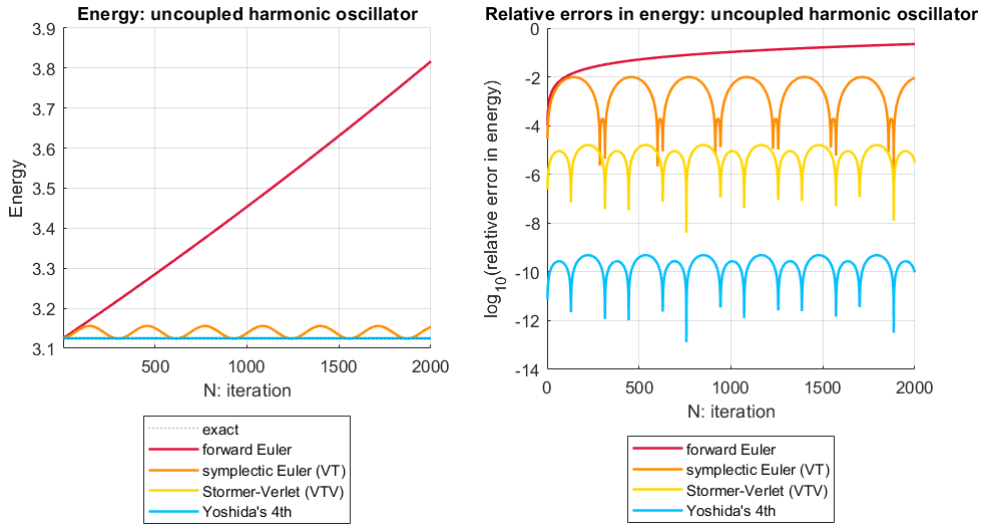


Figure 2.2: The energy in all iterations of the respective numerical schemes (left); the relative errors in energy of each numerical scheme on a logarithmic scale (right). In both cases, the step size $h = 0.01$ was used. A total amount of $N = 2000$ iterations were calculated.

Since the harmonic oscillator is a Hamiltonian system, the total energy i.e. the Hamiltonian is a conserved quantity. In other words, the exact total energy $H_0 = H(q_0, p_0) = 3.125$ remains constant through time. If we consider one iteration of the forward Euler integrator, we obtain $q_1 = 1.97$ and $p_1 = 1.54$, which results in the total energy

$$H(q_1, p_1) = \frac{q_1^2}{2} + \frac{p_1^2}{2} = 3.13$$

associated to this iteration of the numerical scheme. After just one iteration of the explicit Euler scheme, the obtained energy has increased by 0.05. For the symplectic Euler and Stormer-Verlet schemes, the associated energies stay much closer to $H_0$ (3.1252 resp. 3.1250) after one integration step. Figure 2.2 (left) repeats this procedure in each iteration and plots the resulting energy. The graph shows us that forward Euler does not preserve the system's energy at all; instead, it leads to a great increase, clearly violating the energy conservation property of Hamiltonian systems. This affirms our earlier finding (2.32): the error grows linearly. Contrarily, the symplectic integrators are better at preserving the exact energy $H_0$. The energy resulting from Yoshida's 6th order integrator stays even closer to $H_0$ than it does for the other two symplectic schemes. Figure 2.2 (right) shows us the relative errors in energy of the numerical schemes on a logarithmic scale. It affirms what the previous graphs already told us. The symplectic integrators are best at preserving energy out of our choice of numerical methods, and Yoshida's integrator approximates the solution best.

### 2.5.2 The mathematical pendulum

We now consider the mathematical pendulum, with Hamiltonian function

$$H(q, p) = \frac{p^2}{2} + 1 - \cos(q)$$

with $q, p \in \mathbb{R}$, inducing the Hamiltonian equations

$$\begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} p \\ -\sin(q) \end{pmatrix}$$

and we consider the initial values $(q_0, p_0) = \left(\frac{3}{2}, 1\right)$. This is an example of a separable, non-integrable Hamiltonian system: it has no exact solution. This is where numerical methods demonstrate their utility.

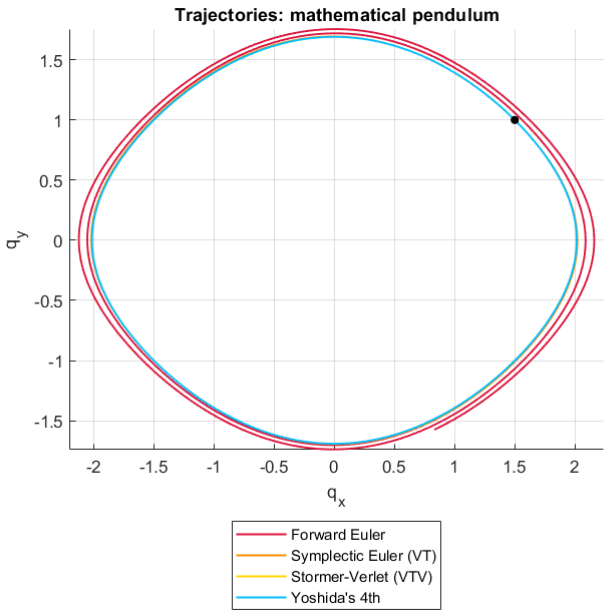


Figure 2.3: The trajectories of the numerical schemes (colored), along with the exact solution (dashed circle). The step size $h = 0.01$ was used in the numerical schemes and a total amount of $N = 2000$ iterations were calculated.

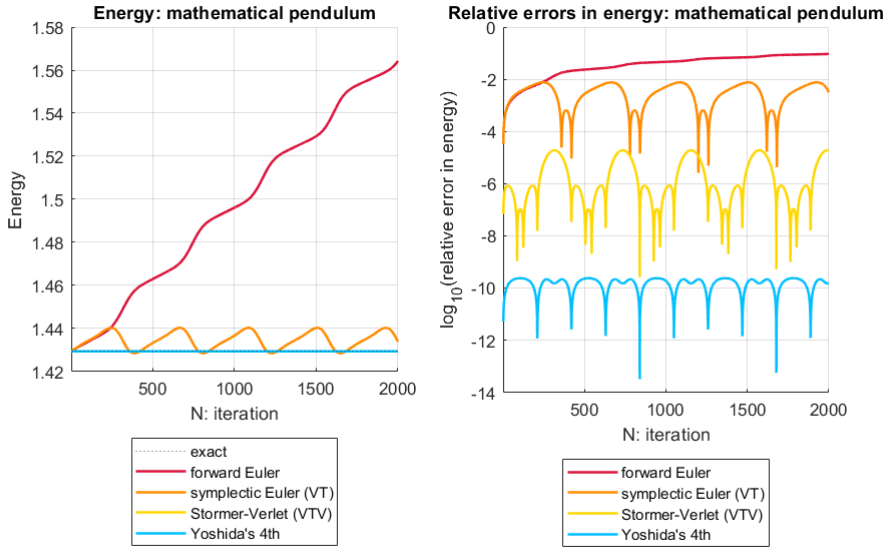


Figure 2.4: The energy in all iterations of the respective numerical schemes (left); the relative errors in energy of each numerical scheme on a logarithmic scale (right). In both cases, the step size $h = 0.01$ was used. A total amount of $N = 2000$ iterations were calculated.

FWe can draw the same conclusions as before from the figures above. The symplectic integrators clearly perform way better at preserving the system's total energy, while forward Euler does not.

### 2.5.3 The $n$-body problem

In this section, we will consider the $n$-body problem for $n \in \{2, 3\}$. It describes the motion of $n$ celestial bodies that are gravitationally attracted to eachother. In general, the Hamiltonian function is given by

$$H(q,p) := \sum_{i=1}^{n} \frac{\|p_i\|_2^2}{2m_i} - \sum_{1 \le i < j \le n} \frac{G m_i m_j}{\|q_i - q_j\|_2} \tag{2.36}$$

where $q = (q_1, \ldots, q_n)$ and $p = (p_1, \ldots, p_n) \in (\mathbb{R}^3)^n$ are supervectors of the $n$ bodies' positions resp. momenta, $m_i$ are the masses of the $n$ bodies, $G$ is the gravitational constant, and $\|\cdot\|_2$ denotes the Euclidean norm. The Hamiltonian (2.36) leads to the equations of motion

$$\begin{pmatrix} \dot{q}_i \\ \dot{p}_i \end{pmatrix} = \begin{pmatrix} \partial_{p_i} H \\ -\partial_{q_i} H \end{pmatrix} = \begin{pmatrix} \dfrac{p_i}{m_i} \\ -\displaystyle\sum_{j \neq i} \dfrac{G m_i m_j}{\|q_i - q_j\|_2^3} (q_i - q_j) \end{pmatrix} \tag{2.37}$$

for $1 \le i \le n$, with initial values $(q_0, p_0) = (q(0), p(0)) \in (\mathbb{R}^3)^n \times (\mathbb{R}^3)^n$. For $n = 2$, we will consider the Kepler-problem and for $n = 3$ we will consider a specific instance of the three body problem.

#### 2.5.3.1 Kepler problem (n=2)

For the Kepler problem, the Hamiltonian (2.36) can be written as

$$H(q_1, q_2, p_1, p_2) = \frac{\|P\|_2^2}{2} - \frac{1}{\|Q\|_2}$$

after proving that the Kepler problem's dynamics are planar, and by choosing a basis $Q \in \mathbb{R}^2$ of this plane where $Q = (q_1, q_2)$ and $P = (p_1, p_2)$, as per [12]. The equations of motion (2.37) can in turn be written as

$$\begin{pmatrix} \dot{Q} \\ \dot{P} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\|Q\|_2^{-3} & 0 \end{pmatrix} \begin{pmatrix} Q \\ P \end{pmatrix}.$$

Both the Hamiltonian $H$ and the angular momentum $L$ are conserved quantities for this system, where $L$ is defined as

$$L(Q, P) = q_1 p_2 - q_2 p_1.;$$

If we switch to polar coordinates

$$(q_1, q_2) = (r \cos\alpha, r \sin\alpha)$$

we can rewrite the energy and angular momentum associated to a given solution as

$$H_0 = \frac{L_0^2}{2r^4} \left( \left( \frac{\partial r}{\partial \alpha} \right)^2 + r^2 \right) - \frac{1}{r}$$
$$L_0 = r^2 \dot{\alpha}$$

where $\dot{r} = \partial_\alpha r \cdot \dot{\alpha}^2$ by the chain rule. By isolating $\partial_\alpha r$ in the identity for $H_0$ and by separation of variables, it follows that

$$r(\alpha) = \frac{L_0^2}{1 + e\cos(\alpha - \alpha_0)}$$

where $e := \sqrt{1 + 2H_0 L_0^2}$ is called the *eccentricity* of the orbit and where $\alpha_0$ is determined by the initial conditions for $Q$ and $P$, which in turn determine the initial conditions $r_0$ and $\alpha_0$. Depending on the values of $e$, $r(\alpha)$ results in a circle ($e = 0$), an ellipse ($0 < e < 1$), a parabola ($e = 1$) or a hyperbola ($e > 1$). Using the formula for $e$, we can now plot the exact solution to the Kepler problem given a set of initial conditions along with some approximations. We will use step size $h = 0.01$, compute a maximum amount of $N = 10000$ iterations per integrator and we will use the initial values

$$q(0) = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} \text{ and } p(0) = \begin{pmatrix} 0 \\ \frac{\sqrt{6}}{2} \end{pmatrix}$$

as proposed in [12, pg. 5].

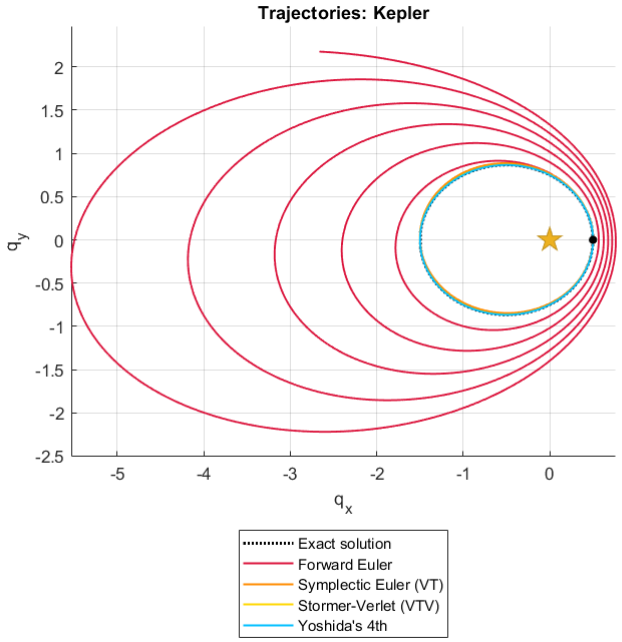


Figure 2.5: The trajectories of the numerical schemes (colored), along with the exact solution (dashed ellipse). The step size $h = 0.01$ was used in the numerical schemes and a total amount of $N = 10000$ iterations were calculated.

This figure immediately demonstrates the qualitative superiority of the symplectic integrators. The forward Euler approximations do not take long to diverge from the exact solution, whereas the symplectic integrators keep a reasonable distance from the exact solution. Moreover, the forward Euler approximations do not visually represent an ellipse, whereas the symplectic approximations do.

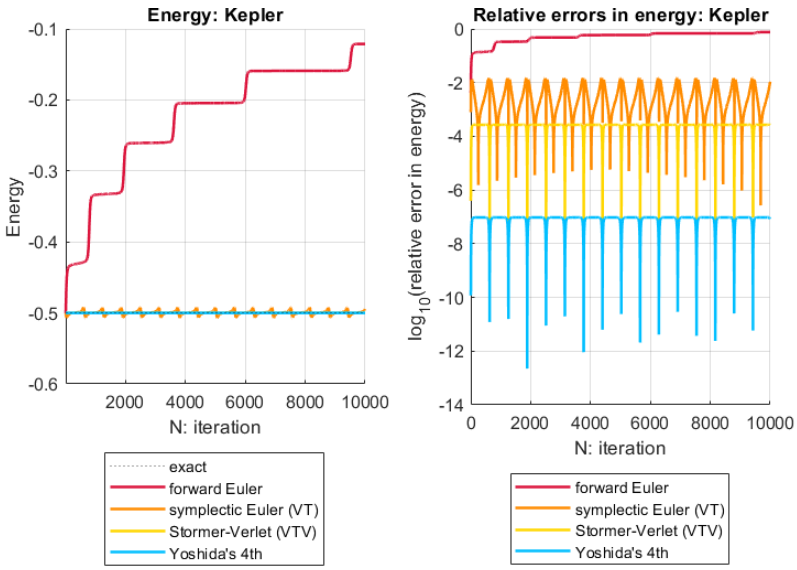


Figure 2.6: The energy in all iterations of the respective numerical schemes (left); the relative errors in energy of each numerical scheme on a logarithmic scale (right). In both cases, the step size $h = 0.01$ was used. A total amount of $N = 10000$ iterations were calculated.

This figure above illustrates the energy corresponding to each iteration of the integrators. We can draw the same conclusion as for the oscillator from these graphs - the symplectic integrators perform way better at conserving the system's energy than the classical scheme does. The energy spikes at each start of a new orbit due to the kinetic energy being largest when the celestial body is closest to the star. Moreover, the relative errors remain bounded even when we compute a large amount of iterations, e.g. $N = 8000$.

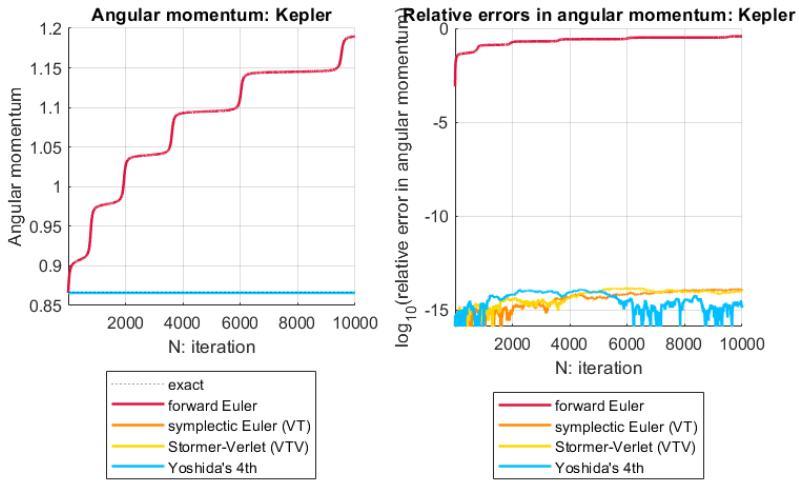


Figure 2.7: The angular momentum in all iterations of the respective numerical schemes (left); the relative errors in angular momentum of each numerical scheme on a logarithmic scale (right). In both cases, the step size $h = 0.01$ was used and a total amount of $N = 10000$ iterations were calculated.

The figure above illustrates the angular momenta corresponding to each iteration of the integrators and leads to the same conclusions we previously made. However, in

this case the symplectic integrators stay even closer the the initial (exact) angular momentum $L_0$ whereas the forward Euler scheme does not conserve it at all.

#### 2.5.3.2 Three body problem (n=3)

For the three body problem, the Hamiltonian (2.36) becomes

$$H(q,p) = \frac{\|p_1\|_2^2}{2m_1} + \frac{\|p_2\|_2^2}{2m_2} + \frac{\|p_3\|_2^2}{2m_3} - \frac{Gm_1m_2}{\|q_1-q_2\|_2} - \frac{Gm_1m_3}{\|q_1-q_3\|_2} - \frac{Gm_2m_3}{\|q_2-q_3\|_2}.$$

The non-integrable character of the three body problem is what sets it apart from the previous numerical simulations: it has no exact solution. Therefore, the best one can do is to approximate its solution by using numerical integrators. It is important to mention that the resulting system of ODEs (2.37) is chaotic for many different initial value conditions $(q(0), p(0)) = (q_0, p_0)$. This means that it is hard to study the long-term dynamics, since they depend sensitively on the initial value conditions. However, this problem can be more or less circumvented by imposing further restrictions on the general three body problem to arrive at *nice* solutions. In our simulation, we will consider a set of initial values that have been known to produce satisfactory trajectories. More precisely, we consider the initial values

$$q_{1,0} = \begin{pmatrix} 0.97000436 \\ -0.24308753 \\ 0 \end{pmatrix} \quad q_{2,0} = -q_{1,0} \quad q_{3,0} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$
$$p_{1,0} = \begin{pmatrix} 0.466203685 \\ 0.43236573 \\ 0 \end{pmatrix} \quad p_{2,0} = p_{1,0} \quad p_{3,0} = \begin{pmatrix} -0.93240737 \\ -0.86473146 \\ 0 \end{pmatrix}$$

as proposed in [7, pg. 1] and we set all three masses and the gravitational constant $G$ equal to 1. Moreover, we will use the step size $h = 0.01$ and compute a maximal amount of $N = 20000$ approximations for each integrator. In the aforementioned paper, Chenciner and Montgomery prove that the resulting orbit of the three bodies is a stable collision-free figure 8 with angular velocity equal to 0. This gives us an approach to geometrically validate the numerically computed trajectories.

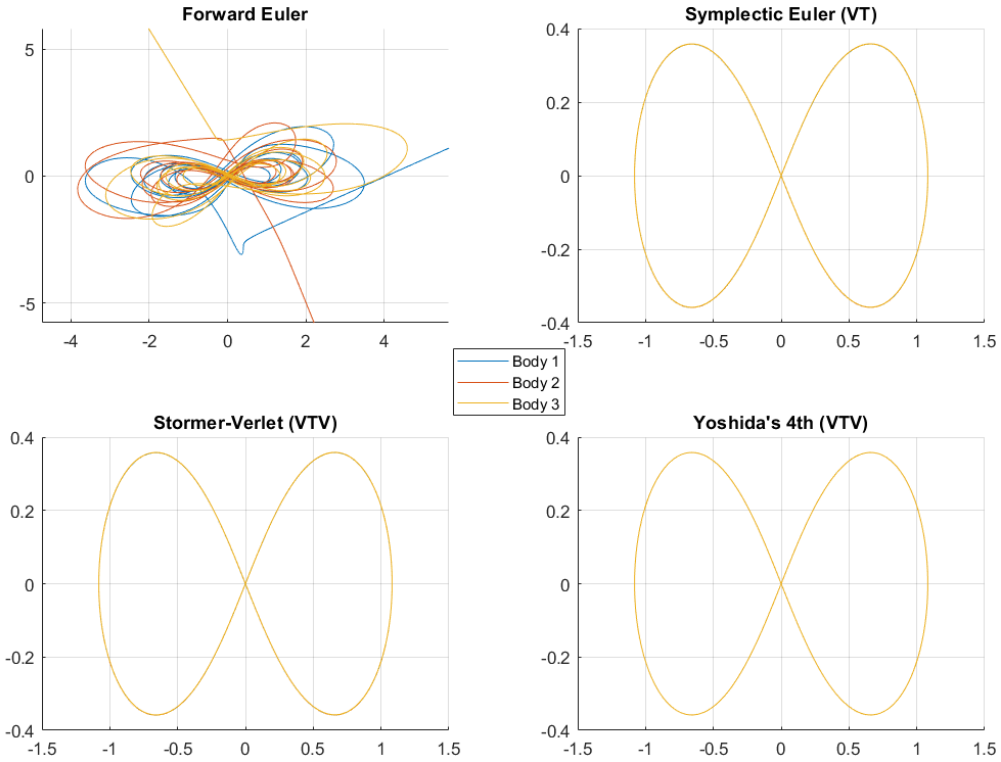


Figure 2.8: The trajectories of the numerical schemes (colored): $q_x$ coordinates on the horizontal axes, $q_y$ coordinates on the vertical axes. The step size $h = 0.01$ was used in the numerical schemes and a total amount of $N = 20000$ iterations were calculated. See [7, pg. 1] for more information on our choice of initial values.

It is immediately clear from Figure 2.8 that the forward Euler integrator does not respect the figure 8 trajectory. In fact, all three bodies completely deviate from the expected orbit after a long time, i.e. after about 20000 iterations in this simulation. The other panels reaffirm the qualitative superiority of the symplectic integrators, in the sense that the trajectories do not deviate from the figure 8 trajectory, even after a long time. We now move on to the energy associated to the system.

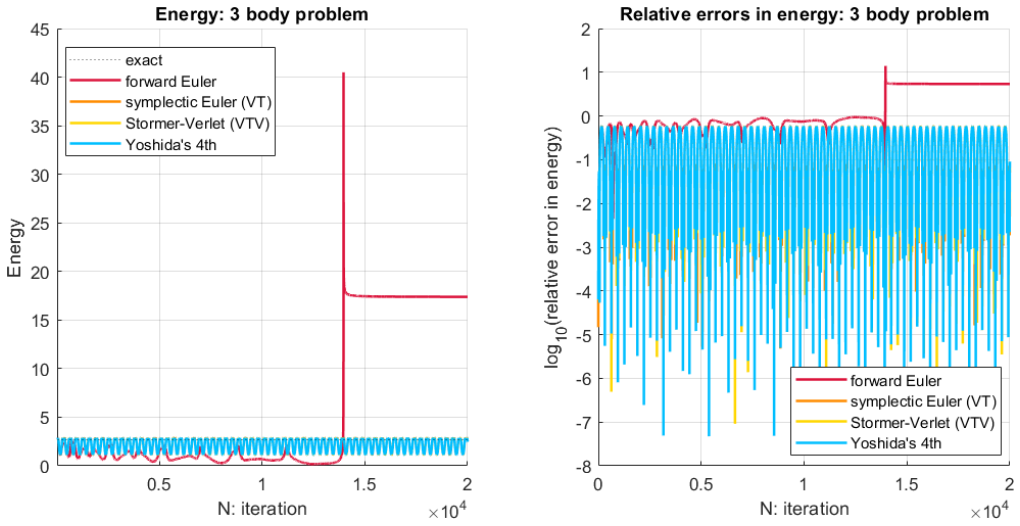


Figure 2.9: The energy in all iterations of the respective numerical schemes (left); the relative errors in energy of each numerical scheme on a logarithmic scale (right). In both cases, the step size $h = 0.01$ was used. A total amount of $N = 20000$ iterations were calculated.

It follows from the initial values that $H_0 = 2.7129$ is the system's exact initial energy. From Figure 2.9 we can see that the energy associated to the forward Euler integrator does not remain close to $H_0$, whereas the energy associated to the symplectic integrators does. Yoshida's integrator performs slightly worse than it did for the previous two examples, since its relative error in energy is no longer the lowest among the integrators.

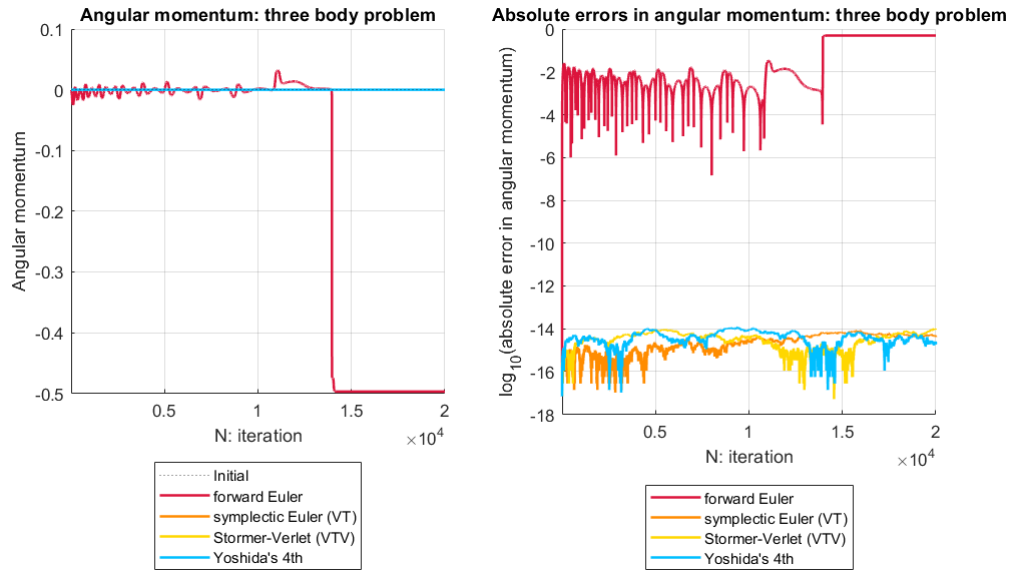


Figure 2.10: The total angular momentum in each iteration (left), and its absolute error (right) on a logarithmic scale. The step size $h = 0.01$ was used and a total amount of $N = 20000$ iterations were calculated.

By our choice of initial values and by the definition of angular momentum, the only non-zero component of the three bodies' angular momenta is the third one. The angular momentum in our case is $(0, 0, 0)$, see [7]. The figure above plots

the magnitudes of the numerically approximated angular momenta. It shows that the symplectic integrators stay much closer to the initial angular momentum than forward Euler does. This reaffirms our earlier findings.

### 2.5.4 Verification of the orders

Finally, we will numerically verify the order of the integrators when applied to the harmonic oscillator. The order can be proven analogously when applied to the other systems. By running the code for different values for the step size $h$, we find the following plots of the global error at the final time $T = N \cdot h$.

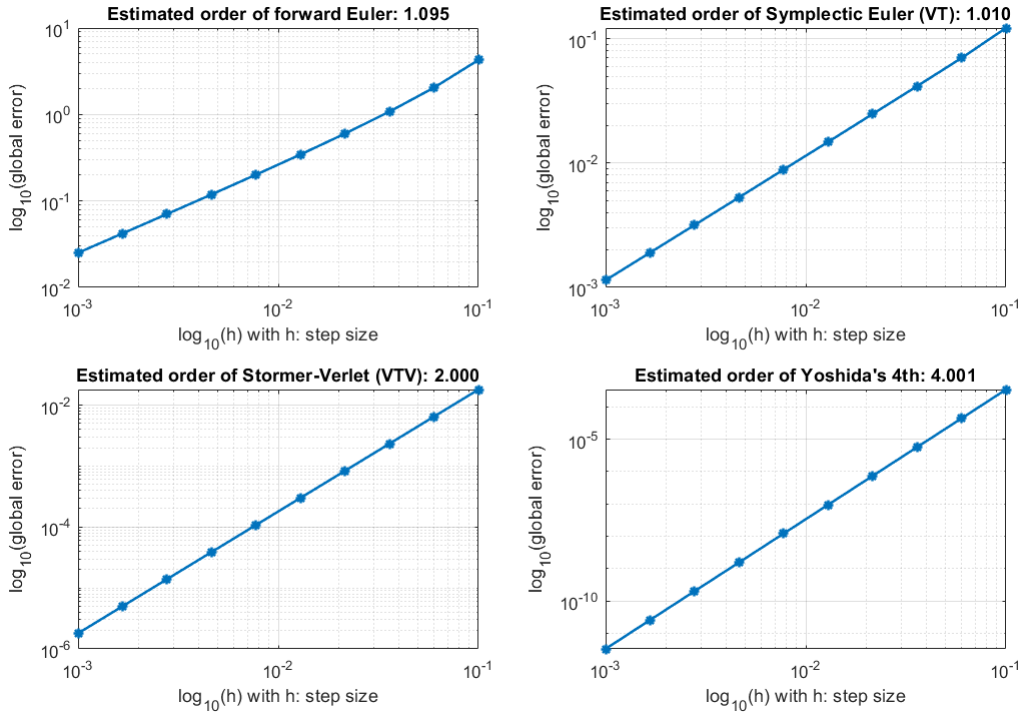


Figure 2.11: The global errors of the four integrators when applied to the harmonic oscillator for different step sizes $h$.

Figure 2.11 shows us that the orders of the forward Euler, symplectic Euler (VT), Stormer-Verlet (VTV), and Yoshida's 4th integrators are indeed 1, 1, 2 and 4. For example, if we consider the plot for Yoshida's integrator, then we see that a decrease of the stepsize by a factor of 10 leads to a decrease of the global error by a factor of about $10^4$. Alternatively, if we fit a straight line through the points on the plot, then its slope is equal to 4.001, also indicating the 4th order of the integrator.

# Chapter 3

# Characteristics-based splitting methods for PDEs

Since there exists no general existence-uniqueness theory for partial differential equations (PDEs), we need to make distinctions between different classes of equations. In this chapter, we study one particular class of PDEs: (quasi)linear PDEs of first order.

## 3.1 Classification of PDEs

In general, a partial differential equation is of the form

$$f(x_1, \ldots, x_n, u, \partial_{x_1} u, \ldots, \partial_{x_n} u, \partial^2_{x_1}, \partial^2_{x_1 x_2} u, \ldots) = 0 \tag{3.1}$$

with $u = u(x_1, \ldots, x_n)$ an unknown function such that each derivative appearing in the equation above exists (i.e. $u$ is *sufficiently smooth*), where

$$\partial^k_{x_{i_1} \ldots x_{i_k}} u = \frac{\partial^k u}{\partial x_{i_1} \ldots \partial x_{i_k}}$$

denotes a $k$-th order partial derivative of $u$. We classify PDEs based on the properties of $f$ in (3.1).

**Definition 3.1.1**
The **order** of a PDE is the highest order of a partial derivative appearing in (3.1).

For example, the equation $\partial_{x_1} u + \partial_{x_2} u = 0$ is of first order, whereas the equation $\partial^3_{x_1 x_1 x_2} u - \partial^2_{x_2 x_3} u = u$ is of third order. Another way to classify PDEs is based on the dependence of $f$ on $u$ and its partial derivatives.

**Definition 3.1.2**
A PDE (3.1) is called **linear** if it is linear in $u$ and all of its partial derivatives. **Quasilinear** PDEs are still linear in the highest order derivatives, but the coefficients are also allowed to depend on $u$ and its lower order derivatives.

## 3.2 Numerical methods for PDEs

In this section, we will discuss finite difference and operator splitting schemes. Later on, we will use finite difference schemes to qualitatively compare the solutions of characteristics-based integrators for linear first order PDEs.

### 3.2.1 Finite difference schemes

By first discretising the spatial variables and then discretising the temporal variable, a grid is created on which the numerical solution to a PDE can be computed. The spatial discretisation is done by replacing the derivatives in the equation by finite difference formulas, from which a system of linear equations can be found. These resulting linear ODEs can be solved using classical integrators, such as the forward Euler scheme. We will first consider PDEs (3.1) where the function $u$ depends on one spatial variable $x$ and a temporal variable $t$. Later, we will discuss how these methods can be adapted for functions $u = u(x, y, t)$ depending on two spatial variables. We omit the discussion of higher dimensional problems and instead refer the reader to [20].

#### 3.2.1.1 Finite difference formulas

Suppose $f : \mathbb{R} \to \mathbb{R}$ is a sufficiently smooth function and consider $h > 0$. By considering $f$'s Taylor expansion

$$f(x \pm h) = f(x) \pm hf'(x) + \frac{h^2}{2}f''(x) \pm \frac{h^3}{6}f'''(x) + \dots$$

we can deduce the identities

$$\begin{aligned}
f'(x) &= \frac{f(x+h) - f(x)}{h} + \mathcal{O}(h) \\
f'(x) &= \frac{f(x) - f(x-h)}{h} + \mathcal{O}(h) \\
f'(x) &= \frac{f(x+h) - f(x-h)}{2h} + \mathcal{O}(h^2) \\
f''(x) &= \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + \mathcal{O}(h^2)
\end{aligned}$$

where in each case, the fractions on the right-hand sides can be used to approximate the corresponding derivative of $f$ up to a certain order. In this context, the order is just the exponent of $h$ inside the big $\mathcal{O}$ symbol, contrarily to how we defined orders in the previous chapter. These fractions are called finite difference formulas. The finite difference formulas in the equations above are respectively called forward resp. backward finite differencing of first order (for the first derivative), central finite differencing of second order (for the first derivative), and central finite differencing of second order (for the second derivative). Analogous equations can also be derived for multivariate functions $f$ by using Taylor's theorem. For example, for $f : \mathbb{R}^2 \to \mathbb{R}$ it can be proven that

$$\frac{\partial f}{\partial x}(x, y) \approx \frac{f(x + h_x, y) - f(x - h_x, y)}{2h_x} \tag{3.2}$$

$$\frac{\partial f}{\partial y}(x, y) \approx \frac{f(x, y + h_y) - f(x, y - h_y)}{2h_y} \tag{3.3}$$

for $h_x, h_y > 0$, which are both first order approximations (in the sense of multivariate big $\mathcal{O}$ symbols, see [14]).

#### 3.2.1.2 Spatial-temporal discretisations

Suppose our two dimensional PDE is defined on some spatial interval $x \in [x_L, x_R]$ and consider a step size $h = m^{-1}$ for some $m \in \mathbb{N}^+$. This allows us to define a spatial lattice with grid points $x_j = x_L + jh$ for $j \geq 0$. The approximations $U_j(t) \approx u(x_j, t)$ will be computed on the points $x_j$.
To illustrate the method, we consider for example the two dimensional diffusion equation [31] given by

$$\frac{\partial u}{\partial t}(x,t) = \frac{\partial^2 u}{\partial^2 x}(x,t) \tag{3.4}$$

defined for $x \in \mathbb{R}$ and $t \in [0, T]$ with initial data $u(x, 0) = f(x)$ for some sufficiently smooth function $f$, and boundary conditions $u(x, t) = u(x + x_R, T) = 0$ for $x \in \mathbb{R}$. Because of this boundary condition, we can restrict (3.4) to the interval $[0, x_R]$. By using the second order central finite difference formula

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

we can discretise (3.4) as

$$U_j'(t) = \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{2h^2}$$

for $j \in \{1, \ldots, m\}$ where $U_0(t) = U_m(t)$ and $U_{m+1}(t) = U_1(t)$. In other words, (3.4) can be written as a linear system of ODEs

$$\frac{1}{h^2}\begin{pmatrix} U_1'(t) \\ U_2'(t) \\ \vdots \\ \vdots \\ U_{m-1}'(t) \\ U_m'(t) \end{pmatrix} = \begin{pmatrix} -2 & 1 & & & & 1 \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \\ 1 & & & & 1 & -2 \end{pmatrix} \begin{pmatrix} U_1(t) \\ U_2(t) \\ \vdots \\ \vdots \\ U_{m-1}(t) \\ U_m(t) \end{pmatrix} \tag{3.5}$$

where

$$U_0 = U_0(t) = \begin{pmatrix} u(x_1, 0) \\ u(x_2, 0) \\ \vdots \\ u(x_m, 0) \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{pmatrix}.$$

We have now discretised the PDE in all dimensions but one: this approach is called the *method of lines* [20, p. 191]. The system (3.5) can be solved by discretising the temporal variable $t$ and applying a numerical integration scheme. If we choose a time step $\tau > 0$ and consider the temporal grid points $t_j = j\tau$ for $j \geq 0$, then the numerical solution can be computed at each point $t_j$. For example, one could use the forward Euler integrator. Finally, approximate solutions will have been computed on each grid point in the spatial-temporal mesh.

#### 3.2.1.3 Example of a 2D spatial discretisation

In the previous section, we discussed how a PDE depending on only one spatial variable can be spatially discretised. We will now consider the case of a PDE in one temporal dimension and two spatial dimensions, and we use an example to demonstrate the adapted strategy. This section is based on [20].

Consider the Liouville equation (1.3) for $n = 1$ applied to the harmonic oscillator (2.34)

$$\frac{\partial \rho}{\partial \tau} + p\frac{\partial \rho}{\partial q} - q\frac{\partial \rho}{\partial p} = 0$$

defined on a 3 dimensional *box* $[\tau_{\min}, \tau_{\max}] \times [q_{\min}, q_{\max}] \times [p_{\min}, p_{\max}]$, with unknown function $\rho = \rho(q, p, t)$. The initial data is given by

$$\rho(q, p, 0) = \frac{1}{\pi \sigma_q \sigma_p} \exp\left(-\frac{(q - q_c)^2}{\sigma_q^2} - \frac{(p - p_c)^2}{\sigma_p^2}\right)$$

where $\sigma_q, \sigma_p > 0$ and we consider the following Dirichlet boundary conditions:

$$\begin{cases} \rho(q_{\min}, p, t) & = \rho(q_{\max}, p, t) = 0 \\ \rho(q, p_{\min}, t) & = \rho(q, p_{\max}, t) = 0. \end{cases} \tag{3.6}$$

We start by creating grids for $q$ and $p$

$$\begin{aligned} q_j &= q_{\min} + jh_q \\ p_j &= p_{\min} + jh_p \end{aligned}$$

where $j \geq 0$ and the step sizes $h_q, h_p > 0$ are chosen such that all $q_j$ resp. $p_j$ are equidistant. We suppose the grids for $q$ and $p$ consist of $N_q$ resp. $N_p$ points. By using the central first order finite difference formulas (3.2) and (3.3), we can discretise (1.3) as

$$R_i'(t) = q_i \frac{\rho_{i,i+1} - \rho_{i,i-1}}{2h_p} - p_i \frac{\rho_{i+1,i} - \rho_{i-1,i}}{2h_q}$$

for $1 \leq i \leq N_q N_p$ with $\rho_{i,j} := \rho(q_i, p_j, t)$. The $\rho_{ij}$ terms at the boundary of the problem can be computed from the Dirichlet conditions. This allows us to compute the matrices approximating $\partial_q$ and $\partial_p$

$$\boldsymbol{D}_z = \frac{1}{h_z} \begin{pmatrix} -1 & 1 & 0 & & \\ -\frac{1}{2} & 0 & \frac{1}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{2} & 0 & \frac{1}{2} \\ & & & -1 & 1 \end{pmatrix}$$

where $z \in \{q, p\}$. The linear system that approximately solves the PDE can be written by means of the *Kronecker product* of matrices [25].

**Definition 3.2.1**
Suppose $\boldsymbol{A} = (a_{ij})$ is an $m_1 \times m_2$ matrix and $\boldsymbol{B} = (b_{kl})$ is an $n_1 \times n_2$ matrix. Then the **Kronecker product** $\boldsymbol{A} \otimes \boldsymbol{B}$ is defined as the $m_1 n_1 \times m_2 n_2$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1m_2}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2m_2}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m_1 1}\boldsymbol{B} & a_{m_1 2}\boldsymbol{B} & \dots & a_{m_1 m_2}\boldsymbol{B} \end{pmatrix}.$$

If we denote by $\boldsymbol{\mathcal{I}}(q)$ and $\boldsymbol{\mathcal{I}}(p)$ the matrices with only the grid points $q_j$ resp. $p_j$ on their diagonals, then by [25, p. 3] we can discretise as follows

$$q\frac{\partial \rho}{\partial p} - p\frac{\partial \rho}{\partial q} \to \boldsymbol{A} := \boldsymbol{D}_p \otimes \boldsymbol{\mathcal{I}}(q) - \boldsymbol{\mathcal{I}}(p) \otimes \boldsymbol{D}_q$$

and by defining

$$R(t) = \begin{pmatrix} \rho_{0,0} \\ \rho_{1,0} \\ \vdots \\ \rho_{N_q-1,0} \\ \rho_{0,1} \\ \rho_{1,1} \\ \vdots \\ \rho_{N_q-1,1} \\ \vdots \\ \rho_{0,N_p-1} \\ \rho_{1,N_p-1} \\ \vdots \\ \rho_{N_q-1N_p-1} \end{pmatrix}$$

we obtain the linear system $R'(t) = \boldsymbol{A}R(t)$ with initial condition $R_0 = R(0)$. After solving this system using an integrator, we find the finite differences approximation to our PDE.

#### 3.2.1.4 Further reading

In this section, we have only briefly introduced finite difference schemes for PDEs. Some possible references that offer deeper or additional insights into these schemes are [15], [20], and [26].

### 3.2.2 Operator splitting

In chapter 2, we discussed splitting methods for ODEs. A similar class of integrators exists for PDEs: in the literature, they are known as operator splitting methods. Their goal is to solve PDEs of the form

$$\frac{\partial u}{\partial t} = f(t, u(x,t)) \text{ with } u(x,0) = u_0(x)$$

where $f$ is interpreted as a *spatial partial differential operator* [3, p. 86], $x \in \mathbb{R}^D$ for $D \in \mathbb{N}^+$ and $t \in \mathbb{R}$. By splitting $f$ into solvable components, an approximate solution to the total problem can be constructed. We will not go into any further detail since these integrators are not the main focus of this thesis. However, they are worth mentioning, since they extend the notion of splitting integrators for ODEs to a more general class of PDEs. Later in this chapter, we restrict our study of integrators to linear PDEs of first order.

## 3.3 Analytic solutions for first order PDEs

Linear and quasilinear PDEs of first order can be solved through what is known as the method of characteristics. In this section, we discuss this strategy and needed conditions for the uniqueness of its results. This section is based on [27] and [30].

### 3.3.1 Definitions

Consider a general quasilinear PDE of first order in two dimensions

$$a(x,y,u)\frac{\partial u}{\partial x} + b(x,y,u)\frac{\partial u}{\partial y} = c(x,y,u) \tag{3.7}$$

where the unknown function $u = u(x,y)$ depends on two variables $x$ and $y$, and the coefficients $a$, $b$ and $c$ are assumed to be adequately smooth. We will consider the solution $u$ to be a surface in $(x,y,u)$-space: such a solution surface is called an *integral surface* of the PDE. Note that (3.7) can be rewritten as

$$(a,b,c)\cdot(u_x,u_y,-1) = 0$$

where $\cdot$ denotes the Euclidean scalar product, and $u_x$ and $u_y$ denote the partial derivatives of $u$ with respect to $x$ resp. $y$. Since $\vec{N}_{(x,y)} := (u_x,u_y,-1)$ is by construction nothing more than the normal vector to $u$ at a point $(x,y)$, the equation above implies that $(a,b,c)$ is orthogonal to $\vec{N}_{(x,y)}$ for any $(x,y)$ and $(a,b,c)$ lies in the tangent plane to $u$ at the point $(x,y)$. As a result, the vector $(a,b,c)$ defines a direction on the integral surface called the *characteristic direction* or *Monge axis* of the PDE.

**Definition 3.3.1**
Let $\gamma$ be a curve in $(x,y,u)$-space such that its tangent vector at each point coincides with the corresponding Monge axis $(a,b,c)$. Then $\gamma$ is called a **characteristic curve** of (3.7) and its projection onto the plane $u = 0$ is called a **characteristic**.

If we choose a parametrisation $(x(s),y(s),u(s))$ of a characteristic curve $\gamma$, then by definition its tangent vector at some point $s$ must satisfy

$$\gamma'(s) = \left(\frac{\mathrm{d}x}{\mathrm{d}s}, \frac{\mathrm{d}y}{\mathrm{d}s}, \frac{\mathrm{d}u}{\mathrm{d}s}\right) = (a,b,c)$$

which means that in order to find the characteristic curves of (3.7), one must solve

the following (autonomous) system of ODEs:

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}s} = a(x,y,u) \\ \dfrac{\mathrm{d}y}{\mathrm{d}s} = b(x,y,u) \\ \dfrac{\mathrm{d}u}{\mathrm{d}s} = u(x,y,u). \end{cases} \tag{3.8}$$

**Definition 3.3.2**
The system of autonomous first order ODEs (3.8) are called the **characteristic equations** of (3.7). Moreover, the non-parametrical equations on the right-hand side of (3.8) are known as the **Lagrange-Charpit equations**.

In other words, the characteristic equations (3.8) describe the curves along which the PDE can be reduced to an ODE.

### 3.3.2 The method of characteristics

Characteristic curves of (3.7) can be used to find solutions to the PDE, given that certain conditions are met. These conditions are the subject of the next subsection. We now present the following results discussing the relevance of these curves.

**Proposition 3.3.3 (Myint-U and Debanth [27], p. 32, Theorem 2.3.1)**
Consider two functions $\varphi = \varphi(x,y,u)$ and $\psi = \psi(x,y,u)$, and let $f = f(\varphi,\psi)$ be an arbitrary function such that $f(\varphi,\psi) = 0$. Then $u = u(x,y)$ satisfies a first order PDE of the form

$$(\varphi_y\psi_u - \varphi_u\psi_y)\frac{\partial u}{\partial x} + (\varphi_u\psi_x - \varphi_x\psi_u)\frac{\partial u}{\partial y} = \varphi_x\psi_y - \varphi_y\psi_x. \tag{3.9}$$

We can use this proposition to prove the following theorem.

**Theorem 3.3.4 (Myint-U and Debanth [27], p. 35, Theorem 2.5.1)**
The general solution to a first order quasilinear PDE (3.7) is given by $f(\varphi,\psi) = 0$ where

1. $f$ is an arbitrary function of $\varphi(x,y,u)$ and $\psi(x,y,u)$;
2. $\varphi(x,y,u) = C_1$ and $\psi(x,y,u) = C_2$ are constant for some $C_1, C_2 \in \mathbb{R}$ and they are solutions to the characteristic equations (3.8), i.e. they are characteristic curves of the PDE.

*Proof.* Since $\varphi(x,y,u) = C_1$ and $\psi(x,y,u) = C_2$ are assumed to be characteristic curves, they must satisfy the identities

$$\begin{aligned} a\varphi_x + b\varphi_y + c\varphi_u &= 0 \\ a\psi_x + b\psi_y + c\psi_u &= 0 \end{aligned}$$

which we can solve for the coefficients $a, b, c$ to arrive at

$$\frac{a}{\varphi_y\psi_u - \varphi_u\psi_y} = \frac{b}{\varphi_u\psi_x - \varphi_x\psi_u} = \frac{c}{\varphi_x\psi_y - \varphi_y\psi_x} \tag{3.10}$$

by computing the cross product $(a, b, c) = (\varphi_x, \varphi_y, \varphi_u) \times (\psi_x, \psi_y, \psi_u)$ to find a non-trivial solution to the system. By Proposition 3.3.3, it holds that $f(\varphi, \psi) = 0$ satisfies a PDE of the form (3.9). If we plug (3.10) into this PDE, then it immediately follows that $f(\varphi, \psi) = 0$ is a solution to the quasilinear PDE (3.7). □

With these theoretical considerations in mind, we now present the step-by-step procedure [30, p. 28] used to find the solution to Cauchy problems for quasilinear PDEs. Consider the PDE (3.7) and an initial curve $\Gamma = \Gamma(x, y, u)$ parametrised by $t$ as $(x(0,t), y(0,t), u(0,t)) = (x_0(t), y_0(t), u_0(t))$ such that the solution $u$ is known along this curve. Our goal is to find the solution to the PDE in a neighbourhood of $\Gamma$. The parametrisation of $\Gamma$ determines the initial values for the characteristic curves and the solution of (3.8) leads to parametrical expressions

$$x = x(s,t), y = y(s,t), u = u(s,t)$$

which can be transformed into a solution $u = u(x, y)$ whenever the conditions of the implicit function theorem are met. Essentially, we compute the solution $u$ of the PDE along the characteristic curves and then propagate these values along the characteristic direction $(a, b, c)$. This procedure constructs the integral surface $u = u(x, y)$, yielding a solution to the Cauchy problem.

**Remark 3.3.5**
Mutatis mutandis, the method of characteristics can also be used to solve linear PDEs of first order. If we consider a PDE

$$a(x,y)\frac{\partial u}{\partial x} + b(x,y)\frac{\partial u}{\partial y} = c_0(x,y)u + c_1(x,y)$$

then the characteristic equations are given by

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}s} = a(x,y,u) \\ \dfrac{\mathrm{d}y}{\mathrm{d}s} = b(x,y,u) \\ \dfrac{\mathrm{d}u}{\mathrm{d}s} = c_0(x,y,u)u + c_1(x,y,u) \end{cases}$$

where the dependence of the coefficients on $u$ is dropped.

It is possible to generalise the method of characteristics as introduced above to higher dimensions. If we consider an $n$-dimensional quasilinear PDE

$$\sum_{i=1}^{n} a_i(x_1, \ldots, x_n, u)\frac{\partial u}{\partial x_i} = c(x_1, \ldots, x_n, u) \tag{3.11}$$

along with initial data given by a surface $\Gamma$ of codimension 2 in $\mathbb{R}^{n+1}$, parametrised as follows

$$\begin{aligned} x_0^i &= x_0^i(t_1, \ldots, t_{n-1}) \text{ for } i \in \{1, 2, \ldots, n\} \\ u_0 &= u_0(t_1, \ldots, t_{n-1}) \end{aligned}$$

then the characteristic equations (3.8) become

$$\begin{aligned}\frac{\mathrm{d}x_i}{\mathrm{d}s} &= a_i(x_1,\dots,x_n,u) \text{ for } i \in \{1,2,\dots,n\},\\ \frac{\mathrm{d}u}{\mathrm{d}s} &= c(x_1,\dots,x_n,u).\end{aligned}$$

By solving this system of ODEs and using the initial data, we find the solution $u = u(x_1,\dots,x_n)$ of (3.11) as a parametric hypersurface in $\mathbb{R}^{n+1}$. The method of characteristics is also valid for first order $n$-dimensional linear PDEs.

### 3.3.3 The existence of unique solutions

This section is based on [30]. Several problems may arise in the method of characteristics as described in the previous section. For example, the characteristic equations generally do not have a unique solution unless certain conditions are met, see Theorem 1.2.1. Another possible problem is the inability of inverting the parametrisation of the integral surface, making it impossible to find an expression in the form $u = u(x,y)$. However, when the conditions of the implicit function theorem [18, p. 661, theorem C.40] are met, such an inversion does exist and the problem is avoided. In other words, the following condition must hold.

**Definition 3.3.6 ([30], p. 36, Definition 2.9)**
Consider a quasilinear PDE of the form (3.7) with initial conditions $x(0,s) = x_0(s)$, $y(0,s) = y_0(s)$, and $u_0(s)$ defined on some initial curve $\Gamma = \Gamma(s)$ on the integral surface. Then the resulting Cauchy problem satisfies the **transversality condition** at a point $s$ on $\Gamma$ if the identity

$$\left(\frac{\mathrm{d}x}{\mathrm{d}s}\frac{\mathrm{d}y}{\mathrm{d}t} - \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}y}{\mathrm{d}s}\right)\Bigg|_{s=0} = \det\begin{pmatrix} a & b \\ \partial_t x_0 & \partial_t y_0 \end{pmatrix} \neq 0$$

is satisfied.

The determinant is 0 whenever the vectors

$$(a,b) \text{ and } \left(\frac{\mathrm{d}x_0}{\mathrm{d}t}, \frac{\mathrm{d}y_0}{\mathrm{d}t}\right)$$

are linearly dependent. The transversality condition states that the characteristic obtained by projecting the initial curve $\Gamma(s)$ onto the plane $u = 0$ nontangentially intersects the projection of $\Gamma$. For example, if the initial curve is also a characteristic curve in its own right, then the method of characteristics cannot nicely propagate information along the characteristic curve, and there exist infinitely many solutions [30, p. 39, example 2.11]. Moreover, similar problems can arise when a characteristic equation crosses the initial curve multiple times or when their projections (i.e. the characteristics) intersect. We present the following theorem, which deals with most of these problems.

**Theorem 3.3.7 ([30], p. 36, Theorem 2.10)**
We consider the Cauchy problem given by

$$\begin{cases} a(x,y,u)\dfrac{\partial u}{\partial x} + b(x,y,u)\dfrac{\partial u}{\partial y} = c(x,y,u) \\ x(0,s) = x_0(s), y(0,s) = y_0(s), u_0(s) = u(0,s) \text{ for } s \in I := ]\alpha,\beta[ \end{cases}$$

and we make the following assumptions:

**(A1)** The functions $a, b, c$ are smooth with respect to each of their variables, and $f = (a, b, c)$ is Lipschitz continuous;

**(A2)** There exist some $\delta > 0$ and $s_0 \in I$ such that for any $s \in ]s_0 - 2\delta, s_0 + 2\delta[$ the transversality condition holds.

Then the Cauchy problem has a unique solution in a neighbourhood of the initial curve, i.e. for any $(t, s) \in ]-\varepsilon, \varepsilon[ \times ]s_0 - \delta, s_0 + \delta[$ where $\varepsilon > 0$. Moreover, if the transversality condition fails on some interval of $s$-values, then the Cauchy problem either has no solution, or infinitely many solutions.

*Proof.* We follow the proof proposed in [30, section 2.5]. Since the characteristic curves of the PDE are nothing more than the solution to the system of ODEs

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= a(x,y,u) \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= b(x,y,u) \\ \frac{\mathrm{d}u}{\mathrm{d}t} &= c(x,y,u) \end{aligned} \tag{3.12}$$

with initial conditions $x(0,s) = x_0(s)$, $y(0,s) = y_0(s)$, $u(0,s) = u_0(s)$, an application of the existence theorem for ODEs results in the unicity of the characteristic curves at each point $s$ on the initial curve. The characteristic curves parametrically describe an *integral surface*, which is smooth as a result of the transversality condition. We will now prove that this surface satisfies the given PDE. To this aim, we will write $\tilde{u}(x,y) = u(t(x,y), s(x,y))$ from which it follows that

$$\begin{aligned} a(x,y,u)\tilde{u}_x(t,s) + b(x,y,u)\tilde{u}_y(t,s) &= a(x,y,u)(u_t \cdot t_x + u_s s_x) + b(x,y,u)(u_t t_y + u_s s_y) \\ &= 1 \cdot u_t + 0 \cdot u_s = u_t = c(x,y,u) \end{aligned}$$

by the chain rule and the characteristic equations (3.12). We will now prove that the constructed integral surface is unique. Recall that by construction, the characteristic curves always start on the integral surface, so we still need to prove that they remain on the surface. Geometrically, this corresponds to proving that the orthogonal projection of a characteristic curve in a point on the surface onto the surface's normal vector at that point is always zero. We will denote by $(x(t), y(t), u(t))$ a characteristic curve and define the integral surface by $u(t) = f(x,y)$. Moreover, we will define a function $\varphi(t) = u(t) - f(x(t), y(t))$. Dropping the $t$-independence for

legibility, it follows that

$$\begin{aligned}\varphi_t(t) &= u_t - f_x(x,y)x_t - f_y(x,y)y_t \\ &= c(x,y,u) - f_x(x,y)a(x,y,u) - f_y(x,y)b(x,y,u) \\ &= c(x,y,u+\varphi) - f_x(x,y)a(x,y,u+\varphi) - f_y(x,y)b(x,y,u+\varphi).\end{aligned}$$

By a direct computation, it can be checked that $\varphi(t) = 0$ is a solution to this equation. Since the equation has smooth coefficients, this solution is unique by the previous part of the theorem. Therefore $u(t) = f(x(t), y(t))$ which proves that the characteristic curves parametrically define a unique integral surface.
For the last part of the theorem, we now consider what happens when the transversality condition fails on some interval of $s$-values. If that is the case, then this means that the characteristic curves run along the initial curve, meaning that we cannot construct a solution surface to the PDE. A first scenario in which this happens is when the initial data is inconsistent with the PDE, in which case there is no solution to the Cauchy problem. A second scenario in which this happens is when the characteristic curves coincide with the initial curve, in which case it is possible to construct infinitely many integral surfaces that contain the characteristics. This proves the theorem. □

It is also possible for the transversality condition to fail on isolated points $s$ on the initial curve. In this case, it becomes hard to formulate general existence and uniqueness results and as a result, this scenario was omitted from the theorem [30, p. 38, remark 2.12].

### 3.3.4 Examples

We now consider some examples that demonstrate the method of characteristics.

**Example (Quasilinear equation)**
Consider the quasilinear first order PDE

$$(y+u)\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = x - y \tag{3.13}$$

and the initial data

$$u(x,1) = 1 + x. \tag{3.14}$$

The characteristics of (3.13) are given by

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}s} = y + u \\ \dfrac{\mathrm{d}y}{\mathrm{d}s} = y \\ \dfrac{\mathrm{d}u}{\mathrm{d}s} = x - y \end{cases}$$

and we can once again use the parametrisation

$$x_0(t) = x(0,t) = t, y_0(t) = y(0,t) = 1, u_0(t) = u(0,t) = 1 + t$$

for the initial curve $\Gamma$ on which (3.14) is defined. We first check the transversality condition

$$\det \begin{pmatrix} 2+s & 1 \\ 1 & 0 \end{pmatrix} = -1 \neq 0$$

which implies that there exists a unique integral surface near $\Gamma$. Solving the characteristic equations leads to

$$\begin{cases} x(s,t) = (1+t)e^s - e^{-s} \\ y(s,t) = e^s \\ u(s,t) = te^s + e^{-s} \end{cases}$$

from which we find the integral surface in terms of $x$ and $y$ given by

$$u(x,y) = \frac{2}{y} + x - y$$

which is once again not differentiable at the $x$-axis and therefore cannot be interpreted as a global solution to (3.7).

**Example (Linear equation: Liouville)**
We consider the Liouville PDE (1.3) for $n = 1$ applied to the harmonic oscillator (2.34), i.e. we consider the first order linear PDE

$$\frac{\partial \rho}{\partial \tau} + p\frac{\partial \rho}{\partial q} - q\frac{\partial \rho}{\partial p} = 0 \tag{3.15}$$

with unknown function $\rho(q,p,\tau)$ and initial data

$$\rho(q,p,0) = \frac{1}{\pi\sigma_q\sigma_p} \exp\left(-\frac{(q-q_c)^2}{\sigma_q^2} - \frac{(p-p_c)^2}{\sigma_p^2}\right). \tag{3.16}$$

Note that in the context of statistical mechanics, it makes sense to refer to (3.16) as the *initial density* associated to the underlying Hamiltonian system.
The characteristic equations of (3.15) are given by

$$\begin{cases} \dfrac{\mathrm{d}q}{\mathrm{d}s} = p \\ \dfrac{\mathrm{d}p}{\mathrm{d}s} = -q \\ \dfrac{\mathrm{d}\tau}{\mathrm{d}s} = 1 \\ \dfrac{\mathrm{d}\rho}{\mathrm{d}s} = 0 \end{cases}$$

and a parametrisation of the initial curve $\Gamma$ on which (3.16) is defined, is given by

$$\begin{cases} q_0(\xi_1,\xi_2) = q((\xi_1,\xi_2,0) = \xi_1 \\ p_0(\xi_1,\xi_2) = p((\xi_1,\xi_2,0) = \xi_2 \\ \tau_0(\xi_1,\xi_2) = \tau(\xi_1,\xi_2,0) = 0 \\ \rho_0(\xi_1,\xi_2) = \rho(\xi_1,\xi_2,0) = \dfrac{1}{\pi\sigma_q\sigma_p} \exp\left(-\dfrac{(\xi_1-q_c)^2}{\sigma_q^2} - \dfrac{(\xi_2-p_c)^2}{\sigma_p^2}\right) \end{cases} \tag{3.17}$$

for the parameters $\xi_1$ and $\xi_2$. The third characteristic equation is solved by $\tau = s$ and the exact solution (2.35) to the first two characteristic equations can equivalently be written as

$$\begin{cases} q_0(\xi_1, \xi_2) = \xi_1 = q\cos(s) - p\sin(s) \\ p_0(\xi_1, \xi_2) = \xi_2 = q\sin(s) + p\cos(s) \end{cases}$$

implying that $\rho(q, p, \tau)$ is given by

$$\rho(q, p, \tau) = \frac{1}{\pi\sigma_q\sigma_p} \exp\left( -\frac{(q\cos(\tau) - p\sin(\tau) - q_c)^2}{\sigma_q^2} - \frac{(q\sin(\tau) + p\cos(\tau) - p_c)^2}{\sigma_p^2} \right)$$

which is therefore the exact solution to (3.15). This solution is plotted at times $t = 0$ and $t = \pi$ in Figure 3.1.

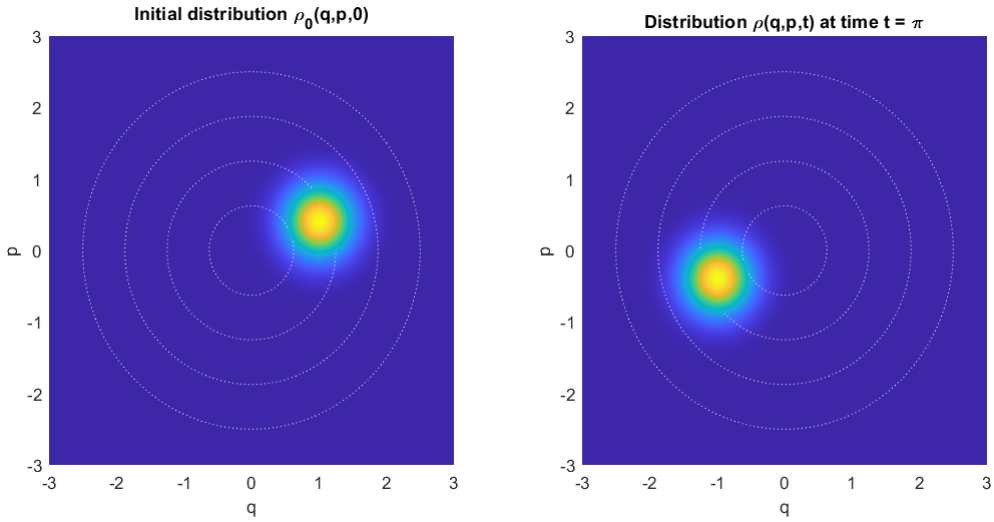


Figure 3.1: The exact solution to the Liouville application applied to the harmonic oscillator at times $t = 0$ (left) and $t = \pi$ (right). The dashed white circles represent characteristic curves of the PDE (3.15).

## 3.4 Characteristics-based splitting

We now extend our discussion of splitting methods for ODEs to the setting of PDEs. To this aim, we will focus on the Liouville equation (1.3). This is a linear PDE of first order, therefore the method of characteristics is applicable. For brevity, we call a PDE's solution obtained through the method of characteristics the *characteristic solution* of the PDE (given some initial data).

### 3.4.1 Setting

To compute the characteristic solution of the Liouville equation (1.3) with respect to some initial data, we must solve the Cauchy problem

$$\begin{cases} \dfrac{\partial \rho}{\partial \tau} = -\{H, \rho\} \\ \rho(\vec{q}, \vec{p}, 0) = \rho_0(\vec{q}, \vec{p}) \end{cases}$$

where $\rho_0$ is a given initial density function defined on some initial curve $\Gamma = \Gamma(s)$, $\vec{q}, \vec{p} \in \mathbb{R}^{2n}$, $\tau \in \mathbb{R}$, and $H : \mathbb{R}^{2n} \to \mathbb{R}$ is the Hamiltonian function of the underlying Hamiltonian system. The characteristic equations are given by

$$\begin{cases} \dfrac{\mathrm{d}q_i}{\mathrm{d}s} = \dfrac{\partial H}{\partial p_i} \\ \dfrac{\mathrm{d}p_i}{\mathrm{d}s} = -\dfrac{\partial H}{\partial q_i} \\ \dfrac{\mathrm{d}\tau}{\mathrm{d}s} = 1 \\ \dfrac{\mathrm{d}\rho}{\mathrm{d}s} = 0 \end{cases} \tag{3.18}$$

for $1 \leq i \leq n$. The first 2 (or rather, $2n$) of these listed equations correspond to the underlying Hamiltonian system's equations of motion. The third equation implies that the parameter $s$ is nothing more than the temporal variable $t$, and the final equation implies that the solution at any time $\tau$ can be found via the initial density function as follows

$$\rho(\vec{q}, \vec{p}, \tau) = \rho_0(\vec{q}(x_0, -\tau), \vec{p}(x_0, -\tau)) \tag{3.19}$$

where $x_0 = (q_0, p_0)$ is the initial value for the Hamiltonian system's equations of motion. The minus sign in front of $\tau$ appears because the Liouville equation propagates the initial density along the Hamiltonian flow induced by $H$, but backward in time.

**Example**
Consider the harmonic oscillator (2.34) and Liouville's associated PDE (3.15) with Gaussian initial density (3.16). Then at each time $t \geq 0$, the solution $\rho(q, p, t)$ with $q, p \in \mathbb{R}$ represents a Gaussian *bump* that moved backwards in time along the oscillator's flow for a time $t$.

If the underlying Hamiltonian system is integrable, then the characteristic equations (3.18) can be solved analytically. This is the case for the harmonic oscillator. If it is not, then numerical integration is needed to compute the characteristic curves and to find the PDE's characteristic solution.
Since the first $2n$ characteristic equations in (3.18) form a Hamiltonian system, we can use symplectic integrators to approximate the solution

$$(Q_i(x_0, -\tau), P_i(x_0, -\tau)) \text{ where } 1 \leq i \leq n$$

and by plugging this into the given initial density function, we find the numerical characteristic solution

$$R(\vec{Q}, \vec{P}, \tau) = \rho_0(\vec{Q}(x_0, -\tau), \vec{P}(x_0, -\tau))$$

at the time $\tau$, where $\vec{Q} = (Q_1, Q_2, \ldots, Q_n)$ and $\vec{P} = (P_1, P_2, \ldots, P_n)$.

### 3.4.2 Numerical implementation in MATLAB

The numerical implementation of our proposed characteristics-based integrators for the Liouville equation can be found in the appendix C.1. We consider a variable

$s$ parametrising the initial curve $\Gamma$ and a variable $t$ parametrising the time interval for integration. Discretising these variables results in two evenly spaced grids $s_1, s_2, \dots s_{N_s}$ and $t_1, t_2, \dots t_{N_t}$ with steps $ds$ and $dt$. We will now discuss our algorithm to solve the characteristic equations.

The algorithm starts by iterating over the $s$-grid. For each point $s_{\text{iter}}$ on the initial curve, the initial value $(q_0^{\text{iter}}, p_0^{\text{iter}}) = (q_0^{\text{iter}}(s), p_0^{\text{iter}}(s))$ is computed (see (3.17), for example) and used to perform a time integration over the $t$-grid. This temporal integration is carried out by a selected integrator: forward Euler, symplectic Euler (VT), Stormer-Verlet (VTV), or Yoshida's 4th order integrator. The approximated values for $(q, p)$ are then stored row-wise in the matrices $\boldsymbol{Q}, \boldsymbol{P}$. For example, to plot the approximate characteristic curve emanating from the $j$-th point $s_j$ on $\Gamma$, one would plot $\boldsymbol{Q}$'s $j$-th row against $\boldsymbol{P}$'s $j$-th row. The columns of these matrices represent the values of the characteristic curves at a specific time $t_k$.

After completing the iteration over the $s$-grid, the approximate characteristic solution can be computed from the initial density, i.e. $\rho(q, p, T) = \rho_0(\boldsymbol{Q}(-T), \boldsymbol{P}(-T))$. This solution is stored in a matrix $\boldsymbol{R}$, where the $k$-th row represents the solution's values along the characteristic curve emanating from the point $s_k$ on the $\Gamma$. Similarly, $\boldsymbol{R}$'s $k$-th column represents the values of the different characteristic curves at the time $t_k$. To fill in the elements of $\boldsymbol{R}$, we loop over the $s$- and $t$-grids as follows.

$$\begin{aligned} s\text{-loop} &: 1 \to N_s \\ t\text{-loop} &: \left\lfloor \frac{N_t}{2} + 1 \right\rfloor \to N_t \end{aligned}$$

The $t$-loop is cut in half to ensure that the computed characteristic curves for our numerical examples (i.e. using the harmonic oscillator) only perform one full orbit through phase space. This is because the harmonic oscillator's exact solution is $2\pi$-periodic in $t$, see (2.35): the resulting characteristic curves pass through the line $s = 0$ at each positive integer multiple of $\pi$. If we considered an iteration $1 \to N_t$, the plots in section 3.5 would be less clear.

Our algorithm was written with the harmonic oscillator in mind, and as a result the underlying Hamiltonian system is assumed to be 2-dimensional. However, our algorithm can be adapted to systems of higher dimensions $2n$. In this case, more matrices $\boldsymbol{Q}_i$, $\boldsymbol{P}_i$ with $1 \leq i \leq n$ would be needed to store the approximations and only one matrix $\boldsymbol{R}$ would be required, since the initial density remains a scalar function.

### 3.4.3 Error analysis

In this section, we will present a concise procedure to perform an error analysis of our characteristics-based integrators. We will denote by $\rho_A = \rho_a(\vec{q}, \vec{p}, t)$ the analytical solution to (3.15) and by $\rho_N = \rho_N(\vec{q}, \vec{p}, t)$ its numerical solution.

#### 3.4.3.1 The $L^1$ error

Since we are considering the Liouville equation, the initial data $\rho_0(\vec{q},\vec{p})$ is a probability density function and as a result, its integral over $\mathbb{R}^{2n}$ satisfies the identity

$$\int_{\mathbb{R}^{2n}} |\rho_0(\vec{q},\vec{p})| d\vec{x} = \int_{\mathbb{R}^{2n}} \rho_0(\vec{q},\vec{p}) d\vec{x} = 1$$

where $\vec{x} := (\vec{q},\vec{p}) \in \mathbb{R}^{2n}$ and $d\vec{x} = dq_1 \dots dq_n dp_1 \dots dp_n$. This identity implies that the initial density $\rho_0 \in L^1(\mathbb{R}^{2n})$, i.e. the integral over $\mathbb{R}^{2n}$ of $|\rho_0|$ is finite. Since the analytical solution $\rho_A(\vec{q},\vec{p},t)$ at a time $t$ can be obtained from the initial density by (3.19), it follows from Liouville's Theorem 1.1.20 that $\rho_A(\vec{q},\vec{p},t) \in L^1(\mathbb{R}^{2n})$ for any $t \in \mathbb{R}$. As a result, it makes sense to study the $L^1$-error of our proposed integrators. This can be done by using the distance function

$$\mathrm{d}(f_1, f_2) := \int_{\mathbb{R}^{2n}} |f_1(\vec{x},t) - f_2(\vec{x},t)| d\vec{x}$$

where $f_1(\vec{x},t)$, $f_2(\vec{x},t) \in L^1(\mathbb{R}^{2n})$ for any fixed $t \in \mathbb{R}$.

**Definition 3.4.1**
The $L^1$ **error** of the characteristics-based integrators at a time $t \in \mathbb{R}$ is defined by

$$\mathrm{err}_{L^1}(t) := \mathrm{d}(\rho_A, \rho_N) = \int_{\mathbb{R}^{2n}} |\rho_A(\vec{x},t) - \rho_N(\vec{x},t)| d\vec{x}.$$

In practice, this definition poses two problems: we cannot numerically integrate functions over unbounded intervals, and $\rho_N$ is computed as a discrete array of numbers: it is not implemented as a continuous function. As a result, we must restrict the problem's phase space to a bounded set $\Omega \subset \mathbb{R}^{2n}$ and approximate the integral by an $n$-fold summation of absolute errors. For example, if we consider the harmonic oscillator (with $n = 1$), then the $L^1$ error at a time $t_k$ can be numerically computed as

$$\mathrm{err}_{L^1}(t_k) \approx \Delta s \sum_{j=1}^{N_s} \left| R^{\mathrm{ex}}_{j,k} - R^{\mathrm{num}}_{j,k} \right|$$

where $R^{\mathrm{ex}}_{j,k}$ represents the element at position $(j,k)$ of the matrix containing the exact solution at grid points $(s_j, t_k)$, and analogously for $R^{\mathrm{num}}_{j,k}$. Intuitively, we expect this error to follow the same trends as observed in section 2.5: the forward Euler integrator has the biggest error, and the symplectic schemes' errors are smaller.

In section 3.5, we consider the Liouville equation (3.15) with Gaussian initial density (3.16). The following proposition implies that the corresponding analytical solution $\rho_A(q,p,t) \in L^1(\mathbb{R}^{2n})$ for any fixed $t \in \mathbb{R}$, and as a result, the $L^1$ error is well-defined.

**Proposition 3.4.2**
Consider constants $\sigma_q$, $\sigma_p > 0$ and $(q_c, p_c) \in \mathbb{R}^2$. The function $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ defined as

$$f(q,p,t) := \frac{1}{\pi \sigma_q \sigma_p} \exp\left( -\frac{(q-q_c)^2}{\sigma_q^2} - \frac{(p-p_c)^2}{\sigma_p^2} \right)$$

is of class $L^1(\mathbb{R}^2)$ for any $t \in \mathbb{R}$, i.e. $\int_\Omega |f| dq dp < +\infty$.

*Proof.* The result follows from the identity

$$\int_{-\infty}^{+\infty} \exp(-x^2)dx = \sqrt{\pi}$$

and by computing the integral

$$\begin{aligned}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left| \frac{1}{\pi\sigma_q\sigma_p} \exp\left(-\frac{(q-q_c)^2}{\sigma_q^2} - \frac{(p-p_c)^2}{\sigma_p^2}\right)\right| dqdp &= \frac{1}{\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\left(-\tilde{q}^2-\tilde{p}^2\right) d\tilde{q}d\tilde{p} \\ &= \frac{\sqrt{\pi}\cdot\sqrt{\pi}}{\pi} \\ &= 1 < +\infty\end{aligned}$$

using the change of variables

$$(\tilde{q},\tilde{p}) = \left(\frac{q-q_c}{\sigma_q}, \frac{p-p_c}{\sigma_p}\right).$$

□

#### 3.4.3.2 Error in energy

Since the characteristic curves can be found by solving the equations of motion associated to the underlying Hamiltonian system, the symplectic nature of the ODE is, in some sense, inherited by the PDE. Consequently, we would like to define a notion of error in energy.

To this aim, recall that in classical mechanics, the state of a system is represented by coordinates $(\vec{q},\vec{p})$ where $\vec{q}$ denotes the position and $\vec{p}$ denotes momenta. For autonomous Hamiltonian systems, the total energy is given by a function

$$H : \mathbb{R}^{2n} \to \mathbb{R} : (\vec{q},\vec{p}) \mapsto H(\vec{q},\vec{p})$$

which we will now use to define the notion of *error in energy* for our proposed integrators. Consider a function $\rho(\vec{x},t)$ such that $\rho(\vec{x},t) \in L^1(\mathbb{R}^{2n})$ for any fixed $t \in \mathbb{R}$. Since $(L^1(\mathbb{R}^{2n}))^* \cong L^\infty(\mathbb{R}^{2n})$ (see [10, chapter 6]), we can define a functional

$$\mathcal{H} : \rho_t \mapsto \langle H \rangle := \int_{\mathbb{R}^{2n}} H(\vec{x})\rho_t(\vec{x})d\vec{x}$$

where $\rho_t(\vec{x}) := \rho(\vec{x},t) \in L^1(\mathbb{R}^{2n})$. This functional can be used to associate an *energy* to a solution. Note that since $H(\vec{x})$ is generally not bounded, the integral in the definition of $\mathcal{H}$ may diverge. However, since in the sequel we will always compute these integrals over compact domains, this will not pose a problem.

**Definition 3.4.3**
If $\rho_A$ and $\rho_N$ denote the analytical resp. numerical solutions to (1.3), then we

define the **error in energy** as

$$\Delta H(t) := \left| \int_{\mathbb{R}^{2n}} H(\vec{x}) \left(\rho_A(\vec{x}, t) - \rho_N(\vec{x}, t)\right) d\vec{x} \right|.$$

In practice, this definition poses the same problems as the $L^1$ errors: we need to restrict $\mathbb{R}^{2n}$ to a bounded subset $\Omega \subset \mathbb{R}^{2n}$ and the integral needs to be approximated using an $n$-fold summation.

The following proposition provides an upper bound for the error in energy.

**Proposition 3.4.4**
If $\rho_A$ and $\rho_N$ denote the analytical resp. numerical solutions to (1.3) and $\Omega \subset \mathbb{R}^{2n}$ is bounded, then the error in energy satisfies the following upper bound:

$$\Delta H(t) \leq |\Omega| \cdot \max_{\vec{x} \in \Omega} \{|H(\vec{x})|\} \cdot d(\rho_A, \rho_N).$$

*Proof.* The result follows from the following computations:

$$\begin{aligned}
\Delta H(\vec{x}) &= \left| \int_\Omega H(\vec{x}) \rho_A(t, \vec{x}) d\vec{x} - \int_\Omega H(\vec{x}) \rho_N(t, \vec{x}) d\vec{x} \right| \\
&\leq \int_\Omega |H(\vec{x})| \cdot |\rho_A(t, \vec{x}) - \rho_N(t, \vec{x})| \, d\vec{x} \\
&\leq \int_\Omega |H(\vec{x})| \, d\vec{x} \cdot \int_\Omega |\rho_A(t, \vec{x}) - \rho_N(t, \vec{x})| \, d\vec{x} \\
&\leq \int_\Omega d\vec{x} \cdot \max_{\vec{x} \in \Omega} \{|H(\vec{x})|\} \cdot d(\rho_A, \rho_N) \\
&= |\Omega| \cdot |\max_{\vec{x} \in \Omega} \{|H(\vec{x})|\} \cdot d(\rho_A, \rho_N).
\end{aligned}$$

□

For example, if we consider the harmonic oscillator and restrict its phase space to the square grid $\Omega = [-a, a] \times [-a, a]$ for some $a > 0$, then we get the upper bound

$$\Delta H(t) \leq a^2 \cdot \frac{a^2}{2} \cdot d(\rho_A, \rho_N) = \frac{a^4}{2} d(\rho_A, \rho_N).$$

## 3.5 Numerical examples

We now consider the Liouville equation (3.15) applied to the harmonic oscillator (2.34). For more details on the MATLAB code that was used, see the Appendices B (integrators) and C (algorithm).

### 3.5.1 Finite difference approximation

In section 3.2.1.3, we semidiscretised (3.15) by imposing the Dirichlet boundary conditions (3.6) and by using the first order finite difference (FD) formulas (3.2) and (3.3). Note that these boundary conditions make sense, since the initial data (a Gaussian) quickly tends to 0 as $q, p \to \infty$. After semidiscretisation, we obtained

a system of the form $R'(t) = \boldsymbol{A}R(t)$ which we will solve using Runge-Kutta's fourth order integrator (A.2).

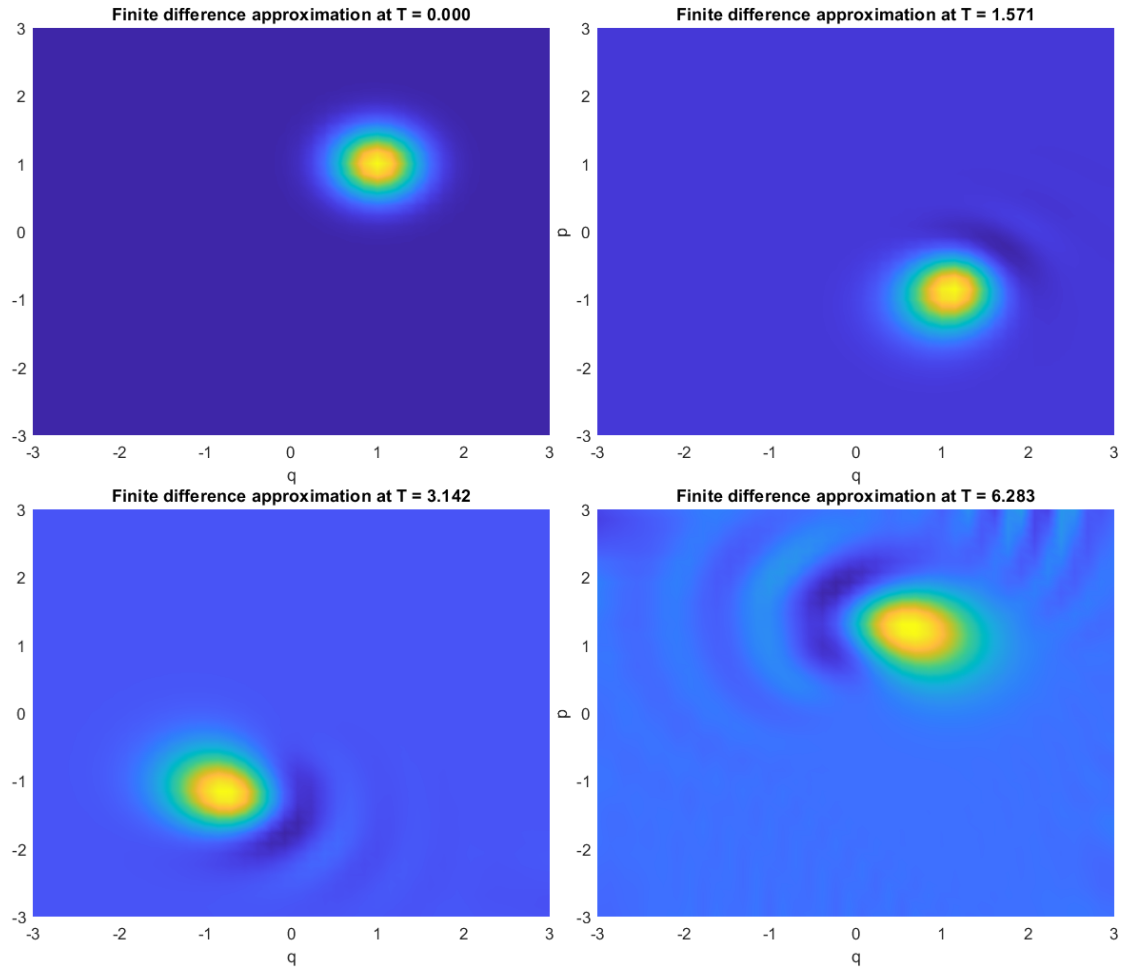


Figure 3.2: The finite difference solutions of the Liouville equation applied to the harmonic oscillator. The spatial discretisation was done by using first order central difference schemes in both $q$- and $p$-directions, and the temporal discretisation was carried out by Runge-Kutta's fourth order integrator.

The figure above shows why the proposed FD scheme is not an optimal choice of integrator for the Liouville equation. As time progresses, the FD approximations start showing *parasitic waves* [24, p. 326] or *spurious oscillations* which are due to the numerical method's stability and which have nothing to do with the problem's actual solution. The numerical solutions no longer represent probability distributions, since they take on negative values. The figure also shows that the Gaussian starts losing it shape after a full orbit: the *bump* looks slightly elongated. Moreover, the FD scheme does not conserve the distribution's mass, as the following graph shows.

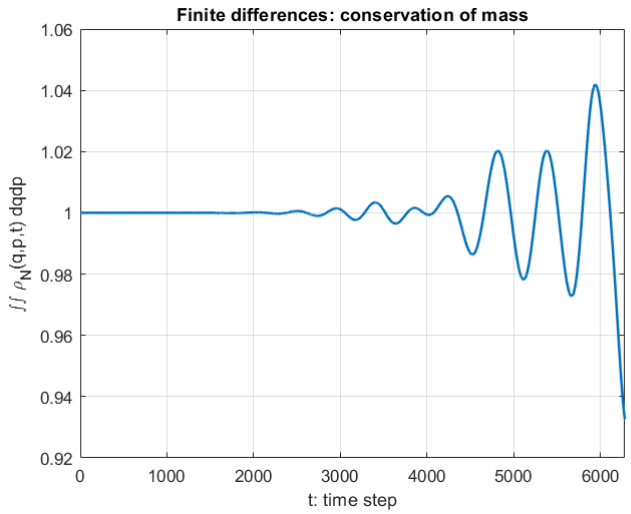


Figure 3.3: The numerical mass in each iteration of the FD scheme applied to the Liouville equation (harmonic oscillator).

For a probability distribution, the total unit mass should be conserved at each possible time $t$. According to the figure above, the mass is clearly not conserved by the FD integrator: the graph shows clear oscillatory behaviour.

### 3.5.2 Characteristics-based integrators

In this section, we will discuss the visual behaviour of the solutions obtained by the algorithm described in section 3.4.2.

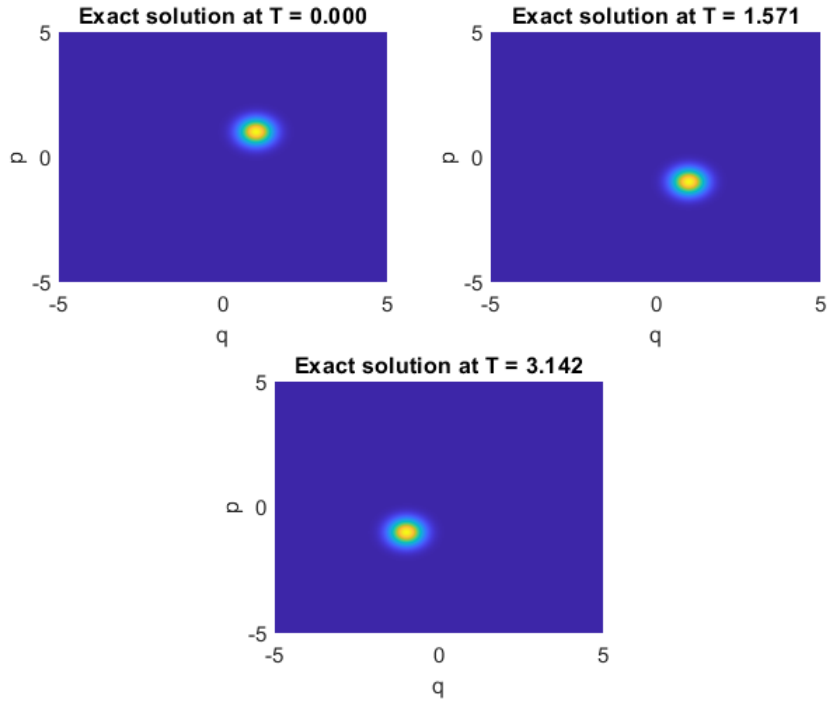


Figure 3.4: The exact solution to the Liouville equation applied to the harmonic oscillator at three different times: $T = 0$, $T = \frac{\pi}{2}$, and $T = \pi$.

This first graph shows the exact solutions at three different times: it shows the initial distribution at time $T = 0$, the distribution after a quarter orbit at time $T = \frac{\pi}{2}$, and the distribution after a half orbit at time $T = \pi$. By construction, the characteristic solution translates the Gaussian unaltered along the characteristic curves.

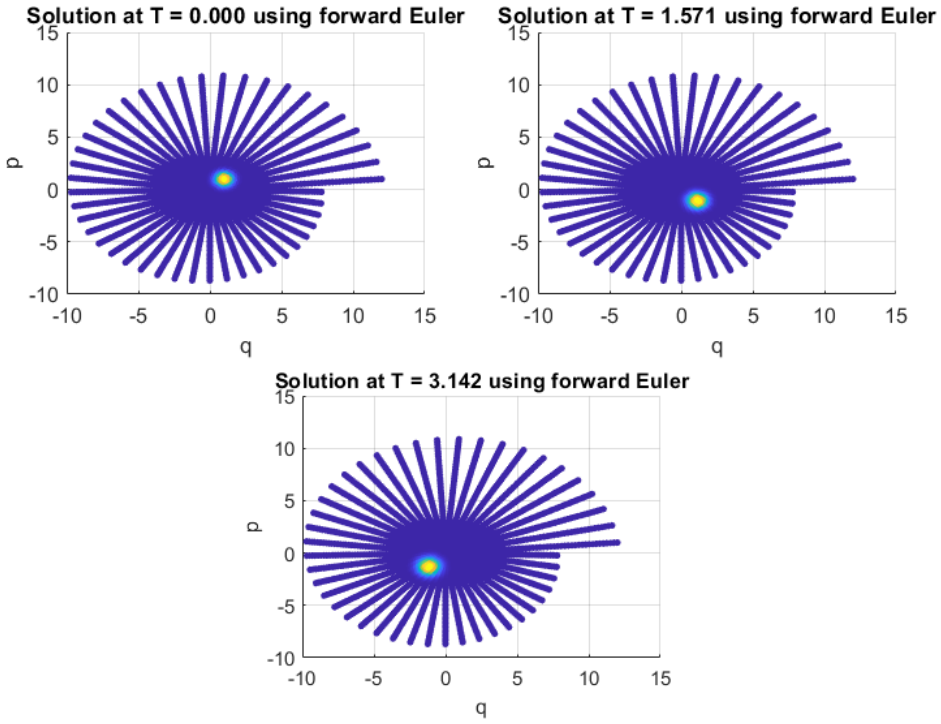


Figure 3.5: The numerical solution to the Liouville equation with the harmonic oscillator as its underlying Hamiltonian system, at times $T = 0$, $T = \frac{\pi}{2}$, and $T = \pi$. The forward Euler integrator was used.

The figure above shows us the numerical solution using the forward Euler integrator and each *radial* line represents a characteristic curve. We clearly see the same spiralling behaviour that was previously observed in Figure 2.1. Similar to the ODE setting, this behaviour is again a negative attribute of the forward Euler integrator: because of the horizontal gap formed by the spiral, the Gaussian will not be translated unaltered over large time intervals. This is because forward Euler does not conserve the symplectic structure present in the characteristic curves. As a result, we also expect the energy (or probability mass) to blow up over increasingly big time intervals, as was the case in the ODE setting.

The next three figures 3.6, 3.7, and 3.8 show the solutions of our algorithm obtained by four different integrators: symplectic Euler (VT), Stormer-Verlet (VTV), and Yoshida's 4th order integrator.

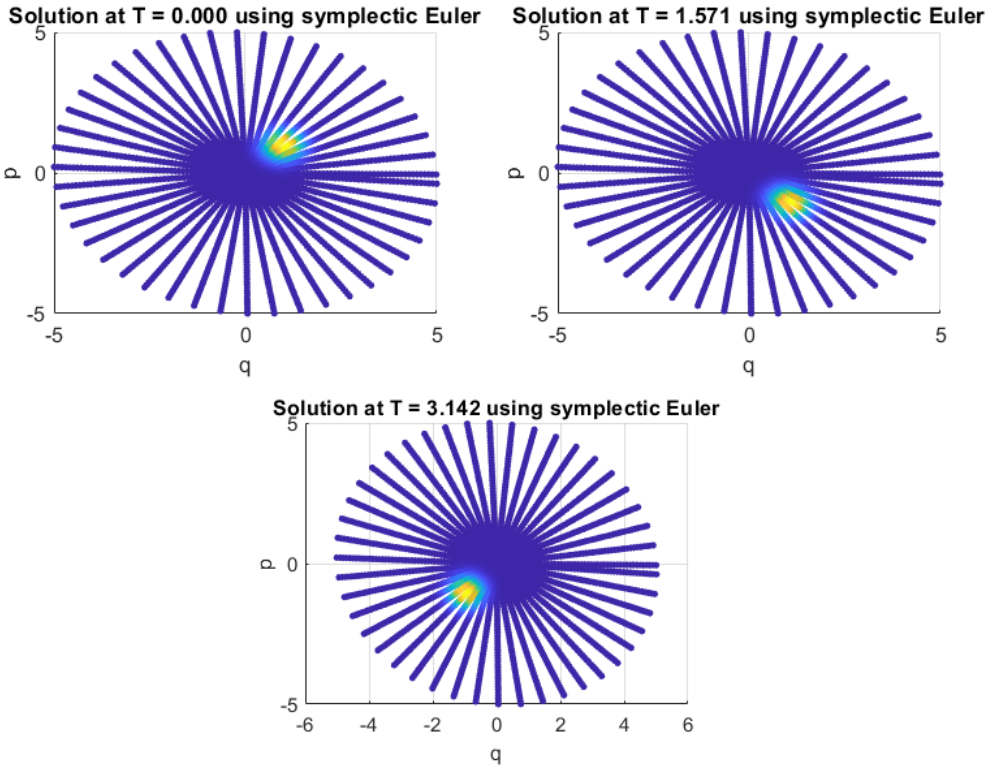


Figure 3.6: The numerical solution to the Liouville equation with the harmonic oscillator as its underlying Hamiltonian system, at times $T = 0$, $T = \frac{\pi}{2}$, and $T = \pi$. The symplectic Euler (VT) integrator was used.

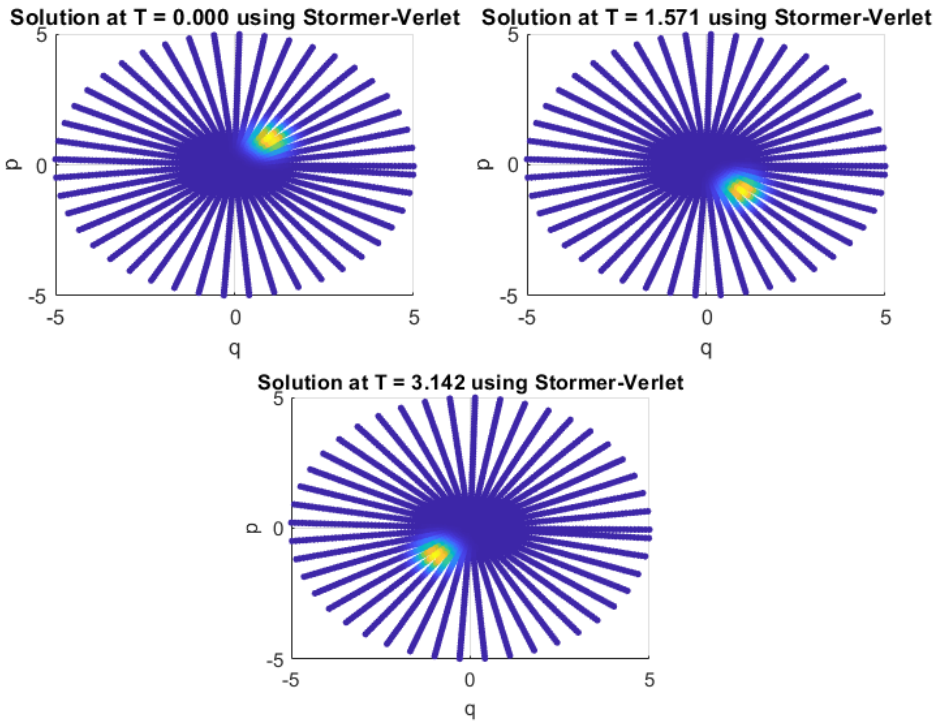


Figure 3.7: The numerical solution to the Liouville equation with the harmonic oscillator as its underlying Hamiltonian system, at times $T = 0$, $T = \frac{\pi}{2}$, and $T = \pi$. The Stormer-Verlet (VTV) integrator was used.

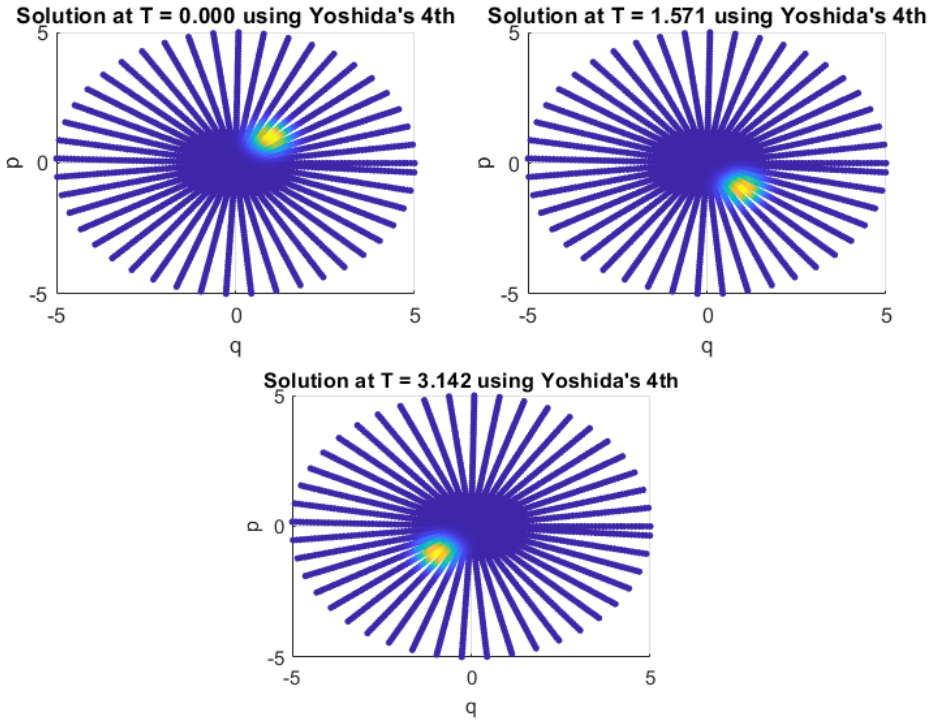


Figure 3.8: The numerical solution to the Liouville equation with the harmonic oscillator as its underlying Hamiltonian system, at times $T = 0$, $T = \frac{\pi}{2}$, and $T = \pi$. Yoshida's 4th order integrator was used.

None of these symplectic integrators show spiralling behaviour, as was the case for forward Euler. Visually, the characteristic curves seem to be radial slices of circles centered at the origin. As a result, we expect that the $L^1$ and energy errors do not blow up, implying that the Gaussian *bump* is translated around the characteristic curves without being undesirably altered or reshaped, as was also the case for the FD solution at $T = 2\pi$.

Moreover, none of the solutions obtained by our algorithm show spurious oscillations. In the FD solution, these oscillations could be observed even after a relatively short time $T = \frac{\pi}{2}$. As a result, our algorithm seems to be more stable, at least when applied to this problem. However, this is only a *visual* conclusion based on the figures, since we did not use the same step sizes in the FD scheme and the proposed algorithm. A more thorough stability analysis is needed to substantiate these claims.

# Conclusion and future outlook

In Chapter 1, we discussed geometrical and analytical preliminaries that are used throughout this thesis. In the following Chapter 2, we performed a literary review on the subject of splitting integrators for ODEs. We focussed in particular on Hamiltonian systems induced by Hamiltonians of the form $H = T + V$. We discussed how splitting integrators preserve the underlying geometrical (symplectic) structure, making them a better choice of numerical scheme as opposed to classical integrators, such as forward Euler. Through a backward error analysis, we showed that symplectic splitting integrators nearly preserve the energy of Hamiltonian systems. In Chapter 3, we discussed the method of characteristics to find analytical solutions to (quasi)linear PDEs of first order. We presented a numerical scheme based on the method of characteristics that can be used to solve linear PDEs of first order. We applied these schemes to the Liouville equation focused on the harmonic oscillator as the main example, with a comparison to a finite difference (FD) scheme. We found that our algorithm's solution appears more stable: the FD solution shows spurious oscillations, whereas the characteristics-based integrators do not, even if further analyses still have to be performed.

The work done in this thesis can be extended in multiple directions. In Chapter 2, we only reviewed the subclass of symplectic integrators. In Chapter 3, the algorithm that we proposed for linear PDEs of first order could be modified to also work for quasilinear PDEs. Moreover, we only discussed the Liouville equation, since its characteristic curves can be computed by solving a Hamiltonian system. The algorithm could also be modified to work for other PDEs that share this property. In the numerical examples of section 3.5, we did not perform a thorough error analysis of our integrators. Such an error analysis would be a valuable addition to this thesis, providing a greater understanding of the accuracy and stability of the characteristics-based integrators. Furthermore, geometrical splitting methods have also been introduced for contact Hamiltonian dynamics, see [4]. The characteristics-based splitting methods proposed in Chapter 3 could be extended to the Liouville equation for contact systems.

Revisiting the introduction's political metaphor, we could say that this thesis is a testament to the power of *dividing and conquering* difficult problems. By decomposing differential equations into simpler, solvable subproblems, we demonstrated the efficacy of splitting methods in formulating answers to the original equation. We divided, we integrated, and we conquered differential equations - *veni, vidi, vici*!

# List of Figures

# Bibliography


[1] V. I. Arnold. *Mathematical Methods of Classical Mechanics*. Jan. 1989. DOI: 10.1007/978-1-4757-2063-1. URL: https://doi.org/10.1007/978-1-4757-2063-1.

[2] Sergio Blanes, Fernando Casas, and Ander Murua. "On the Linear Stability of Splitting Methods". In: *Foundations of Computational Mathematics* 8 (June 2008). DOI: 10.1007/s10208-007-9007-8.

[3] Sergio Blanes, Fernando Casas, and Ander Murua. *Splitting Methods for differential equations*. Jan. 3, 2024. URL: https://arxiv.org/abs/2401.01722.

[4] Alessandro Bravetti et al. "Numerical integration in Celestial Mechanics: a case for contact geometry". In: *Celestial Mechanics and Dynamical Astronomy* 132 (Jan. 2020). DOI: 10.1007/s10569-019-9946-9.

[5] John Butcher. "Numerical Methods for Ordinary Differential Equations, Second Edition". In: *Numerical Methods for Ordinary Differential Equations, Second Edition* (Mar. 2008), pp. 1–463. DOI: 10.1002/9780470753767.

[6] Xueying Chen, Jerry Q. Cheng, and Min-ge Xie. *Divide-and-conquer methods for big data analysis*. 2021. arXiv: 2102.10771 [stat.ML]. URL: https://arxiv.org/abs/2102.10771.

[7] Alain Chenciner and Richard Montgomery. "A Remarkable Periodic Solution of the Three-Body Problem in the Case of Equal Masses". In: *Annals of Mathematics* 152.3 (2000), pp. 881–901. ISSN: 0003486X, 19398980. URL: http://www.jstor.org/stable/2661357 (visited on 07/18/2025).

[8] Vicente Cortés and Alexander S. Haupt. *Mathematical Methods of Classical Physics*. Jan. 2017. DOI: 10.1007/978-3-319-56463-0. URL: https://doi.org/10.1007/978-3-319-56463-0.

[9] Ana Cannas Da Silva. *Lectures on Symplectic Geometry*. Jan. 2008. DOI: 10.1007/978-3-540-45330-7. URL: https://doi.org/10.1007/978-3-540-45330-7.

[10] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. ISBN: 9781118626399. URL: https://books.google.be/books?id=wI4fAwAAQBAJ.

[11] Geeks for Geeks. *Introduction to Divide and Conquer Algorithm*. 2025. URL: https://www.geeksforgeeks.org/dsa/introduction-to-divide-and-conquer-algorithm/ (visited on 08/05/2025).

[12] Ernst Hairer, Gerhard Wanner, and Christian Lubich. *Geometric Numerical Integration*. Jan. 2006. DOI: 10.1007/3-540-30666-8. URL: https://doi.org/10.1007/3-540-30666-8.

[13] Andreas Henriksson. "Liouville's theorem and the foundation of classical mechanics". In: *Lithuanian Journal of Physics* 62.2 (July 2022). DOI: 10.3952/physics.v62i2.4740. URL: https://arxiv.org/abs/1905.06185.

[14] Rodney Howell. "On Asymptotic Notation with Multiple Variables". In: (Jan. 2007).

[15] Arieh Iserles and Gilbert Strang. "The Optimal Accuracy of Difference Schemes". In: *Transactions of The American Mathematical Society - TRANS AMER MATH SOC* 277 (June 1983), pp. 779–779. DOI: 10.1090/S0002-9947-1983-0694388-9.

[16] Jacek Jachymski, Izabela Jóźwik, and Małgorzata Terepeta. "The Banach Fixed Point Theorem: selected topics from its hundred-year history". In: *Revista de la Real Academia de Ciencias Exactas Físicas y Naturales Serie A Matemáticas* 118.4 (July 2024). DOI: 10.1007/s13398-024-01636-6. URL: https://doi.org/10.1007/s13398-024-01636-6.

[17] Hiroshi Kinoshita, Haruo Yoshida, and Hiroshi Nakai. "Symplectic integrators and their application to dynamical astronomy". In: *Celestial Mechanics and Dynamical Astronomy* 50.1 (Jan. 1, 1991), pp. 59–71. DOI: 10.1007/bf00048986. URL: https://doi.org/10.1007/bf00048986.

[18] John M. Lee. *Introduction to Smooth Manifolds*. 2013. DOI: 10.1007/978-1-4419-9982-5. URL: https://julianchaidez.net/materials/reu/lee_smooth_manifolds.pdf.

[19] Bram Lens. *Splitting Thesis GitHub Repository*. 2025. URL: https://github.com/bramlens/splitting-thesis (visited on 08/15/2025).

[20] Randall Leveque. "Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems". In: Jan. 2007. DOI: 10.1137/1.9780898717839.

[21] Jervin Zen Lobo and Terence Johnson. "Solution of Differential Equations using Exponential of a Matrix". In: *IOSR Journal of Mathematics* 5 (Jan. 2013), pp. 12–17. DOI: 10.9790/5728-0531217.

[22] Robert Mclachlan and G. Quispel. "Geometric integrators for ODEs". In: *J. Phys. A* 39 (May 2006), pp. 5251–. DOI: 10.1088/0305-4470/39/19/S01.

[23] Robert Mclachlan and G. Quispel. "Splitting methods". In: *Acta Numer.* 11 (Jan. 2002), pp. 341–. DOI: 10.1017/S0962492902000053.

[24] Robert I. McLachlan. *Explicit Symplectic Methods applied to PDE's*. Accessed: 2025-08-01. 1993. URL: https://www.massey.ac.nz/~rmclachl/ExplicitSymplectic.pdf.

[25] Mohammad Mirhosseini, Hossein Rahami, and A. Kaveh. "Analytical Solution of Laplace and Poisson Equations Using Conformal Mapping and Kronecker Products". In: *International Journal of Civil Engineering* 14 (June 2016). DOI: 10.1007/s40999-016-0037-y.

[26] K. W. Morton and D. F. Mayers. *Numerical Solution of Partial Differential Equations: An Introduction*. 2nd ed. Cambridge University Press, 2005.

[27] Tyn Myint-U and Lokenath Debnath. *Linear Partial Differential Equations for Scientists and Engineers*. Jan. 2007. ISBN: 978-0-8176-4393-5. DOI: 10.1007/978-0-8176-4560-1.

[28] R. K. Pathria and Paul D. Beale. *Statistical Mechanics*. 3rd ed. Amsterdam ; Boston: Elsevier/Academic Press, 2011. ISBN: 978-0-12-382188-1.

[29] J.C. Pim. "On Yoshida's Method For Constructing Explicit Symplectic Integrators For Separable Hamiltonian Systems". Accessed at 2025-08-01. Bachelor's thesis. University of Groningen, July 2019. URL: https://fse.studenttheses.ub.rug.nl/20185/1/bMATH_2019_PimJC.pdf.

[30] Yehuda Pinchover and Jacob Rubinstein. "An Introduction to Partial Differential Equations". In: Cambridge University Press, 2005.

[31] Tsegaye Simon and Purnachandra Rao Koya. "Application of Some Finite Difference Schemes for Solving One Dimensional Diffusion Equation". In: *American Scientific Research Journal for Engineering, Technology, and Sciences* 26.3 (Nov. 2016), pp. 140–154. URL: https://asrjetsjournal.org/American_Scientific_Journal/article/view/1936.

[32] John Stillwell. *Naive lie theory*. Jan. 1, 2008. DOI: 10.1007/978-0-387-78214-0. URL: https://doi.org/10.1007/978-0-387-78214-0.

[33] Gerald Teschl. *Ordinary differential equations and dynamical systems*. Vol. 140. American Mathematical Soc., 2012.

[34] Haruo Yoshida. "Construction of higher order symplectic integrators". In: *Physics Letters A* 150.5 (1990), pp. 262–268. ISSN: 0375-9601. DOI: https://doi.org/10.1016/0375-9601(90)90092-3. URL: https://www.sciencedirect.com/science/article/pii/0375960190900923.

[35] Haruo Yoshida. "Recent progress in the theory and application of symplectic integrators". In: *Celestial Mechanics and Dynamical Astronomy* 56.1-2 (Jan. 1993), pp. 27–43. DOI: 10.1007/bf00699717. URL: https://doi.org/10.1007/bf00699717.

[36] Ge Zhong and Jerrold E. Marsden. "Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators". In: *Physics Letters A* 133.3 (1988), pp. 134–139. ISSN: 0375-9601. DOI: https://doi.org/10.1016/0375-9601(88)90773-6. URL: https://www.sciencedirect.com/science/article/pii/0375960188907736.

# Appendix A

# Runge-Kutta integrators

## A.1 Definition and examples

In this thesis, we will study a specific class of numerical integrators that can be used for ODEs and PDEs. One classical family of such integrators consists of the Runge-Kutta methods. We briefly recall their definition.

**Definition A.1.1**
Consider a system of $n$ ODEs $U'(t) = f(t, U(t))$. with initial value $U(0) = u_0$ in $\mathbb{R}^n$. Then for $m \in \mathbb{N}$, an $m$-stage **Runge-Kutta integrator** is defined by a matrix

$$\boldsymbol{M_m} := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \\ b_1 & b_2 & \cdots & b_m \end{pmatrix} \in \mathbb{R}^{m+1 \times m} \tag{A.1}$$

in the following way. Let $c_i$ denote the sum of $\boldsymbol{M_m}$'s $i$-th row for $1 \leq i \leq m$. Then the numerical scheme is given by

$$\begin{cases} u_k = u_{k-1} + \sum_{j=1}^{m} b_j q_j \\ q_i = hf\left(t_{k-1} + c_i h, u_{k-1} + \sum_{j=1}^{m} a_{ij} q_j\right) \text{ for } 1 \leq i \leq m \end{cases}$$

where $q_i = q_{i,k}$ depends on the current step of the algorithm and $h > 0$ is the so-called step size. The algorithm starts from the initial value $u_0$, i.e. $u_k = u_0$ for $k = 0$. Moreover, a Runge-Kutta method is called **explicit** if $\boldsymbol{M_m}$'s first $m$ rows form a lower triangular matrix. Otherwise, the method is called **implicit**.

The forward Euler method, also known as the explicit Euler method, is perhaps the simplest example of a (first order) Runge-Kutta integrator. It is defined by the matrix

$$\boldsymbol{M_1} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and the iterations can be computed by $u_k = u_{k-1} + hf(t_{k-1}, u_{k-1})$. Another example

is the *classical* Runge Kutta method of fourth order, given by the matrix

$$\boldsymbol{M}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{pmatrix}. \tag{A.2}$$

## A.2 Order conditions

In a similar fashion to the section on higher order splitting integrators, we can derive order conditions for Runge-Kutta integrators by comparing coefficients of the relevant Taylor expansions, see [5, section 23]. For an $s$-stage integrator, this results in the **first order** condition

$$\sum_{i=1}^{s} b_i = 1, \tag{A.3}$$

the **second order** conditions

$$\begin{cases} (\text{A.3}) \\ \sum_{j=1}^{s} b_j c_j = \frac{1}{2}, \end{cases} \tag{A.4}$$

the **third order** conditions

$$\begin{cases} (\text{A.4}) \\ \sum_{j=1}^{s} b_j c_j^2 = \frac{1}{3} \\ \sum_{j,k=1}^{s} b_j a_{jk} c_k = \frac{1}{6} \end{cases}$$

and so on for $s > 3$. Clearly, as the required order $p$ increases, so do the amount of order conditions required to attain it. In [5], other strategies to formulate these order methods are also presented.

# Appendix B

# Code chapter 2

Every program that was written for this thesis can be found on my GitHub page [19], along with an explanatory `readme` file. In this appendix, our MATLAB implementation of the four integrators used in section 2.5 can be found.

## B.1 forwardEuler.m

```
function approx = forwardEuler(data)
% This function computes the forward (explicit) Euler
   approximations for a
% given ODE. It takes the following input:
% -> data: a data struct containing the relevant data of
   the problem;
% and it produces the following output values:
% -> approx: a dim x N+1 matrix containing the
   approximations for each
%             variable, row-wise.

f = data.f;

% Matrix containing the approximations, column-wise
% -> rows 1 until dim/2: approximations for the q-
   variable
% -> rows dim/2 + 1 until dim: approximations for the p-
   variable
approx      = sparse(data.dim, data.N+1);
approx(:,1) = data.initial;

% Implementing the integrator
for iteration = 2:data.N+1
    approx(:,iteration) = approx(:,iteration-1) + data.h*
       data.f(approx(:,iteration-1));
end

% For consistency: putting the rows in the right order,
   if needed
```

```
if data.dim == 4
    q2 = approx(3,:);
    p1 = approx(2,:);
    approx(2,:) = q2;
    approx(3,:) = p1;
end

end
```

## B.2 symplecticEuler.m

```
function approx = symplecticEuler(data)
% This function computes the symplectic Euler (VT)
   approximations for a
% given ODE. It takes the following input:
% -> data: a data struct containing the relevant data of
   the problem;
% and it produces the following output values:
% -> approx: a dim x N+1 matrix containing the
   approximations for each
%            variable, row-wise.

% Matrix containing the approximations, column-wise
% -> rows 1 until dim/2: approximations for the q-
   variable
% -> rows dim/2 + 1 until dim: approximations for the p-
   variable
approx      = sparse(data.dim, data.N + 1);
approx(:,1) = data.initial;

% Implementing the integrator
for iter = 2:data.N + 1
    approx(1:(0.5*data.dim), iter)             = approx
       (1:(0.5*data.dim), iter-1) + data.h * data.gradT(
       approx((0.5*data.dim + 1):data.dim, iter-1));
    approx((0.5*data.dim + 1):data.dim, iter) = approx
       ((0.5*data.dim + 1):data.dim, iter-1) - data.h *
       data.gradV(approx(1:(0.5*data.dim), iter));
end

end
```

## B.3 stormerVerlet.m

```
function approx = stormerVerlet(data)
% This function computes the Stormer-Verlet (VTV)
   approximations for a
% given ODE. It takes the following input:
% -> data: a data struct containing the relevant data of
   the problem;
% and it produces the following output values:
% -> approx: a dim x N+1 matrix containing the
   approximations for each
%            variable, row-wise.

% Matrix containing the approximations, column-wise
% -> rows 1 until dim/2: approximations for the q-
   variable
% -> rows dim/2 + 1 until dim: approximations for the p-
   variable
approx      = sparse(data.dim, data.N + 1);
approx(:,1) = data.initial;

% Implementing the integrator
for iter = 1:data.N

    % Half step
    p_half = approx((0.5*data.dim + 1):data.dim, iter) -
       0.5*data.h*data.gradV(approx(1:0.5*data.dim, iter)
       );

    % Updating position q
    approx(1:0.5*data.dim, iter + 1) = approx(1:0.5*data.
       dim, iter) + data.h*data.gradT(p_half);

    % Updating momenta p
    approx((0.5*data.dim+1):data.dim, iter + 1) = p_half
       - 0.5*data.h*data.gradV(approx(1:0.5*data.dim,
       iter + 1));

end

end
```

## B.4 yoshida4th.m

```
function approx = yoshida4th(data)
% This function computes Yoshida's 4th order integrator's
    approximations
% for a given ODE. It takes the following input:
% -> data: a data struct containing the relevant data of
   the problem;
% and it produces the following output values:
% -> approx: a dim x N+1 matrix containing the
   approximations for each
%            variable, row-wise.

% Two vectors containing the q- and p-approximations
q = zeros(0.5*data.dim, data.N+1);
p = zeros(0.5*data.dim, data.N+1);

% Initial data
q(:,1) = data.initial(1:0.5*data.dim);
p(:,1) = data.initial(0.5*data.dim+1:data.dim);

% 4th order coefficients
w1 = -nthroot(2,3) / (2 - nthroot(2,3));
w2 = 1 / (2 - nthroot(2,3));
c  = [0.5*w2, 0.5*(w1 + w2), 0.5*(w1 + w2), 0.5*w2];
d  = [w2, w1, w2];

% Integration
for iter = 1:data.N

    % First step: half for position, full for momentum
    Q1 = q(:, iter) + c(1) * data.gradT(p(:, iter)) *
       data.h;
    P1 = p(:, iter) - d(1) * data.gradV(Q1) * data.h;

    % Second step
    Q2 = Q1 + c(2) * data.gradT(P1) * data.h;
    P2 = P1 - d(2) * data.gradV(Q2) * data.h;

    % Third step
    Q3 = Q2 + c(3) * data.gradT(P2) * data.h;
    P3 = P2 - d(3) * data.gradV(Q3) * data.h;

    q(:, iter + 1) = Q3 + c(4) * data.gradT(P3) * data.h;
    p(:, iter + 1) = P3;
end

% Gathering the approximations
```

```
approx(1:0.5*data.dim, :)              = q;
approx(0.5*data.dim + 1:data.dim, :) = p;

end
```

# Appendix C

# Code chapter 3

Every program that was written for this thesis can be found on my GitHub page [19], along with an explanatory `readme` file. In this appendix, our MATLAB implementation of the method of characteristics used in section 3.5 can be found.

## C.1 solverMOC.m

```
function [Q, P, R] = solverMOC(data, int, T)
% This function numerically solves a linear, first order
   PDE using the
% method of characteristics; it is specifically
   implemented to solve the
% Liouville equation applied to a Hamiltonian system. It
   takes the input:
% -> data: a data struct containing all relevant
   information about the PDE,
%          underlying Hamiltonian system, etc;
% -> int: a string expressing which integrator is to be
   used to solve the
%         characteristic equations;
% -> T: a time at which we want to compute the MoC
   solution
%       (standardisation: -2*pi <= T <= 2*pi)
% and it produces the following output:
% -> Q, P, R: matrices representing the solution's values
    throughout time
%             (row-wise).

% Matrices that will contain the approximations
Q = sparse(length(data.sRange), length(data.tRange));
P = sparse(length(data.sRange), length(data.tRange));
R = sparse(length(data.sRange), length(data.tRange));

% Extracting variables from the data struct for
   legibility
s = data.sRange;
```

```
t = data.tRange; dt = t(2) - t(1);
data.h = dt;

% Method of characteristics
for iter = 1:length(data.sRange) % Picking a point on the
    initial curve

    % Computing the initial values
    s  = data.sRange(iter);
    q0 = data.q0(s);
    p0 = data.p0(s);
    data.initial = [q0; p0];

    % Solving the resulting system of ODEs by the chosen
       integrator for
    % chosen s on the initial curve
    if int == "fe"
        sol = forwardEuler(data);
    elseif int == "se"
        sol = symplecticEuler(data);
    elseif int == "sv"
        sol = stormerVerlet(data);
    elseif int == "y4"
        sol = yoshida4th(data);
    end

    % Saving the approximations in the matrices
    Q(iter,:) = sol(1,:);
    P(iter,:) = sol(2,:);
end

% Computation of the solution at selected time T
tPlot = floor(T/dt);
for t_iter = floor((length(data.tRange)/2)+1):length(data
   .tRange)
    for s_iter = 1:length(data.sRange)
        R(t_iter, s_iter) = data.rho0(Q(s_iter, t_iter -
           tPlot), P(s_iter, t_iter - tPlot));
    end
end

end
```