ISBN 978-82-326-6865-6 (printed ver.)
ISBN 978-82-326-6657-7 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (electronic ver.)

NTNU
Norwegian University of
Science and Technology

Doctoral thesis

Benjamin Tapley

Doctoral theses at NTNU, 2021:189

NTNU
Norwegian University of
Science and Technology
Thesis for the degree of
Philosophiae Doctor
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU

Doctoral theses at NTNU, 2021:189

Benjamin Tapley

Structure-preserving numerical methods with application to problems in particle dynamics

NTNU
Norwegian University of
Science and Technology

Benjamin Tapley

Structure-preserving numerical methods with application to problems in particle dynamics

Thesis for the degree of Philosophiae Doctor

Trondheim, June 2021

Norwegian University of Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

© Benjamin Tapley

ISBN 978-82-326-6865-6 (printed ver.)
ISBN 978-82-326-6657-7 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (electronic ver.)

Doctoral theses at NTNU, 2021:189

PRINTED IN
NORWAY
NO - 1598

NORDIC SWAN ECOLABEL
Printed matter
2041 0731

Printed by Skipnes Kommunikasjon AS

Structure-preserving numerical methods with application to problems in particle dynamics

Benjamin K. Tapley

14 June 2021

Preface

This thesis is a compilation of projects that I have contributed to over the duration of my four-year PhD at NTNU. I have been fortunate enough to have spent 16 months of my PhD undertaking research trips and collaborating with mathematicians from around the world and therefore have many people to thank.

First and foremost, I am very grateful to my supervisor Elena Celledoni for sharing her broad expertise with me, her constant support and our countless discussions. The same can be said about my co-supervisors Brynjulf Owren and Helge Andresson whom have also contributed significantly towards this work and have been a strong source of encouragement. In addition, I am extremely grateful for all the opportunities that Elena and Brynjulf have extended to me including invitations for research stays abroad, conferences and cabin trips.

I would like to extend my gratitude to my many co-authors and colleagues whom I have shared many interesting discussions with and learned an immeasurable amount from including Reinout Quispel, Dave McLaren, Pambos Evripidou, Peter Van der Kamp, Christian Offen, Robert McLachlan and Laurel Ohm. In particular, I am very appreciative of Robert McLachlan and Carola Schönlieb for their hospitality during my visits to the University of Massey in 2018 and the University of Cambridge in 2017. A special thank you goes to Reinout Quispel, not only for his generous hospitality during my trips to the University of La Trobe in 2018 and 2020, but also for our productive and enjoyable collaboration over the last three years, which has significantly shaped this thesis.

I also acknowledge the European Union Horizons 2020 project (CHiPS) that has funded my research trips to the University of Cambridge, La Trobe University, Massey University as well as to five international conferences and more local conferences. I would like to acknowledge the Isaac Newton Institute for Mathematical Sciences of the University of Cambridge, for support and hospitality during the programme Geometry, compatibility and structure preservation in computational differential equations in 2019, as well as Darwin College for accepting me as a visiting student during my stay in Cambridge in 2017.

I'd also like to thank my friends from NTNU, including those from the mathematics department, ISFiT, UKA, NTNUI Svømmegruppa, Propulse, my housemates and otherwise, who have all made my time in Trondheim as memorable as it has been.

Lastly and most importantly are my mum, dad and brother, whom without their love and support throughout my 26 years, I would not be where I am today.

Ben Tapley
Trondheim, 14 June, 2021

Contents

Introduction

1.1 Structure-preserving numerical integration

Since the discovery of calculus in the 1700s, it has long been understood that differential equations govern almost everything that happens in our universe. From biology, cosmology, finance and engineering, knowing how to solve differential equations is paramount to mankind's ability to understand and control the world around us. As most differential equations that are relevant to real-world applications are not easily amenable to analytic techniques, one must look for numerical methods to calculate fast, accurate and reliable solutions. Since the birth of commercial computers in the late 20th century research into numerical methods for ordinary differential equations (ODEs), to which we will now limit our discussion to, has exploded. Some of the most successful numerical methods for solving ODEs are the Runge-Kutta and linear multi-step methods and there have been thousands of researchers whom have dedicated their careers to advance our collective knowledge in these areas. Such methods are efficient, accurate and can be applied to a general initial value problem (allowing for some regularity assumptions [16]) of the form

$$\dot{x}(t) = f(x(t)) \in \mathbb{R}^n, \quad x(0) = x_0. \tag{1.1.1}$$

The word *general* is an important distinction here and we will call a numerical method that can be applied to such an ODE a *general-purpose* method. When designing general-purpose methods one is often concerned with reducing the local and global error, which given some time h, is usually defined as the distance (in some norm) between the numerical solution $\Phi_h(x_0)$ and the exact solution $x(h)$. That is, after one step

$$\text{local error} = \|\Phi_h(x_0) - x(h)\|, \tag{1.1.2}$$

or after n steps

$$\text{global error} = \|\Phi_h^n(x_0) - x(nh)\|. \tag{1.1.3}$$

The theory of general-purpose numerical methods has been extremely successful and, due to the immense amount of research on this topic over the last 50 years, is nearing a certain level of maturity. More recently, however, there has been a growing interest in studying the numerical solution of ODEs with a prescribed structure. By *structure*, we mean a particular property of the ODE that relates to a feature of the exact solution. Some examples of common ODEs with structure are as follows.

1. Hamiltonian ODEs:
$$\dot{x}(t) = J^{-1}\nabla H(x(t)), \tag{1.1.4}$$
where J is constant and skew-symmetric and for the Hamiltonian function $H : \mathbb{R}^{2n} \to \mathbb{R}$.

2. Volume preserving ODEs:
$$\dot{x}(t) = f(x(t)) \tag{1.1.5}$$
where $\nabla \cdot f(x(t)) = 0$.

3. Contractive ODEs:
$$\dot{x}(t) = f(x(t)) - Ax(t) \tag{1.1.6}$$
where $\nabla \cdot f(x(t)) = 0$ and A is positive semi-definite.

4. ODEs with multiple first integrals, e.g., in $n = 3$ dimensions the Nambu system:
$$\dot{x}(t) = \nabla H(x(t)) \times \nabla K(x(t)), \tag{1.1.7}$$
where $H, K : \mathbb{R}^n \to \mathbb{R}$.

While it is often an impossible task to find the exact solution $x(t)$ that solves the above ODEs, knowing their structure can reveal a number of qualitative features that pertain to the exact solution. For the above examples, those features are as follows.

1. Hamiltonian ODEs: The trajectory of the ODE preserves the Hamiltonian function and the symplectic 2-form , that is
$$\frac{\mathrm{d}}{\mathrm{d}t}H(x(t)) = 0 \quad \text{and} \quad \omega = \sum \mathrm{d}p_i \wedge \mathrm{d}q_i \tag{1.1.8}$$
is preserved along the trajectory of (1.1.4) [27,45].

2. Volume preserving ODEs: The exact solution $x(t)$ satisfies
$$\det\left(\frac{\partial x(t)}{\partial x_0}\right) = 1 \tag{1.1.9}$$
that is, volumes in phase space remain constant along the flow of (1.1.5) [25,55].

3. Contractive ODEs: The exact solution $x(t)$ satisfies

$$\det\left(\frac{\partial x(t)}{\partial x_0}\right) = \mathrm{e}^{-\mathrm{Tr}(A)\,t} \tag{1.1.10}$$

   that is, volumes in phase space monotonically contract along the flow of (1.1.6) [21,32].

4. The Nambu system: The functions (first integrals) $H(x(t))$ and $K(x(t))$ are preserved along the trajectory of (1.1.7) [47], that is

$$\frac{\mathrm{d}}{\mathrm{d}t}H(x(t)) = 0 \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}t}K(x(t)) = 0. \tag{1.1.11}$$

In addition to the above examples, there are other classes of ODEs of the form $\dot{x}(t) = f(x(t))$ whose structure is more subtle in the sense that one cannot generally write the vector field $f(x(t))$ in a way that elucidates the geometric properties of its solution. Examples of two such classes of ODEs that are relevant to this thesis are as follows.

1. Measure preserving ODEs: Where one or more measures of the form

$$\int \frac{\mathrm{d}x_1 \wedge \ldots \wedge \mathrm{d}x_n}{m(x)} \tag{1.1.12}$$

   are preserved along the flow, for some function $m : \mathbb{R}^n \to \mathbb{R}$.

2. ODEs with one or more Darboux polynomials: Where one or more pairs of polynomial functions $p(x(t))$ and $c(x(t))$ exist satisfying

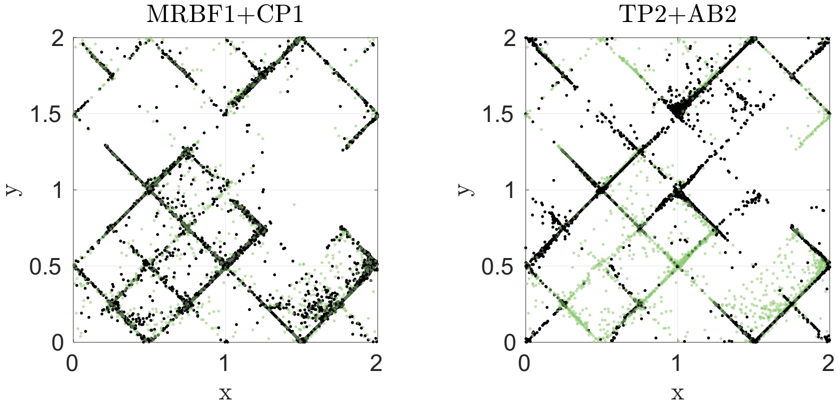
$$\dot{p}(x(t)) = c(x(t))p(x(t)). \tag{1.1.13}$$

   In this case, $p(x(t))$ is preserved along their zero level sets i.e., when $p(x(0)) = 0$ [20].

Generally, when general-purpose numerical methods are applied to one of the above ODEs the geometric features of the exact solution (e.g., (1.1.8) - (1.1.13)) are not preserved in the numerical solution. The non-conservation of such properties becomes an issue especially when integrating ODEs over long times. This is because these geometric properties usually have a clear physical meaning that is important for the numerical solution to inherit. For this reason "geometric" and "structure-preserving" are used synonymously when describing numerical integration methods. We will now illustrate the application of geometric integration methods with two examples from the thesis.

1.1.1 Example 1: Geometric integration of particle suspensions

The first example is presented in figure 1.1, which shows the final positions of 10^4 small spheroidal particles having evolved in a viscous cellular flow field for six of seconds of simulation time. In figure 1.1a we have used a geometric method called MRBF1+CP1 to calculate the particle positions, which are represented by black dots and figure 1.1b uses a higher order general-purpose method to do the same. In both figures, the green dots represent the "exact" solution. The method MRBF1+CP1 was purpose-built for the particular equations that govern the dynamics of these small particles, which are of the form (1.1.6). By this, we mean that the numerical solution replicates a number of physical features that the exact solution possesses, such as the constant contractivity of phase space volume from equation (1.1.10), among others. These features are not present in the numerical solution of the general-purpose method shown in figure 1.1b and the consequence of this is that the particles erroneously cluster in regions where they shouldn't as seen by the mismatch of green and black dots. Of increased interest is the fact that the general-purpose method is order-two and is more costly than the geometric method, which is order-one and faster. This is a perfect example where only focusing on reducing the error in a conventional sense (e.g., equations (1.1.2) and (1.1.3)) is less important than preserving important physical properties that pertain to the exact solution as we do in geometric numerical integration.



(a) Geometric method

(b) General-purpose method

Figure 1.1: A comparison of a geometric method versus a general-purpose method for computing the spatial distribution of particles in a viscous flow field. The green dots represent the positions of the particles of the "exact" solution and the black dots are calculated by the geometric method (figure (a)) and a higher-order general-purpose method (figure (b)).

1.1.2 Example 2: Geometric integration of a free rigid-body

In our second example, we consider the numerical integration of the free rigid-body equations (also known as the Euler top),

$$\dot{y}_1 = \alpha_1 y_2 y_3, \tag{1.1.14}$$

$$\dot{y}_2 = \alpha_2 y_3 y_1, \tag{1.1.15}$$

$$\dot{y}_3 = \alpha_3 y_1 y_2, \tag{1.1.16}$$

with $\alpha_1 = -1/2$, $\alpha_2 = -1/3$, $\alpha_3 = 5/6$ and for initial conditions satisfying $\|\mathbf{y}(0)\| = 1$. Here, the vector $\mathbf{y}$ represents the angular momentum of a freely spinning rigid-body. We note that this system can be formulated as a Nambu system (1.1.7) and possesses the following first integrals

$$H(\mathbf{y}) = \frac{(\alpha_3 - \alpha_2)\, {y_1}^2}{\alpha_1} + \frac{(\alpha_1 - \alpha_3)\, {y_2}^2}{\alpha_2} + \frac{(\alpha_2 - \alpha_1)\, {y_3}^2}{\alpha_3}, \tag{1.1.17}$$

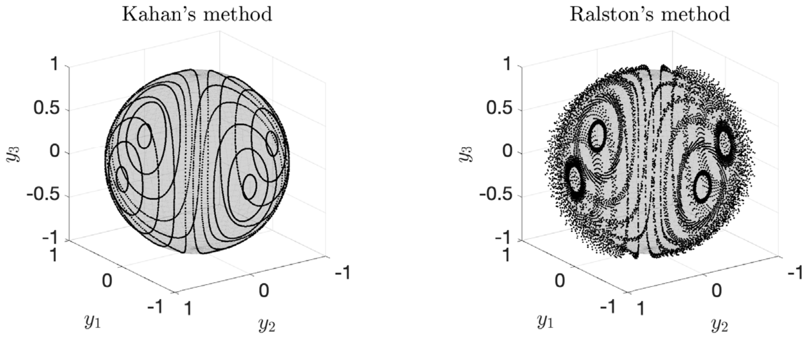
$$K(\mathbf{y}) = y_1^2 + y_2^2 + y_3^2, \tag{1.1.18}$$

which correspond to the conservation of kinetic energy and momentum, respectively. To the above system we first apply Kahan's method (which is introduced in section 1.3.4) then Ralston's method, which are both second-order explicit Runge-Kutta methods for this system. Figure 1.2a presents the solution of Kahan's method. We observe here that the orbits remain closed and very close to the isosurface $H(\mathbf{y}) = 1$. This remarkable property is explained by the fact that the Kahan map possesses the following rational integrals when applied to this system

$$H_h(\mathbf{y}) = \frac{H(\mathbf{y})}{1 - \alpha_3 \alpha_2 h^2 {y_1}^2/4}, \tag{1.1.19}$$

$$K_h(\mathbf{y}) = \frac{K(\mathbf{y})}{1 - \alpha_3 \alpha_2 h^2 {y_1}^2/4}, \tag{1.1.20}$$

which depend on the time-step h. These integrals were derived using the method of discrete Darboux polynomials, which is discussed in this thesis and introduced in section 1.3. In other words, the Kahan map is an integrable map and preserves nearby modified integrals of the continuous system, which explains its favorable solution compared to the Ralston method, which does not preserve any integrals. The non-preservation of these integrals result in a numerical solution that drifts off the isosurface $K(\mathbf{y}) = 1$, as seen in figure 1.2b. Such a solution becomes increasingly non-physical as $t \to \infty$. For a comprehensive overview of geometric numerical integration we refer to the books [5, 15] among the extensive literature.



(a) Geometric method

(b) General-purpose method

Figure 1.2: The numerical solution of the rigid body equations for 500 steps using a step size of $h = 1/2$. The grey surface denotes $K(\mathbf{y}) = 1$.

1.1.3 Outline of thesis

This thesis is composed of a number of chapters spanning multiple topics in the broad field of geometric integration and numerical analysis. Each chapter is based on an article that is either published or in the submission process. The exception is chapter 8 which is based on an unsubmitted pre-print.

The thesis begins with two chapters that are based on geometric numerical methods for one-way coupled particle suspensions with Stokes drag force. Chapter 2 is based on a paper [48] that develops a basic splitting method for a particular spheroidal particle model. In chapter 3, we propose a geometric method that is designed to preserve a number of properties of the exact solution for particle suspensions and is based on the previous chapter. The algorithm is implemented for spherical and spheroidal particle models and give excellent results when calculating spatial distributions of particle suspensions [49]. These methods are applicable to other particle shapes within the same setting, for example, slender rigid particles, given that one can calculate the forces and torques on the particle. This brings us to the next two chapters, which were written in collaboration with Dr Laurel Ohm, now at the Courant Institute, New York. The first of these two chapters focus on a validation and comparisons of a new particle model inspired by slender body theories due to Ohm [37, 38]. In the sequel, we further develop the model to make it suitable to numerical implementation and inversion. Having done so, we perform numerical tests and experiments on the model and propose an algorithm based on linear algebra and quadrature for computing the forces and torques on slender rigid particles [1]. This chapter concludes our discussion of particles in Stokes flow.

The next three chapters lie within the related fields of integrable systems and geometric numerical methods. Here, we introduce the idea of discrete Darboux polynomials. These articles were written with the group at La Trobe University, Melbourne, lead by Prof Reinout Quispel whom conceived the idea. In the first of these three chapters, the idea of discrete Darboux polynomials for rational maps are introduced and a systematic algorithm for deriving rational integrals of a map is presented [8]. In the second article, the method is elaborated on and implemented in a number of novel examples, where we uncover many interesting properties of some well known integrable maps [7]. The third chapter is based on a preprint and extends the theory of discrete Darboux polynomials to Runge-Kutta maps.

The final chapter in this thesis lies within an adjacent area of geometric integration and was written with Dr Christian Offen and Prof Robert McLachlan, at Massey University, New Zealand. Here, we investigate the use of symplectic integration for the numerical solution of Lie-Poisson PDEs when written in their Clebsch variables. That is, when reformulated as a Hamiltonian system on a symplectic manifold. We focus our numerical examples from fluid dynamics, namely the Burgers equation and related PDEs [31].

The remainder of the introduction will serve to introduce the topics of particles in Stokes flow and discrete Darboux polynomials.

1.2 Small particles in viscous flow

One of the central topics to this thesis is numerical methods for calculating the dynamics of particle suspensions in viscous flows. Particles immersed in viscous flow could be interpreted as, for example, paper fibers [19], biopolymers [17], soot [42], plankton [50], ice crystals [22], pollen [26], microcontaminants [13] or indoor pollutants [35]. Accurately computing the dynamics of fluid-structure interactions is a difficult and computationally demanding task (e.g., using the boundary integral method [41, 57]) so to be able to simulate suspensions of hundreds of thousands of particles, one needs to lay out a few assumptions to derive a model for the forces and torques that is computationally tractable. To this end, we will consider particles of size less than the smallest fluid length scale (e.g., the Kolmogorov scale for turbulent flows). Under this assumption, the local flow around the particle is Stokesian to good approximation and it is justified to assume that the particles do not influence the surrounding fluid field. We also assume that the particle suspension is dilute enough that particle-particle collisions are infrequent and therefore ignored and that the particles are heavy enough that Brownian motion is negligible. This is called a one-way coupled system and has been a topic of interest for many influential studies, including [4, 12, 14, 18, 28, 34, 39, 40, 44, 46, 51–53] and references therein.

When considering the literature in particle-laden flows over the past four decades, there has been an astounding number of studies dedicated to developing fast and accurate numerical methods for the direct numerical simulation of the Navier-Stokes equations. This has resulted in advanced and continually improving software and databases [56] for this purpose. However, far less attention has been given towards the development of fast and accurate solvers for the equations of motion that govern the dynamics of the particulate phase. We note that of the aforementioned references use general-purpose methods for this purpose such as Runge-Kutta integration and standard polynomial interpolation. One of the topics in this thesis is to develop specialized methods for the equations of motion for the particulate phase.

In the remainder of this section we will outline the equations of motion and introduce the particle models that feature in the next four chapters of the thesis as well as briefly introduce splitting methods.

1.2.1 The equations of motion

The dynamics of such particles are governed by the rigid body equations with a hydrodynamic Stokes force and torque. The non-dimensionalised translational dynamics is given by

$$\dot{\mathbf{v}} = St^{-1}\mathbf{F}, \tag{1.2.1}$$

$$\dot{\mathbf{x}} = \mathbf{v}, \tag{1.2.2}$$

where $\mathbf{x}$ is the particle's location, $\mathbf{v}$ the velocity of its center of mass and St is the particle Stokes number, which is a dimensionless measure of the particle's relative inertia. Under our assumptions, the only relevant force is the hydrodynamic Stokes drag, which is linear in the slip velocity*. This is given by

$$\mathbf{F} = K(\mathbf{u}(\mathbf{x}) - \mathbf{v}), \tag{1.2.3}$$

where $\mathbf{u}(\mathbf{x})$ is the fluid velocity at the particle's location $\mathbf{x}$ and K is a positive definite resistance tensor that depends on the particle shape, which we will discuss in the next section. The angular velocity $\boldsymbol{\omega}$ evolves according to

$$J\dot{\boldsymbol{\omega}} = J\boldsymbol{\omega} \times \boldsymbol{\omega} - \mathbf{T}, \tag{1.2.4}$$

*The slip velocity is defined as the difference between the background fluid velocity and the particle velocity $\mathbf{u}(\mathbf{x}) - \mathbf{v}$

where J is the diagonal body frame moment of inertia tensor and $\mathbf{T}$ is the hydrodynamic torque. The rotation matrix $Q \in SO(3)$, which specifies the particle's orientation, transforms a vector in the body frame (i.e., a frame that is co-rotating and co-translating with the particle) to one in a co-translating frame and is calculated by solving the matrix ODE

$$\dot{Q} = Q\widehat{\boldsymbol{\omega}}, \tag{1.2.5}$$

where $\widehat{\cdot} : \mathbb{R}^3 \to \mathfrak{so}(3)$ is defined by

$$\begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \mapsto \widehat{\boldsymbol{\omega}} = \begin{pmatrix} 0 & -\omega_1 & \omega_2 \\ \omega_1 & 0 & -\omega_3 \\ -\omega_2 & \omega_3 & 0 \end{pmatrix}, \tag{1.2.6}$$

such that $\widehat{\boldsymbol{\omega}}\mathbf{v} = \boldsymbol{\omega} \times \mathbf{v}$.

1.2.2 Particle models

Spherical particles

The spherical particle model is understandably the simplest and can be described purely by the translational equations of motion (1.2.1) and (1.2.2) due to its isotropic shape. Maxey and Riley derived the equations of motion for a finite-sized sphere immersed in viscous flow [30]. For small particles in low Reynolds number flow, however, this expression simplifies to one of the form (1.2.3) with $K = I$. It is in this setting that the equations of motion are more suited to analytical methods such as calculating Lyapunov exponents [3], analyzing caustics [54] or perturbative methods [29], all of which shed light on physical features of the exact solution. For example, the centrifuge effect, which was first discovered by Maxey [29], tells us that particles disperse in regions of high fluid vorticity. It is due to features of the exact solution, like the centrifuge effect, that explain phenomena we observe in particle suspensions, such as the preferential concentration in turbulent flows [46]. As these collective properties are of interest to researchers studying the physics of particle-laden flows, it is also of interest to preserve these features in the numerical solution. This will be addressed in chapter 3.

Spheroidal particles

The surface of a spheroid (i.e., an ellipsoid with one rotational axis of symmetry) is defined by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} + \frac{z^2}{c^2} = 1, \tag{1.2.7}$$

where a and c are the distinct semi-axis lengths. The spheroid's shape can be uniquely characterized by the dimensionless aspect ratio $\lambda = c/a > 0$, which distinguishes between spherical ($\lambda = 1$), prolate ($\lambda > 1$) and oblate ($\lambda < 1$) particles (the latter two shapes are also called as rods and disks). Like the spherical case, the geometry is relatively simple and there exist closed form expressions for the resistance tensor, where in a reference frame co-translating with the particle, is given by $K = Q^T K_b Q$ for the diagonal body frame resistance tensor K_b given by Brenner [6] and Oberbeck [36]. Due to the anisotropy of the particle, one must now keep track of the particle's rotational variables by solving equations (1.2.4) and (1.2.5). Here, a closed form expression for the torque vector $\mathbf{T}$ is given by Jeffery [23]. As the equations of motion for spheroidal particles in turbulence are more complex than the spherical case, there has been less theoretical studies on their collective behavior in complex flows. There have, however, been many numerical studies on the statistical behavior of these particles in a variety of flow fields, e.g., [12, 34]. None-the-less, the numerical methods we develop for the spherical particles are still applicable to non-spherical particles and are addressed in chapters 2 and 3.

Slender particles

Another particular particle shape that we are concerned with in this thesis are slender particles, i.e., particles with a very high aspect ratio. Denoting by ϵ the maximum radius of a particle and $2L$ its length, then a slender particle is characterized by $\epsilon/L \ll 1$. The forces and torques on such a particle can be modeled using a number of techniques, such as bead models [43] or if the particle is straight, the aforementioned spheroid model. Slender body theory (SBT), on the other hand, offers a more computationally efficient and accurate approach to these long fibers and is based on integrating fundamental solutions to the Stokes equations along the particle centerline and exploiting the small parameter ϵ. For a particle whose centerline is parametrised by a C^1 non-intersecting curve $\boldsymbol{X}(s) : [-L, L] \to \mathbb{R}^3$, SBT results in expressions for the force density $\mathbf{f}(s)$ of the following form

$$\mathbf{v} - \boldsymbol{X}(s) \times \boldsymbol{\omega} - \mathbf{u}(\boldsymbol{X}(s)) = \alpha I + \int_{-L}^{L} K_\epsilon(s, t)\mathbf{f}(t)\mathrm{d}t \tag{1.2.8}$$

where α is a regularization parameter and $K_\epsilon(s, t)$ is a shape-dependent integration kernel. Such equations, where the force density is the quantity of interest is called a Fredholm integral equation [2] and are typically ill-posed. In our application, the force and torque on the particle is required to determine the particle's dynamics. These are given by integrating the force density along the center line

$$\mathbf{F} = \int_{-L}^{L} \mathbf{f}(s)\mathrm{d}s \quad \text{and} \quad \mathbf{T} = \int_{-L}^{L} \boldsymbol{X}(s) \times \mathbf{f}(s)\mathrm{d}s. \tag{1.2.9}$$

By numerically inverting the integral operator in equation (1.2.8) and solving for **F** and **T** we can derive expressions for the resistance tensor K and similarly for the torque that are linear in the slip velocity. Due to this, the dynamical equations are amenable to the same numerical methods proposed in chapters 2 and 3. In chapters 4 and 5 we consider these slender particles. Inspired by the work of Ohm [37], an integral model is proposed and its computation, validation and comparisons to other models are studied as well as an algorithm for computing its dynamics.

1.2.3 Splitting methods for particles

Splitting methods are the central topic of chapters 2 and 3 and are also applied in 4 and 5. A splitting method is a numerical method for integrating an ODE that can be decomposed into a sum of two or more integrable vector fields, for example

$$\dot{x} = f(x) + g(x) \in \mathbb{R}^n. \tag{1.2.10}$$

Denoting by $\phi_h^{[f]}(x_0)$ and $\phi_h^{[g]}(x_0)$ the exact flows of the vector fields $f(x)$ and $g(x)$, respectively, then one can then create a numerical approximation $x_1 \approx x(h)$ to the exact solution at time h by computing alternating compositions of $\phi_h^{[f]}$ and $\phi_h^{[g]}$. For example the first order Lie-Trotter splitting method

$$x_1 = \phi_h^{[f]} \circ \phi_h^{[g]}(x_0) \tag{1.2.11}$$

or the second order Strang splitting method

$$x_1 = \phi_{\frac{h}{2}}^{[f]} \circ \phi_h^{[g]} \circ \phi_{\frac{h}{2}}^{[f]}(x_0). \tag{1.2.12}$$

Splitting methods are particularly favorable when they can be applied due to the fact they are explicit and can be used to design structure-preserving methods. In our application, we also observe better stability compared to other explicit methods of equal order. A good introduction to splitting methods is found in [33].

The ODEs for particles with Stokes drag can be concisely written as a the following dissipative ODE (i.e., similar to the form (1.1.6))

$$\dot{y} = f(y) - St^{-1}\left(Ay + b\right) \tag{1.2.13}$$

Where $f(y)$ here represents the free rigid-body ODEs and $Ay + b$ represents the Stokes force and torque. These equations have a natural splitting into the following two subsystems

$$\dot{y} = f(y), \quad \text{and} \quad \dot{y} = St^{-1}\left(Ay + b\right). \tag{1.2.14}$$

The vector field $f(y)$ represents the free rigid body equations, whose flow is known exactly [9] and the other vector field is affine and can therefore be computed by the variation of parameters formula. This particular splitting method is developed and analysed in chapters 2 and 3.

1.3 The method of Discrete Darboux polynomials

Understanding the integrability properties of ODEs has long been an area of interest in the mathematical sciences. Darboux polynomials, also known as second integrals or weak integrals, have been a successful tool for studying integrable systems, especially for those with one or more rational integrals. In chapters 6, 7 and 8, we depart our discussion of small particle dynamics and instead focus our attention on *discrete* Darboux polynomials, which lies at the intersection of geometric numerical integration and discrete integrable systems. One can always view a numerical method applied to an ODE as a discrete map. However, understanding the integrability properties (e.g., preserved integrals or measures) of discrete maps is not always straight forward, yet it is important for designing structure-preserving methods. As we have seen in example 2, one can explain interesting properties of the Kahan map when applied to free rigid-body ODEs by knowing its preserved integrals.

The section begins by briefly introducing Darboux polynomials for ODEs then discrete Darboux polynomials for maps. We then present two more examples of the use of discrete Darboux polynomials, one with the forward Euler method and one with the Kahan map.

1.3.1 Darboux polynomials (ODE case)

A Darboux polynomial for a polynomial ODE $\dot{x} = f(x) \in \mathbb{R}^n$ is a function $P(x)$ that satisfies

$$\dot{P}(x) = C(x)P(x) \tag{1.3.1}$$

where the polynomial $C(x)$ is called the cofactor of $P(x)$. Darboux polynomials are important because they can give insight into the structural properties of ODEs that are otherwise difficult to find. For example, given an ODE, it is not an easy task to know if it has a preserved quantity or not. However, if one can find two Darboux polynomials, say $P_1(x)$ and $P_2(x)$ that correspond to the cofactors $C_1(x)$ and $C_2(x)$, then it is easy to show that $P_1(x)^{\alpha_1}P_2(x)^{\alpha_2}$ is a Darboux polynomial with cofactor $\alpha_1C_1(x) + \alpha_2C_2(x)$. Moreover, if

$$\alpha_1C_1(x) + \alpha_2C_2(x) = 0 \tag{1.3.2}$$

then

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(P_1(x)^{\alpha_1}P_2(x)^{\alpha_2}\right)=0 \tag{1.3.3}$$

defines a first integral. So the problem of finding first integrals can be made simpler by finding second integrals. A good introduction to Darboux polynomials is found in [20].

1.3.2 Discrete Darboux Polynomials (mapping case)

In chapters 6, 7 and 8 we study *discrete* Darboux polynomials, which, as the name suggests, is a discrete analogue of equation (1.3.1) for a rational map $\Phi_h(x)$, defined by the following

$$p(\Phi_h(x))=c(x)p(x), \tag{1.3.4}$$

where $p(x)$ is polynomial and $c(x)$ is rational. Similarly to the continuous case, discrete Darboux polynomials can be used to find integrals of the map Φ_h by recognizing that $p_1(x)$ and $p_2(x)$ are discrete Darboux polynomials with cofactors $c_1(x)$ and $c_2(x)$ then $p_1(x)^{\alpha_1}p_2(x)^{\alpha_2}$ is also a discrete Darboux polynomial with cofactor $c_1(x)^{\alpha_1}c_2(x)^{\alpha_2}$. Moreover, if

$$c_1(x)^{\alpha_1}c_2(x)^{\alpha_2}=1 \tag{1.3.5}$$

is satisfied then

$$p_1(x)^{\alpha_1}p_2(x)^{\alpha_2} \tag{1.3.6}$$

defines a (possibly rational or even irrational) first integral of the map $\Phi_h(x)$. A systematic method of detecting and determining integrals of mappings and ordinary difference equations has long been an difficult problem. The method of discrete Darboux polynomials is a step towards addressing this.

1.3.3 An example

Consider the following ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1\\ x_2\\ x_3\\ x_4\end{pmatrix}=\begin{pmatrix} x_2{}^2+\left(x_1+2\,x_4+1\right)x_2+x_4{}^2+x_4\\ -2\,x_1x_2+x_3+x_4\\ \left(-2\,x_1-1\right)x_2+(\sigma+1)\,x_3+\sigma\,x_4\\ 2\,x_1x_2+x_2-x_3\end{pmatrix} \tag{1.3.7}$$

where $\sigma\in\mathbb{R}$.

Now consider integrating the above system with the forward Euler method. This method is arguably the most well-known method in numerical analysis, for example, it is a Runge-Kutta method, affinely equivariant and preserves linear integrals. However, when an ODE possesses non-linear integrals, it is not always known when the forward Euler method can preserve these integrals. However, using the method of discrete Darboux polynomials we can show that in some cases, it can preserve a nearby modified integral.

Denote by $\phi_h : \mathbb{R}^4 \to \mathbb{R}^4$ the map defined by the forward Euler discretisation of the above ODE. Using our method we find that ϕ_h has the following discrete Darboux polynomials $p_i(x)$ that correspond to the discrete cofactor $c_i(x)$

	$p_i(x)$	$c_i(x)$
$i=1$	x_3+x_4	$1+\sigma h$
$i=2$	x_2+x_4	$1+h$

(1.3.8)

Using equation (1.3.5) we see that

$$H_h(x) = \frac{(x_2+x_4)^{\sigma_h}}{x_3+x_4} \tag{1.3.9}$$

defines an h-dependent integral of ϕ_h with

$$\sigma_h = \frac{\ln\left(1+\sigma h\right)}{\ln\left(1+h\right)}. \tag{1.3.10}$$

That is, $H_h(\phi_h(x)) = H_h(x)$. In fact, as we will show in chapter 8, it turns out that all Runge-Kutta methods possess a similar integral but with a different exponent σ_h and gives rise due to the existance of iteration-index-dependent integrals. As the forward Euler method approaches the exact solution as $h \to 0$, we can take the continuum limit

$$\lim_{h\to 0}\left(H_h(x)\right) = H(x) := \frac{\left(x_2+x_4\right)^{\sigma}}{x_3+x_4}. \tag{1.3.11}$$

Indeed we find that the original ODE does possess the non-rational integral $H(x)$. In other words, the forward Euler method doesn't preserve the non-rational integral of the ODE exactly, but instead preserves the integral with a modified exponent.

1.3.4 Kahan's method

When applied to an ODE with that possesses one or more preserved quantities, (e.g., those of the form (1.1.4) or (1.1.7)), a preserved measure (1.1.12) or a Darboux polynomial (1.1.13) it is rare that a general-purpose numerical method shares these features. There are of course exceptions, including linear integrals, for example. Remarkably, however, Kahan's method has been shown to preserve many interesting geometric features of the exact solution when applied to quadratic ODEs:

$$\dot{x}_i = \sum_{j,k} a_{i,j,k} x_j x_k + \sum_j b_{i,j} x_j + c_i. \tag{1.3.12}$$

Kahan introduced his method in 1993 [24] and is defined by the map $\phi_h : x \mapsto \phi_h(x) = x'$ where x' is given by the following linearly implicit equation

$$\frac{x'_i - x_i}{h} = \sum_{j,k} a_{i,j,k} \frac{x'_j x_k - x_j x'_k}{2} + \sum_j b_{i,j} \frac{x'_j - x_j}{2} + c_i. \tag{1.3.13}$$

Celledoni et al. [10, 11] have studied this map extensively and have shown that it can preserve certain integrability and geometric properties of original ODE, often by showing the existence of preserved *modified* integrals or measures. As we will show in chapters 6, 7 and 8 the Kahan map also preserves certain (often modified) measures and Darboux polynomials of the ODE under study. This can be used to uncover new properties of the Kahan map.

We illustrate this with an example. Consider the Kahan discretisation of the following family of Nambu systems in the variables $\mathbf{x} = (x, y, z)^T$

$$\dot{\mathbf{x}} = c\left(\nabla H \times \nabla K_\alpha\right), \tag{1.3.14}$$

where $c = y^{2-\alpha}$, $H = \frac{x}{y}$, $K_\alpha = y^\alpha Q(\mathbf{x})$, $Q(\mathbf{x})$ is an arbitrary homogeneous and quadratic polynomial in $\mathbf{x}$ and $\alpha \in \mathbb{R}$ is a free parameter. Note that vector field is scaled by the factor c to make it quadratic and therefore applicable to the Kahan map. Using our method, we can show that the Kahan map preserves the rational integral H exactly, due to x and y being discrete Darboux polynomials of the same cofactor. However, what about the integral K_α? This integral is proportional to y^α which makes K_α generally non-rational. However, our algorithm can also detect for which values of α does the Kahan map preserve the integral K_α. This gives us the following solutions for α and the corresponding additional second integral of the Kahan discretisation summarized in the following table

α	Integrals
-2	H and K_{-2}
-1	H and K_{-1}
0	H and $\tilde{K}_0$
1	H and $\tilde{K}_1$
2	H and $\tilde{K}_2$

So when $\alpha = -2, -1$, (i.e., for rational K) the Kahan map preserves both integrals exactly, whereas for $\alpha = 0, 1, 2$, the Kahan map preserves a modified integral of the form

$$\tilde{K}_i = \frac{K}{1 + O(h^2)}. \tag{1.3.15}$$

One can also show that the Kahan map also preserves a measure (i.e., of the form equation (1.1.12)) when applied to this ODE. It is straight forward to show that a preserved measure corresponds to a Darboux polynomial whose cofactor is the Jacobian determinant of the map. That is,

$$m(\Phi_h(\mathbf{x})) = \det\left(\frac{\partial \Phi_h(\mathbf{x})}{\partial \mathbf{x}}\right) m(\mathbf{x}). \tag{1.3.16}$$

Indeed, we find non-trivial solutions for $m(\mathbf{x})$. We conclude that Kahan map yields an integrable discretisation for this ODE.

1.4 Summary of papers

Paper 1: A novel approach to rigid spheriod models using operator splitting

Benjamin K Tapley, Elena Celledoni, Brynjulf Owren, Helge I. Andersson

Numerical Algorithms 81, no. 4 (2019): 1423-1441.

In this application-focused paper we consider the numerical integration of the equations of motion for a spheroidal particle immersed in viscous flow (i.e., equations (1.2.1)-(1.2.5)), where the resistance tensor K and torque vector $\mathbf{T}$ come from the spheroidal particle model and are given by Brenner [6] and Jeffrey [23]. We develop a splitting scheme based on splitting the ODE into a free rigid body vector field and a vector field that takes into account the viscous Stokes force as in equation (1.2.14). We study the convergence via numerical tests for a variety of Stokes numbers.

Paper 2: Computational geometric methods for preferential clustering of particle suspensions

Benjamin K Tapley, Helge I Andersson, Elena Celledoni, Brynjulf Owren

Submitted to *Journal of Computational Physics*

In this paper, we develop a geometric method for simulating suspensions of spherical and non-spherical particles in a discrete flow field such as a numerical solution to the Navier-Stokes equations. We study the effect of breaking the divergence-free condition in simulations and propose a simple and effective method for diverge-free interpolation using matrix-valued radial basis functions. Furthermore, the equations of motion possess many features that are used to explain the preferential clustering of particles that we observe in experiment. We propose a composition method, based on the splitting scheme of the previous chapter, that preserves a number of relevant physical features in the numerical solution. We conduct numerical experiments in a cellular flow field and show that low-order fast geometric methods can outperform higher-order expensive general-purpose methods.

Paper 3: A slender body model for thin rigid fibers: validation and comparison

Laurel Ohm, Benjamin K Tapley, Helge I Andersson,
Elena Celledoni and Brynjulf Owren

Proc. of MEKiT'19, 10th Nat. Conf. on Comp. Mech., 2019

In this article, we consider a model, based on slender body theory, for calculating the forces and torques for a slender fiber in Stokes flow. We implement basic numerical methods to validate the accuracy of the model and compare it to other known slender body models. We also compare the dynamics of a prolate ellipse to using the slender body model to known dynamics using the Jeffery model.

Paper 4: An integral model based on slender body theory, with applications to curved rigid fibers

Helge I Andersson, Elena Celledoni, Laurel Ohm,
Brynjulf Owren and Benjamin K Tapley

Physics of Fluids 33 (4), 041904

This paper is an extension of the previous. Here, we further develop the previously proposed model to better suit it to numerical inversion. The result is a second-kind Fredholm integral equation. We implement a spectral quadrature method for calculating the force and torque on a rigid slender particle. We propose a fast and algorithm for computing the dynamics of rigid fibers. As the Fredholm integral equation needs to be inverted, we explore the invertibility and convergence properties of the numerical method and show that the algorithm is well conditioned.

Paper 5: Using discrete Darboux polynomials to detect and determine preserved measures and integrals of rational maps

Elena Celledoni, Charalambos A Evripidou, David I McLaren, Brynjulf Owren,
G R W Quispel, Benjamin K Tapley and Peter H van der Kamp

Journal of Physics A: Mathematical and Theoretical 52 31LT01

This letter introduces the idea of discrete Darboux polynomials, which is an analogue to Darboux polynomials of ODEs but applied to mappings. We present results on the preservation of Darboux polynomials by the Kahan map as well as a number of examples where one can determine expressions for the rational integrals that are preserved by this map using the method of discrete Darboux polynomials. We also present results on an algorithm that can detect extra Darboux solutions on a given map with free parameters. The algorithm can tell us the conditions that these free parameters must satisfy to yield extra Darboux polynomial solutions, which can lead to extra integrals.

Paper 6: Detecting and determining preserved measures and integrals of rational maps

Elena Celledoni, Charalambos A Evripidou, David I McLaren, Brynjulf Owren G R W Quispel and Benjamin K Tapley

Submitted to *Journal of Physics A: Mathematical and Theoretical*

This paper is the sequel to the previous letter. We present many novel examples of the integrability properties of Kahan map using the method of discrete Darboux polynomials. In particular, we show examples of the Kahan map preseving an irrational integral, a quartic rational integral and many more. We also present details of the algorithm that detects conditions of the free parameters that yield extra Darboux polynomial solutions and apply it to a number of systems including the coupled Euler tops, the extended McMillan map and a new class of Nambu systems for which the Kahan map preserves integrability.

Paper 7: On the preservation of affine second integrals by Runge-Kutta methods

Benjamin K Tapley

Pre-print

Most of the examples we have considered in the previous two papers have come about from Kahan's method applied to integrable ODEs. In this paper, we generalise the theory of discrete Darboux polynomials by considering their preservation by Runge-Kutta methods. By limiting our discussion to affine Darboux polynomials, we are able to make more general statements about how Darboux polynomials are preserved when discretised. In particular, we show that all Runge-Kutta methods preserve all affine second integrals with a modified discrete cofactor. We also discuss the preservation of higher affine integrals and show that Runge-Kutta methods can preserve some rational integrals for certain ODEs.

Paper 8: Symplectic integration of PDEs using Clebsch variables

Robert I McLachlan, Christian Offen and Benjamin K Tapley

Journal of Computational Dynamics 6.1, pp. 111–130

In this paper we consider the numerical integration of PDEs that can be formulated as Lie-Poisson systems. Our approach is to reformulate the PDE as a Hamiltonian system on a symplectic manifold by writing the system in the so-called Clebsch variables. The advantage is that this lifted system has symplectic structure, to which we apply a symplectic integrator. Compared to integration on the original Poisson manifold, our approach leads to better conservation properties and we observe better stability.

Bibliography

[1] H. I. Andersson, E. Celledoni, L. Ohm, B. Owren, and B. K. Tapley, *An integral model based on slender body theory, with applications to curved rigid fibers*, arXiv preprint arXiv:2012.11561, (2020).

[2] K. Atkinson and W. Han, *Theoretical numerical analysis*, vol. 39, Springer, 2005.

[3] J. Bec, *Fractal clustering of inertial particles in random flows*, Physics of fluids, 15 (2003), pp. L81–L84.

[4] P. S. Bernard, M. F. Ashmawey, and R. A. Handler, *An analysis of particle trajectories in computer-simulated turbulent channel flow*, Physics of Fluids A: Fluid Dynamics, 1 (1989), pp. 1532–1540.

[5] S. Blanes and F. Casas, *A concise introduction to geometric numerical integration*, CRC press, 2017.

[6] H. Brenner, *The stokes resistance of an arbitrary particle*, Chemical Engineering Science, 18 (1963), pp. 1–25.

[7] E. Celledoni, C. Evripidou, D. McLaren, B. Owren, G. Quispel, and B. Tapley, *Detecting and determining preserved measures and integrals of rational maps*, arXiv preprint arXiv:1902.04685, (2019).

[8] E. Celledoni, C. Evripidou, D. I. McLaren, B. Owren, G. R. W. Quispel, B. Tapley, and P. H. van der Kamp, *Using discrete darboux polynomials to detect and determine preserved measures and integrals of rational maps*, Journal of Physics A: Mathematical and Theoretical, 52 (2019), p. 31LT01.

[9] E. Celledoni, F. Fassò, N. Säfström, and A. Zanna, *The exact computation of the free rigid body motion and its use in splitting methods*, SIAM Journal on Scientific Computing, 30 (2008), pp. 2084–2112.

[10] E. Celledoni, R. I. McLachlan, D. I. McLaren, B. Owren, and G. Quispel, *Integrability properties of kahan's method*, Journal of Physics A: Mathematical and Theoretical, 47 (2014), p. 365202.

[11] E. CELLEDONI, R. I. MCLACHLAN, B. OWREN, AND G. R. W. QUISPEL, *Geometric properties of kahan's method*, Journal of Physics A: Mathematical and Theoretical, 46 (2012), p. 025201.

[12] N. R. CHALLABOTLA, L. ZHAO, AND H. I. ANDERSSON, *Orientation and rotation of inertial disk particles in wall turbulence*, Journal of Fluid Mechanics, 766 (2015).

[13] D. W. COOPER, *Particulate contamination and microelectronics manufacturing: an introduction*, Aerosol Science and Technology, 5 (1986), pp. 287–299.

[14] J. W. DEARDORFF AND R. L. PESKIN, *Lagrangian statistics from numerically integrated turbulent shear flow*, The Physics of Fluids, 13 (1970), pp. 584–595.

[15] E. HAIRER, C. LUBICH, G. WANNER, *Geometric Numerical Integration, Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, second ed., 2006.

[16] E. HAIRER, S.P. NØRSETT, G. WANNER, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer-Verlag, Berlin Heidelberg, second ed., 1993.

[17] P. ERNI, C. CRAMER, I. MARTI, E. J. WINDHAB, AND P. FISCHER, *Continuous flow structuring of anisotropic biopolymer particles*, Advances in colloid and interface science, 150 (2009), pp. 16–26.

[18] A. ESHGHINEJADFARD, S. A. HOSSEINI, AND D. THÉVENIN, *Fully-resolved prolate spheroids in turbulent channel flows: a lattice boltzmann study*, AIP Advances, 7 (2017), p. 095007.

[19] P. A. F. LUNDELL, L.D. SOĎERBERG, *Fluid mechanics of papermaking*, Annu. Rev. Fluid Mech, 43 (2011), p. 195–217.

[20] A. GORIELY, *Integrability and nonintegrability of dynamical systems*, vol. 19, World Scientific, 2001.

[21] V. GRIMM AND G. QUISPEL, *Geometric integration methods that unconditionally contract volume*, Applied numerical mathematics, 58 (2008), pp. 1103–1112.

[22] A. HEYMSFIELD, *Precipitation development in stratiform ice clouds: a microphysical and dynamical study*, J. Atmos. Sci., 34 (1977), p. 67–81.

[23] G. B. JEFFERY, *The motion of ellipsoidal particles immersed in a viscous fluid*, Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character, 102 (1922), pp. 161–179.

[24] W. KAHAN, *Unconventional numerical methods for trajectory calculations*, Unpublished lecture notes, 1 (1993), p. 13.

[25] F. KANG AND S. ZAI-JIU, *Volume-preserving algorithms for source-free dynamical systems*, Numerische Mathematik, 71 (1995), pp. 451–463.

[26] R. V. H. L. SABBAN, *Measurements of pollen grain dispersal in still air and stationary, near homogeneous, isotropic turbulence*, J. Aerosol Sci, 42 (2011), p. 67–82.

[27] B. LEIMKUHLER AND S. REICH, *Simulating hamiltonian dynamics*, no. 14, Cambridge university press, 2004.

[28] D. LOPEZ AND E. GUAZZELLI, *Inertial effects on fibers settling in a vortical flow*, Physical Review Fluids, 2 (2017), p. 024306.

[29] M. R. MAXEY, *The motion of small spherical particles in a cellular flow field*, The Physics of Fluids, 30 (1987), pp. 1915–1928.

[30] M. R. MAXEY AND J. J. RILEY, *Equation of motion for a small rigid sphere in a nonuniform flow*, The Physics of Fluids, 26 (1983), pp. 883–889.

[31] R. I. MCLACHLAN, C. OFFEN, AND B. K. TAPLEY, *Symplectic integration of pdes using clebsch variables*, Journal of Computational Dynamics, 6 (2019), pp. 111–130.

[32] R. I. MCLACHLAN AND G. QUISPEL, *Numerical integrators that contract volume*, Applied numerical mathematics, 34 (2000), pp. 253–260.

[33] R. I. MCLACHLAN AND G. R. W. QUISPEL, *Splitting methods*, Acta Numerica, 11 (2002), p. 341.

[34] P. MORTENSEN, H. ANDERSSON, J. GILLISSEN, AND B. BOERSMA, *Dynamics of prolate ellipsoidal particles in a turbulent channel flow*, Physics of Fluids, 20 (2008), p. 093302.

[35] W. W. NAZAROFF, *Indoor particle dynamics*, Indoor air, 14 (2004), pp. 175–183.

[36] A. OBERBECK, *Uber stationare flussigkeitsbewegungen mit berucksichtigung der inner reibung*, J. reine angew. Math., 81 (1876), pp. 62–80.

[37] L. OHM, *Mathematical foundations of slender body theory*, Doctoral dissertation, University of Minnesota, 2020.

[38] L. OHM, B. K. TAPLEY, H. I. ANDERSSON, E. CELLEDONI, AND B. OWREN, *A slender body model for thin rigid fibers: validation and comparisons*, arXiv preprint arXiv:1906.00253, (2019).

[39] Y. PAN AND S. BANERJEE, *Numerical simulation of particle interactions with wall turbulence*, Physics of Fluids, 8 (1996), pp. 2733–2755.

[40] L. M. PORTELA AND R. V. A. OLIEMANS, *Eulerian-Lagrangian DNS/LES of particle-turbulence interactions in wall-bounded flows*, International Journal for Numerical Methods in Fluids, 43 (2003), pp. 1045–1065.

[41] C. POZRIKIDIS ET AL., *Boundary integral and singularity methods for linearized viscous flow*, Cambridge university press, 1992.

[42] K. P. R.C. MOFFETT, *In-situ measurements of the mixing state and optical properties of soot with implications for radiative forcing estimates*, PNAS, 106 (2009), pp. 72–77.

[43] R. F. ROSS AND D. J. KLINGENBERG, *Dynamic simulation of flexible fibers composed of linked rigid bodies*, The Journal of chemical physics, 106 (1997), pp. 2949–2960.

[44] D. W. ROUSON AND J. K. EATON, *On the preferential concentration of solid particles in turbulent channel flow*, Journal of Fluid Mechanics, 428 (2001), p. 149.

[45] J. M. SANZ-SERNA, *Symplectic integrators for hamiltonian problems: an overview*, Acta numerica, 1 (1992), pp. 243–286.

[46] K. D. SQUIRES AND J. K. EATON, *Preferential concentration of particles by turbulence*, Physics of Fluids A: Fluid Dynamics, 3 (1991), pp. 1169–1178.

[47] L. TAKHTAJAN, *On foundation of the generalized nambu mechanics*, Communications in Mathematical Physics, 160 (1994), pp. 295–315.

[48] B. TAPLEY, E. CELLEDONI, B. OWREN, AND H. I. ANDERSSON, *A novel approach to rigid spheroid models in viscous flows using operator splitting methods*, Numerical Algorithms, (2019), pp. 1–19.

[49] B. K. TAPLEY, H. I. ANDERSSON, E. CELLEDONI, AND B. OWREN, *Computational geometric methods for preferential clustering of particle suspensions*, 2021.

[50] J. K. T.J. PEDLEY, *Hydrodynamic phenomena in suspensions of swimming microorganisms*, Annu. Rev. Fluid Mech, 24 (1992), p. 13–58.

[51] W. UIJTTEWAAL AND R. OLIEMANS, *Particle dispersion and deposition in direct numerical and large eddy simulations of vertical pipe flows*, Physics of Fluids, 8 (1996), pp. 2590–2604.

[52] B. A. VAN HAARLEM, *The dynamics of particles and droplets in atmospheric turbulence-A numerical study*, PhD thesis, Delft University of Technology, 2000.

[53] Q. WANG AND K. D. SQUIRES, *Large eddy simulation of particle-laden turbulent channel flow*, Physics of Fluids, 8 (1996), pp. 1207–1223.

[54] M. WILKINSON AND B. MEHLIG, *Caustics in turbulent aerosols*, EPL (Europhysics Letters), 71 (2005), p. 186.

[55] G. WUISPEL, *Volume-preserving integrators*, Physics Letters A, 206 (1995), pp. 26–30.

[56] H. YU, K. KANOV, E. PERLMAN, J. GRAHAM, E. FREDERIX, R. BURNS, A. SZALAY, G. EYINK, AND C. MENEVEAU, *Studying lagrangian dynamics of turbulence using on-demand fluid particle tracking in a public turbulence database*, Journal of Turbulence, (2012), p. N12.

[57] H. ZHAO, A. H. ISFAHANI, L. N. OLSON, AND J. B. FREUND, *A spectral boundary integral method for flowing blood cells*, Journal of Computational Physics, 229 (2010), pp. 3726–3744.

A novel approach to rigid spheriod models using operator splitting

Benjamin K Tapley, Elena Celledoni, Brynjulf Owren, Helge I Andersson

Numerical Algorithms 81, no. 4 (2019): 1423-1441.

A novel approach to rigid spheriod models using operator splitting

Abstract. Calculating cost-effective solutions to particle dynamics in viscous flows is an important problem in many areas of industry and nature. We implement a second-order symmetric splitting method on the governing equations for a rigid spheroidal particle model with torques, drag and gravity. The method splits the operators into a vector field that is conservative and one that takes into account the forces of the fluid. Error analysis and numerical tests are performed on perturbed and stiff particle-fluid systems. For the perturbed case, the splitting method greatly improves the solution accuracy, when compared to a conventional multi-step method, and the global error behaves as $\mathcal{O}(\varepsilon h^2)$ for roughly equal computational cost. For stiff systems, we show that the splitting method retains stability in regimes where conventional methods blow up. In addition, we show through numerical experiments that the global order is reduced from $\mathcal{O}(h^2/\varepsilon)$ in the non-stiff regime to $\mathcal{O}(h)$ in the stiff regime.

2.1 Introduction

Simulating the dynamics of particles in a fluid is of importance to many industrial applications such as paper making [11], pharmaceutical processing [27] and soot emission from combustion processes [34] as well as natural processes including the transportation of plankton in the sea [35], the formation of ice clouds [18] and the dispersion of pollen in the atmosphere [22]. With growing needs for larger models and longer simulation times, there is an increasing demand for effective numerical methods that minimise computational cost. Over the past 50 years, splitting methods have been used to model problems in molecular biology, physics and fluid dynamics, for example, and have been shown to supersede classical integration schemes in terms of both quantitative and qualitative accuracy [31]. In this paper, we employ splitting methods on the axisymmetric rigid-body equations with Stokes viscous force, torque and gravity. Splitting methods are often used when the differential equation has geometric properties that should be preserved under disretisation, such as being Hamiltonian or divergence-free; or possessing a symmetry or a first integral. The idea behind splitting methods is to split the system into two or more simpler sub-systems and compute the numerical flow as the composition of the analytic flows of the subsystems at discrete time-steps. As these methods are purpose-built for the problem under study, they have the ability to mimic the qualitative behaviour of the continuous solution resulting in efficiency and stability improvements over standard, all-purpose integration techniques.

The particle-fluid system is modelled under the assumptions that the particle size is smaller than the smallest fluid length scale (e.g., the Kolomogrov scale) and that the particle shape can be approximated by a triaxial ellipsoid. Under the first assumption, the particle-Reynolds number is likely to be low and the fluid can be approximated by Stokes flow conditions where the dominant forces are drag, torque and gravity. We adopt the second assumption for numerous reasons. Due to the inherent complexity of fluid dynamics, ellipsoids are the only shape where the fluid forces are exactly known at leading order without making overly restrictive assumptions. For example, slender body theory can tell us the forces on the particle only but only if the particle is very long and thin [1, 23, 30] and perturbation theory can tell us the translational [14] and rotational [29] forces only for nearly spherical particles. Other than these two cases, the only shape where the forces are known at leading order are ellipsoids, which are modelled by Stokes viscous force, derived by Brenner [3], and torques, derived by Jeffery [21]. Such models have been adopted in studies such as [17, 25, 28]. Additionally, modelling general non-spherical particles as axisymmetric spheroids, such as rigid rods [17] or disks [25], is a common leading order approximation, for example, ice-cloud particles are hexagonal plates and columns but are modelled as oblate and prolate spheroids [18]. For a comprehensive review on particle modelling the reader is referred to [12]. In this paper we pay particular attention to two cases, one where the fluid forces are seen as a perturbation to an otherwise free rigid-body system and the second is a stiff system, where the fluid forces dominate the free rigid-body equations.

For non-spherical particles, the orientation couples with the translational dynamics and therefore greatly increases the model complexity. As a result, a system of 13 coupled ordinary differential equations (ODEs) need to be solved per time-step: three each for the position, velocity and angular momentum vectors and four for the rotation quaternion. A typical approach to solving these ODEs has been to integrate the system using Runge-Kutta methods and/or linear multistep methods such as a second-order explicit Adams-Bashforth method [25,28]. These methods, although straightforward to implement, present a number of drawbacks when calculating long-time numerical solutions to ODEs: (1) stability restrictions on the time-step h; (2) not time symmetric; and (3) limited ability to conserve properties specific to the underlying physics of the system.

Such issues can only be overcome by enforcing small time-steps, thus increasing the total cost of the solution method, which limits the feasibility of large (e.g., $N > 10^6$ particles) or long (e.g., $T \in [0, 10^3]$ seconds) simulations [12].

Alternatively, one could approach the problem with a purpose-built algorithm, such as a splitting method, which takes advantage of particular properties of the vector field under study. Here, we show that when compared to a conventional two-step Adams-Bashforth method, the splitting method is both cheaper, more accurate and more robust thus allowing for larger time-steps to achieve the same accuracy.

The next section of the paper reviews relevant theory in particle modelling. We then introduce the numerical splitting method and present an error analysis. Section 2.5 presents some numerical experiments and the last section is dedicated to conclusions.

2.2 Governing equations

To describe the forces on the particle we first establish three reference frames. First, we define an *inertial frame* by variables $\mathbf{x} = (x, y, z)^{\mathrm{T}}$ that is an inertial coordinate system as shown in figure 2.1. Secondly, we define a *translating frame* by variables $\mathbf{x}'' = (x'', y'', z'')^{\mathrm{T}}$ that is translating with the particle and has its origin co-located with the particles center of mass. Lastly, we introduce a *body frame* denoted by variables $\mathbf{x}' = (x', y', z')^{\mathrm{T}}$ that is translating and rotating with the particle. Henceforth, all primed and double primed variables are respectively defined in the body and translating frame and unprimed variables are defined in the inertial frame.

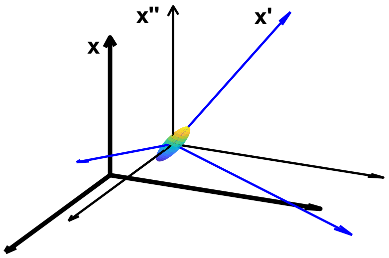


Figure 2.1: A prolate spheroid ($\lambda = 3$) with coordinate lines of the inertial frame (thick black arrows), translating frame (thin black arrows) and the body frame (thin blue arrows).

Jeffery and Brenner derived forces for general rigid ellipsoids, which have three distinct semi-axis lengths; however, for simplicity we will focus on spheroids, which are axisymmetric. In the body frame, a spheroid is defined by

$$\frac{x'^2}{a^2} + \frac{y'^2}{a^2} + \frac{z'^2}{c^2} = 1, \tag{2.2.1}$$

where a and c are the distinct semi-axis lengths. The particle shape is characterised by the dimensionless aspect ratio $\lambda = c/a > 0$, which distinguishes between spherical ($\lambda = 1$), prolate ($\lambda > 1$) and oblate ($\lambda < 1$) particles (the latter two shapes are also called as rods and disks). The axisymmetric moment of inertia tensor for a spheroid in the body frame is

$$I' = ma^2 \text{diag}\left(\frac{(1+\lambda^2)}{5}, \frac{(1+\lambda^2)}{5}, \frac{2}{5}\right), \tag{2.2.2}$$

where $m = \frac{4}{3}\pi\lambda a^3 \rho_p$ is the particle mass and ρ_p is the particle density.

A spheroid immersed in a fluid will experience forces on its surface that have magnitude governed by many parameters such as the particles density ρ_p, semi-major axis length a, aspect ratio λ, fluid density ρ_f, dynamic viscosity ν and fluid relaxation time τ_f, which is defined in section 2.2.3. Hence, it is a logical step to non-dimensionalise our equations by introducing a dimensionless Stokes number. The particle Stokes number is formally defined as the ratio of the particle and fluid relaxation times $St = \tau_p/\tau_f$. In this paper, we will adopt the definition

$$St = \frac{D\lambda^2 a^2}{\nu \tau_f}, \tag{2.2.3}$$

where $D = \frac{\rho_p}{\rho_f}$ is the particle-fluid density ratio. The Stokes number is a dimensionless measure of the relative importance of particle inertia, that is, as $St \to \infty$ the particle behaves as a free body and as $St \to 0$ the particle behaves as if itself were part of the fluid. Henceforth, all equations are presented in their non-dimensional form and all parameters have dimension equal to 1.

The linear momentum, angular momentum and position can be described by the column vectors $\mathbf{p}, \mathbf{m}', \mathbf{x} \in \mathbb{R}^3$, and the orientation can be represented using Euler parameters [15], i.e. a vector $q = (e_0, e_1, e_2, e_3) \in \mathbb{R}^4$ satisfying the constraint

$$1 = e_0^2 + e_1^2 + e_2^2 + e_3^2, \tag{2.2.4}$$

that uniquely determines the orientation of the body frame relative to the axes of translating frame (and hence to the inertial frame subject to an additional translation). The Euler parameters were first used for particle modelling by Fan

[10] and are used in place of the conventional Euler angles to avoid singularities. Each q uniquely determines a rotation matrix $Q \in SO(3)$ that transforms a vector in the body frame $\mathbf{x}'$ to a vector in the translating frame $\mathbf{x}''$ via

$$\mathbf{x}'' = Q\mathbf{x}'. \tag{2.2.5}$$

There is a 2-to-1 correspondence between Euler parameters and 3 × 3 rotation matrices given by the so called Euler-Rodriguez map $\mathscr{E} : q \mapsto Q$ [5]. Setting $\mathbf{e} = (e_1, e_2, e_3)$, the rotation matrix $\mathscr{E}(q) = Q$ is constructed via

$$Q = \mathbb{1} + 2e_0\hat{\mathbf{e}} + 2\hat{\mathbf{e}}\hat{\mathbf{e}}, \tag{2.2.6}$$

where $\mathbb{1}$ is the 3 × 3 identity matrix and we have introduced the hat map $\widehat{\cdot} : \mathbb{R}^3 \to \mathfrak{so}(3)$ defined by

$$\begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \mapsto \hat{\boldsymbol{\omega}} = \begin{pmatrix} 0 & -\omega_1 & \omega_2 \\ \omega_1 & 0 & -\omega_3 \\ -\omega_2 & \omega_3 & 0 \end{pmatrix}, \tag{2.2.7}$$

where $\mathfrak{so}(3)$ is the Lie algebra of $SO(3)$ containing 3 × 3 skew-symmetric matrices satisfying $\boldsymbol{\omega} \times \mathbf{v} = \hat{\boldsymbol{\omega}}\mathbf{v}$ for $\boldsymbol{\omega}, \mathbf{v} \in \mathbb{R}^3$. This gives the following expression for Q explicitly in terms of the Euler parameters

$$Q = \begin{pmatrix} e_0^2 + e_1^2 - e_2^2 - e_3^2 & 2(e_1e_2 - e_0e_3) & 2(e_1e_3 + e_0e_2) \\ 2(e_1e_2 + e_0e_3) & e_0^2 - e_1^2 + e_2^2 - e_3^2 & 2(e_2e_3 - e_0e_1) \\ 2(e_1e_3 - e_0e_2) & 2(e_2e_3 + e_0e_1) & e_0^2 - e_1^2 - e_2^2 + e_3^2 \end{pmatrix}. \tag{2.2.8}$$

2.2.1 Translational dynamics

The Stokes viscous force, derived in [3] and gravity force terms, are given in their non-dimensional form by

$$\mathbf{F}_h = \frac{3\lambda}{4St} QK'Q^{\mathrm{T}}(\mathbf{u} - \mathbf{v}), \tag{2.2.9}$$

$$\mathbf{F}_g = -\, m\mathbf{g}, \tag{2.2.10}$$

where $\mathbf{v}$ is the inertial frame linear velocity, which is related to linear momentum via $\mathbf{p} = m\mathbf{v}$. Note that in our non-dimensional formalism we take $m = 1$ to be a dimensionless constant; however, we will leave m in our equations for consistency with the literature. The inertial frame fluid velocity vector $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ is taken at the location of the particle $\mathbf{x}$ and the inertial frame gravity acceleration vector is $\mathbf{g} = (0, 0, g)^T$ for some positive constant g that is typically defined as $g = 1 - 1/D$ to account for the buoyancy force. The body frame resistance tensor K', derived by Oberbeck [26], is given by

$$K' = 16\pi\lambda\ \mathrm{diag}\left(\frac{1}{\chi_0+\alpha_0}, \frac{1}{\chi_0+\beta_0}, \frac{1}{\chi_0+\lambda^2\gamma_0}\right) \tag{2.2.11}$$

where the constants χ_0, α_0, β_0 and γ_0 were calculated for ellipsoidal particles by Siewert [4] and are presented in table 3.A.1. Note that the inertial frame resistance tensor K is calculated from the similarity transformation $K = QK'Q^{\mathrm{T}}$.

	$\lambda < 1$	$\lambda = 1$	$\lambda > 1$
χ_0	$\frac{\lambda^2(\pi-\kappa_0)}{\sqrt{1-\lambda^2}}$	2	$\frac{-\kappa_0\lambda}{\sqrt{\lambda^2-1}}$
$\alpha_0 = \beta_0$	$\frac{-\lambda\left(\kappa_0-\pi+2\lambda\sqrt{1-\lambda^2}\right)}{2(1-\lambda^2)^{3/2}}$	$\frac{2}{3}$	$\frac{\lambda^2}{\lambda^2-1} + \frac{\lambda\kappa_0}{2(\lambda^2-1)^{3/2}}$
γ_0	$\frac{\left(\lambda(\kappa_0-\pi)+2\sqrt{1-\lambda^2}\right)}{(1-\lambda^2)^{3/2}}$	$\frac{2}{3}$	$\frac{-2}{\lambda^2-1} - \frac{\lambda\kappa_0}{(\lambda^2-1)^{3/2}}$
κ_0	$2\arctan\left(\frac{\lambda}{\sqrt{1-\lambda^2}}\right)$	1	$\ln\left(\frac{\lambda-\sqrt{\lambda^2-1}}{\lambda+\sqrt{\lambda^2-1}}\right)$

Table 2.1: The values for the constants χ_0, α_0, β_0 and γ_0 for $\lambda < 1$, $\lambda = 1$ and $\lambda > 1$.

It will be convenient for the formulation of the methods to rewrite equation (2.2.9) as

$$\mathbf{F}_h = -A_1\mathbf{p} + \mathbf{b}_1, \tag{2.2.12}$$

where

$$A_1 = \frac{3\lambda}{4mSt}K \quad \text{and} \quad \mathbf{b}_1 = mA_1\mathbf{u}(\mathbf{x}, t). \tag{2.2.13}$$

Here, $\mathbf{b}_1$ is implicitly dependent on time through the fluid. This leads to the following ODE for momentum

$$\dot{\mathbf{p}} = -A_1\mathbf{p} + \mathbf{b}_1 - m\mathbf{g}. \tag{2.2.14}$$

The inertial frame position vector $\mathbf{x}$ is calculated by solving

$$\dot{\mathbf{x}} = \mathbf{v}. \tag{2.2.15}$$

2.2.2 Rotational dynamics

The rotational dynamics of an ellipsoidal particle are governed by the free rigid-body equations [20] with torques $\mathbf{N}' = (N'_x, N'_y, N'_z)^{\mathrm{T}}$ that describe the rotational forces acting on an ellipsoid in creeping Stokes flow in the body frame [21]. These are presented in their non-dimensional form

$$N'_x = \frac{16\pi\lambda}{3(\beta_0 + \lambda^2\gamma_0)} \left[(1-\lambda^2)S'_{yz} + (1+\lambda^2)(\Omega'_x - \omega'_x)\right], \tag{2.2.16}$$

$$N'_y = \frac{16\pi\lambda}{3(\alpha_0 + \lambda^2\gamma_0)} \left[(\lambda^2 - 1)S'_{zx} + (1+\lambda^2)(\Omega'_y - \omega'_y)\right], \tag{2.2.17}$$

$$N'_z = \frac{32\pi\lambda}{3(\alpha_0 + \beta_0)}(\Omega'_z - \omega'_z), \tag{2.2.18}$$

where $\boldsymbol{\omega}' = (\omega'_x, \omega'_y, \omega'_z)^{\mathrm{T}}$ is the body frame angular velocity, which is related to body frame angular momentum by $\mathbf{m}' = I'\boldsymbol{\omega}'$. The dimensionless body frame shear $\mathbf{S}' = (S'_{yz}, S'_{zx}, S'_{xy})^{\mathrm{T}}$ and fluid rotation $\boldsymbol{\Omega}' = (\Omega'_x, \Omega'_y, \Omega'_z)^{\mathrm{T}}$ terms are

$$S'_{ij} = \frac{1}{2}\left(\frac{\partial u'_i}{\partial x'_j} + \frac{\partial u'_j}{\partial x'_i}\right) \quad \text{and} \quad \Omega'_i = \frac{1}{2}(\nabla' \times \mathbf{u}')_i. \tag{2.2.19}$$

We write equations (5.C.6), (5.C.7) and (5.C.8) compactly as

$$\mathbf{N}' = -A'_2\mathbf{m}' + \mathbf{b}'_2, \tag{2.2.20}$$

where

$$A'_2 = \frac{12\lambda^2}{St}\mathrm{diag}\left(\frac{(1+\lambda^2)}{(\beta_0 + \lambda^2\gamma_0)}, \frac{(1+\lambda^2)}{(\alpha_0 + \lambda^2\gamma_0)}, \frac{2}{(\alpha_0 + \beta_0)}\right) I'^{-1}, \tag{2.2.21}$$

and

$$\mathbf{b}'_2 = \frac{12\lambda^2}{St}\mathrm{diag}\left(\frac{(1-\lambda^2)}{(\beta_0 + \lambda^2\gamma_0)}, \frac{(\lambda^2 - 1)}{(\alpha_0 + \lambda^2\gamma_0)}, 0\right)\mathbf{S}' + A_2 I'\boldsymbol{\Omega}'. \tag{2.2.22}$$

Here, $\mathbf{b}'_2$ is implicitly dependent on time through the shear and rotation terms. The dimensionless equation governing the angular momentum of the particle in the body frame is therefore

$$\dot{\mathbf{m}}' = \mathbf{m}' \times \boldsymbol{\omega}' - A'_2\mathbf{m}' + \mathbf{b}'_2, \tag{2.2.23}$$

where the cross-product term is the Poisson bracket for the free rigid-body [20] that arises from the fact that $\mathbf{m}'$ is represented in the (non-inertial) body frame. The rotation matrix Q is calculated by solving the matrix ODE

$$\dot{Q} = Q\widehat{\boldsymbol{\omega}}', \tag{2.2.24}$$

which arises from the quaternion formulation for the rigid-body, see [5] for details. When designing a splitting method, it is notationally convenient to express the ODEs as vector equations. To do so we will denote $\mathbf{q}_i$ to be the ith column of Q^{T}, then

$$\dot{\mathbf{q}}_i = -\widehat{\boldsymbol{\omega}}'\mathbf{q}_i \quad \text{for} \quad i = 1,2,3 \tag{2.2.25}$$

which represents three vector equations. It is important to stress, that to ensure that the orthogonality of Q is preserved, it is equation (3.2.4) that is being solved during the implementation of the splitting method and not equation (2.2.25).

2.2.3 Fluid field

This paper is only concerned with the performance of numerical methods in calculating solutions to particle dynamics, so as to measure this in isolation of the costs associated with discrete fluid field interpolation, an analytic fluid field that is known everywhere in time and space is used. The inertial frame fluid velocity vector $\mathbf{u} = (u, v, w)^{\mathrm{T}}$ is modelled by an analytic solution to the Navier-Stokes equations derived by Ethier and Steinman [9]

$$u = -\alpha_f[\mathrm{e}^{\alpha_f x}\sin(\alpha_f y \pm \beta_f z) + \mathrm{e}^{\alpha_f z}\cos(\alpha_f x \pm \beta_f y)]\mathrm{e}^{-\beta_f^2 t}, \tag{2.2.26}$$

$$v = -\alpha_f[\mathrm{e}^{\alpha_f y}\sin(\alpha_f z \pm \beta_f x) + \mathrm{e}^{\alpha_f x}\cos(\alpha_f y \pm \beta_f z)]\mathrm{e}^{-\beta_f^2 t}, \tag{2.2.27}$$

$$w = -\alpha_f[\mathrm{e}^{\alpha_f z}\sin(\alpha_f x \pm \beta_f y) + \mathrm{e}^{\alpha_f y}\cos(\alpha_f z \pm \beta_f x)]\mathrm{e}^{-\beta_f^2 t}, \tag{2.2.28}$$

for positive constants α_f and β_f. The fluid model has time scale $\tau_f = \beta_f^{-2}$ and is chosen as it has non-zero, non-trivial velocities that depend on every direction in each component of $\mathbf{u}$ and its Jacobian $\nabla\mathbf{u}$, and is derived from the full Navier-Stokes equation (i.e., without neglecting the convective, diffusive, unsteady or pressure terms). We assert that this fluid field provides a reasonable test of the solution methods in a non-trivial fluid and insights into their performance when the flow is transitioned to a realistic field, for example in [17,25,28]. In addition, we will conduct long-time experiments on an oscillating shear flow field defined by $\mathbf{u}_S = (0, 0, x\cos(2\pi t)/\tau_f)^{\mathrm{T}}$.

2.3 Numerical methods

2.3.1 Splitting

Splitting methods can be used when an ODE can be expressed as the sum of two or more operators,

$$\dot{\mathbf{y}}(t) = f(\mathbf{y}) = f_1(\mathbf{y}) + f_2(\mathbf{y}), \tag{2.3.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ and $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}^n$. Ideally, the splitting is chosen in such a way that the flows* $\varphi_h^{[1]}$ and $\varphi_h^{[2]}$ of the systems $\dot{\mathbf{y}}(t) = f_1(\mathbf{y})$ and $\dot{\mathbf{y}}(t) = f_2(\mathbf{y})$ can be computed exactly. In this case, numerical approximations can be generated by

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]}, \quad \text{or} \quad \Phi_h^* = \varphi_h^{[2]} \circ \varphi_h^{[1]}, \tag{2.3.2}$$

which are known as Lie-Trotter splittings [19] and are each others adjoints. Taylor expansion shows that the method is first-order. Another numerical method can be generated by

$$\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}, \tag{2.3.3}$$

which is the Strang splitting method [13]. Note that this can be written as the composition of the above Lie-Trotter methods with half time-steps $\Phi_h^{[S]} = \Phi_{h/2} \circ \Phi_{h/2}^*$, hence the method is of second-order and is symmetric [6, pg. 45]. Similarly, $\Phi_h^{[S]} = \Phi_{h/2}^* \circ \Phi_{h/2}$ is also a second-order symmetric method. Symmetric methods of arbitrarily high order can be generated by composition of the above methods, however, we refer the reader to [16, 24] for a more complete description of high-order splitting methods. For a full review of splitting theory, we refer the reader to [31].

2.3.2 System of differential equations

Let $\mathbf{y}(t) = (\mathbf{p}^{\mathrm{T}}, \mathbf{m}'^{\mathrm{T}}, \mathbf{q}_1^{\mathrm{T}}, \mathbf{q}_2^{\mathrm{T}}, \mathbf{q}_3^{\mathrm{T}}, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{18}$ be the solution to the ODE in the form of equation (2.3.1). The particles dynamics is governed by the following system of first-order coupled ODEs

$$\left.\begin{aligned} \dot{\mathbf{p}} &= -A_1\mathbf{p} + \mathbf{b}_1 - m\mathbf{g}, \\ \dot{\mathbf{m}}' &= \mathbf{m}' \times \boldsymbol{\omega}' - A_2'\mathbf{m}' + \mathbf{b}_2', \\ \dot{\mathbf{q}}_i &= -\widehat{\boldsymbol{\omega}}'\mathbf{q}_i, \quad \text{for} \quad i = 1,2,3 \\ \dot{\mathbf{x}} &= \mathbf{v}, \end{aligned}\right\} f(\mathbf{y}) \tag{2.3.4}$$

where the RHS of the equations in (2.3.4) arises due to the vector field $f(\mathbf{y})$. The kinetic and potential energies K and U, and Hamiltonian H are given by

$$K(\mathbf{y}) = \frac{1}{2}\mathbf{p}^{\mathrm{T}} m^{-1}\mathbf{p} + \frac{1}{2}\mathbf{m}'^{\mathrm{T}} I'^{-1}\mathbf{m}', \tag{2.3.5}$$

$$U(\mathbf{y}) = \frac{1}{2}\left(\mathbf{q}_1^{\mathrm{T}}\mathbf{q}_1 + \mathbf{q}_2^{\mathrm{T}}\mathbf{q}_2 + \mathbf{q}_3^{\mathrm{T}}\mathbf{q}_3\right) + m\mathbf{x}^{\mathrm{T}}\mathbf{g}, \tag{2.3.6}$$

$$H(\mathbf{y}) = K(\mathbf{y}) + U(\mathbf{y}), \tag{2.3.7}$$

*We denote by φ_h the flow operator such that $\mathbf{y}(h) = \varphi_h(\mathbf{y}_0)$ is the solution of the ODE at time $t = h$ with initial conditions $\mathbf{y}_0$ at $t = 0$.

where $\sum_{i=1}^{3} \mathbf{q}_i^{\mathrm{T}} \mathbf{q}_i = 3$ is a constant. The gradient of the Hamiltonian is

$$\nabla H(\mathbf{y}) = \left(\mathbf{v}^{\mathrm{T}}, \boldsymbol{\omega}'^{\mathrm{T}}, \mathbf{q}_1^{\mathrm{T}}, \mathbf{q}_2^{\mathrm{T}}, \mathbf{q}_3^{\mathrm{T}}, m\mathbf{g}^{\mathrm{T}}\right)^{\mathrm{T}}, \tag{2.3.8}$$

and is related to the solution vector $\mathbf{y}$ by the following non-injective mapping

$$\nabla H = M\mathbf{y} + \mathbf{g}_1, \tag{2.3.9}$$

where the matrix $M := \mathrm{diag}(m^{-1}\mathbb{1}, I^{-1}, \mathbb{1}, \mathbb{1}, \mathbb{1},) \in \mathbb{R}^{18\times 18}$ is diagonal and singular and $\mathbf{g}_1 = (0, \cdots, 0, m\mathbf{g}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{18}$. Now $\dot{\mathbf{y}}$ can be written as

$$\dot{\mathbf{y}} = f(\mathbf{y}) = S\nabla H - A\mathbf{y} + \mathbf{b}, \tag{2.3.10}$$

where $S \in \mathbb{R}^{18\times 18}$ is a skew-symmetric matrix given by

$$S = \begin{pmatrix} & & & & & -\mathbb{1} \\ & \widehat{\mathbf{m}}' & & & & \\ & & -\widehat{\boldsymbol{\omega}}' & & & \\ & & & -\widehat{\boldsymbol{\omega}}' & & \\ & & & & -\widehat{\boldsymbol{\omega}}' & \\ \mathbb{1} & & & & & \end{pmatrix}, \tag{2.3.11}$$

$A \in \mathbb{R}^{18\times 18}$ is a diagonal matrix given by

$$A = \mathrm{diag}(A_1, A_2', , \ldots, ,), \tag{2.3.12}$$

$\mathbf{b} \in \mathbb{R}^{18}$ is a vector given by

$$\mathbf{b} = (\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}'^{\mathrm{T}}_2, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^{18}, \tag{2.3.13}$$

and $\in \mathbb{R}^{3\times 3}$ is the zero matrix. Note from equations (2.2.13) and (5.C.9) that matrices A_1 and A_2' are positive definite, hence A is positive semi-definite and therefore represents a linear dissipation. Additionally, vectors $\mathbf{b}_1$ and $\mathbf{b}_2'$ represent the forces of the fluid on the particle, hence $\mathbf{b}$ is a non-conservative force term. As the energy of such a system is necessarily non-constant, we can calculate the exact energy dissipation by taking the time derivative of the Hamiltonian

$$\dot{H} = \nabla H^{\mathrm{T}} \dot{\mathbf{y}} = \nabla H^{\mathrm{T}} (-A\mathbf{y} + \mathbf{b}), \tag{2.3.14}$$

where we have used the fact that $\nabla H^{\mathrm{T}} S \nabla H = 0$ for skew-symmetric matrix S. With the forethought that we would like a dissipation-preserving splitting scheme, we split $f(\mathbf{y})$ into the following two sub-systems

$$\dot{\mathbf{y}} = f_1(\mathbf{y}) = S\nabla H, \tag{2.3.15}$$

$$\dot{\mathbf{y}} = f_2(\mathbf{y}) = -A\mathbf{y} + \mathbf{b}. \tag{2.3.16}$$

The first system is Hamiltonian and hence $\dot{H}^{[1]} = 0$ while the second system dissipates energy according to $\dot{H}^{[2]} = \nabla H^{\mathrm{T}} f_2(\mathbf{y}) = \nabla H^{\mathrm{T}}(-A\mathbf{y}+\mathbf{b})$. Hence, the numerical flow given by equation (2.3.3) preserves, up to the order of the method, the energy dissipation of the continuous system given by equation (2.3.14). Equations (2.3.15) and (2.3.16) correspond to the following systems of ODEs

$$\left.\begin{aligned} \dot{\mathbf{p}} &= -\mathbf{g} \\ \dot{\mathbf{m}}' &= -\widehat{\omega}'\mathbf{m}' \\ \dot{\mathbf{q}}_i &= -\widehat{\omega}'\mathbf{q}_i \\ \dot{\mathbf{x}} &= \mathbf{v} \\ (\dot{t} &= 1) \end{aligned}\right\} f_1(\mathbf{y}) \quad \text{and} \quad \left.\begin{aligned} \dot{\mathbf{p}} &= -A_1\mathbf{p}+\mathbf{b}_1 \\ \dot{\mathbf{m}}' &= -A_2'\mathbf{m}'+\mathbf{b}_2' \\ \dot{\mathbf{q}}_i &= \mathbf{0} \\ \dot{\mathbf{x}} &= \mathbf{0} \\ (\dot{t} &= 0) \end{aligned}\right\} f_2(\mathbf{y}), \tag{2.3.17}$$

where $f_1(\mathbf{y})$ represents a free rigid-body vector field with gravity, while $f_2(\mathbf{y})$ represents a purely energy dissipative (exponential decaying) vector field with a non-conservative force that leaves Q and $\mathbf{x}$ constant. Note that we freeze the flow of time in the second system to remove any implicit time dependence that $\mathbf{b}_1$ and $\mathbf{b}_2'$ may have through the fluid vector field.

Solutions to $f_1(\mathbf{y})$

The original system of ODEs is split such that the resulting sub-systems have solutions that can be computed analytically. The first system is solved using the well known solutions for axisymmetric rigid bodies [6, chapt. VII.5]. Note that this method can be generalised to triaxial ellipsoids, see for example [5,20,33]. First, the angular velocity $\boldsymbol{\omega}'$ is solved by

$$\boldsymbol{\omega}'(h) = R_z'(\mu h)\boldsymbol{\omega}_0', \tag{2.3.18}$$

where $\mu = \omega_z'(0)\frac{I_x'-I_z'}{I_x'}$ and $R_z'(\mu h)$ is a planar rotation of angle μh about the z' axis of the body frame

$$R_z'(\mu h) = \begin{pmatrix} \cos(\mu h) & \sin(\mu h) & 0 \\ -\sin(\mu h) & \cos(\mu h) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.3.19}$$

This immediately yields the angular momentum

$$\mathbf{m}'(h) = I'\boldsymbol{\omega}'(h). \tag{2.3.20}$$

Next, setting $w(h) = (0, \boldsymbol{\omega}'(h)^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^4$, the rotation matrix $Q = \mathscr{E}(q)$ is solved by computing the quaternion

$$q(h) = q_0 \cdot \exp_{\mathrm{q}}\left(\frac{h}{2} w(h/2)\right), \tag{2.3.21}$$

where $\exp_q$ is the quaternion exponential and the $\cdot$ represents multiplication of two quaternions (see [5] for details). Here, $w(h/2)$ is evaluated at a half time-step to maintain symmetry. The linear momentum $\mathbf{p}$ is solved by

$$\mathbf{p}(h) = -m\mathbf{g}h + \mathbf{p}_0, \tag{2.3.22}$$

and the position $\mathbf{x}$ is calculated by integrating the velocity

$$\mathbf{x}(h) = -\frac{1}{2}\mathbf{g}h^2 + \frac{1}{m}\mathbf{p}_0 h + \mathbf{x}_0. \tag{2.3.23}$$

These solutions to $\mathbf{p}$, $\mathbf{m}'$, Q and $\mathbf{x}$ at time h are represented by the flow map $\varphi_h^{[1]}$ in equation (2.3.3).

Solutions to $f_2(\mathbf{y})$

The $\mathbf{m}'$ and $\mathbf{p}$ equations in $f_2(\mathbf{y})$ of equation (2.3.17) are solved using the variation of constants formula

$$\mathbf{p}(h) = \exp\left(A_1 h\right)\left(\mathbf{p}_0 + A_1^{-1}\mathbf{b}_1\right) - A_1^{-1}\mathbf{b}_1, \tag{2.3.24}$$

$$'(h) = \exp\left(A_2' h\right)\left('_0 + A'^{-1}_2\mathbf{b}_2'\right) - A'^{-1}_2\mathbf{b}_2'. \tag{2.3.25}$$

Where vectors $\mathbf{b}_1$ and $\mathbf{b}_2'$ are constant in this system as we have enforced $\dot{t} = 0$. Additionally, the rotation matrix Q and the position vector $\mathbf{x}$ are also kept constant in this system. These solutions at time h are represented by the flow map $\varphi_h^{[2]}$ in equation (2.3.3).

2.4 Error Analysis

The dissipative system $f_2(\mathbf{y})$ that represents the fluid forces is inversely proportional to the Stokes number St which can be taken to be small ($St << 1$) or large ($St >> 1$), depending on the application. In addition, the choice of λ can greatly effect the magnitude of matrix A and vector $\mathbf{b}$. In fact it can be shown that $||A|| \leq c_1\lambda^4/St$ for $\lambda > 1$ and $||A|| \leq c_2\sqrt{\lambda}/St$ for $\lambda < 1$ (see table 3.A.1) and for some positive constants c_1 and c_2. This leads us to consider at least two main cases: one where $f_2(\mathbf{y}) = \varepsilon\tilde{f}_2(\mathbf{y})$ is a perturbation and another where $f_2(\mathbf{y}) = \frac{1}{\varepsilon}\tilde{f}_2(\mathbf{y})$ is a stiff term for $0 < \varepsilon << 1$. For the remainder of this section we will set $\mathbf{b} = \mathbf{0}$ (i.e. that f_2 consists only of a linearly dissipative term) and assume that gravity is negligible such that $\nabla H \approx M\mathbf{y}$. We will use backward error analysis to study the error in the non-stiff case, and we will illustrate the behaviour of the error in the stiff case by numerical tests. We will let γ^i represent the eigenvalues of the dissipation matrix A, of which six are non-zero and are the diagonal elements of matrices $A_1' = Q^{\mathrm{T}}A_1Q$ and A_2', given in equations

(2.2.13) and (5.C.9) respectively.

The local error for the energy H is given by the scalar

$$\delta_H(\mathbf{y}_0) = H(\mathbf{y}(h)) - H(\mathbf{y}_1). \tag{2.4.1}$$

If gravity is negligible, the particles energy is only kinetic, hence $H = \frac{1}{2}\mathbf{y}^{\mathrm{T}} M\mathbf{y} = \frac{1}{2}\nabla H^{\mathrm{T}}\mathbf{y}$. Using the fact that the numerical approximation $\mathbf{y}_1$ differs from the exact solution $\mathbf{y}(h)$ by the local error $\mathbf{y}_1 = \mathbf{y}(h) + \boldsymbol{\delta}(\mathbf{y}_0)$, it follows that the local energy error reduces to

$$\delta_H(\mathbf{y}_0) = -\nabla H^{\mathrm{T}}\boldsymbol{\delta}(\mathbf{y}_0) + \mathcal{O}(||\boldsymbol{\delta}||^2). \tag{2.4.2}$$

The next section will be dedicated to calculating the local solution error $\boldsymbol{\delta}(\mathbf{y}_0)$ and local energy error $\delta^{[H]}(\mathbf{y}_0)$ for the numerical method for the perturbed case. For the stiff case we will explore the global error using numerical experiments.

2.4.1 Non-stiff case

Here, we will look at a modified vector field that coincides exactly with the flow of the numerical method and compare this to the exact vector field. For $\gamma^i << 1$, we can write the ODE as $\dot{\mathbf{y}} = f_1(\mathbf{y}) + f_2(\mathbf{y}) = f_1(\mathbf{y}) + \varepsilon\tilde{f}_2(\mathbf{y})$. Here, we have introduced the scaled variables, denoted by the tilde, in our case $\varepsilon\tilde{f}_2 = \varepsilon\tilde{A}\mathbf{y}$. For arguments sake, we will analyse the error for the Lie-Trotter splitting as the results are more concise and analogous to the Strang splitting method. The numerical flow corresponding to the Lie-Trotter operator is

$$\Phi_h^{[LT]}(\mathbf{y}_0) = \varphi_h^{[1]} \circ \varphi_h^{[2]}(\mathbf{y}_0), \tag{2.4.3}$$

The local error can be determined by taking the difference between the exact and numerical flow over one time-step starting from the initial conditions $\mathbf{y}(0) = \mathbf{y}_0$. It follows that the local error for the Lie-Trotter method is

$$\boldsymbol{\delta}^{[LT]}(\mathbf{y}_0) = \varphi_h(\mathbf{y}_0) - \Phi_h^{[LT]}(\mathbf{y}_0) = \frac{h^2}{2}[f_1, f_2]\mathbf{y}_0 + \mathcal{O}(h^3), \tag{2.4.4}$$

where we have Taylor expanded the flows and use the bilinear Lie bracket of vector fields [6, chapt. IV], which expressed in coordinates is given by

$$[f_1, f_2] = \sum_{i,j=1}^{n} \left(f_1^i \partial_i f_2^j - f_2^i \partial_i f_1^j\right)\partial_j, \tag{2.4.5}$$

where f_1^j is the jth element of f_1 and $\partial_i = \partial/\partial y^i$. Inserting equations (2.3.15) and (2.3.16) into (2.4.4) we can write the local error explicitly

$$\boldsymbol{\delta}^{[LT]}(\mathbf{y}) = \frac{\varepsilon h^2}{2}(S\nabla^2 H\tilde{A}\mathbf{y} - \tilde{A}S\nabla H - \nabla\tilde{A}(S\nabla H, y, \cdot) - \nabla S(\tilde{A}\mathbf{y}, \nabla H, \cdot)) + \mathcal{O}(h^3), \tag{2.4.6}$$

where the tri-linear tensor ∇S is calculated by taking the gradient of S and satisfies the skew-symmetric relationship $\nabla S(\mathbf{u},\mathbf{v},\mathbf{w}) = -\nabla S(\mathbf{u},\mathbf{w},\mathbf{v})$ in its last two components for vectors $\mathbf{u},\mathbf{v},\mathbf{w} \in \mathbb{R}^{18}$, hence $\nabla S(\tilde{A}\mathbf{y}, \nabla H, \cdot)$ is interpreted as a column vector. If we insert equation (2.4.6) into equation (2.4.2) we can compute the local energy error

$$\begin{aligned}\delta_H^{[LT]}(\mathbf{y}) =& \frac{\varepsilon h^2}{2}(\nabla H^{\mathrm{T}} S \nabla^2 H \tilde{A}\mathbf{y} - \nabla H^{\mathrm{T}} \tilde{A} S \nabla H - \nabla \tilde{A}(S\nabla H, y, \nabla H) - \nabla S(\tilde{A}\mathbf{y}, \nabla H, \nabla H)) + \mathcal{O}(h^3),\\ =& \frac{\varepsilon h^2}{2}(\nabla H^{\mathrm{T}} S \nabla^2 H \tilde{A}\mathbf{y} - \nabla H^{\mathrm{T}} \tilde{A} S \nabla H - \nabla \tilde{A}(S\nabla H, y, \nabla H)) + \mathcal{O}(\varepsilon h^3),\end{aligned} \tag{2.4.7}$$

where, we have used the fact that $\nabla S(\tilde{A}\mathbf{y}, \nabla H, \nabla H)$ vanishes due to the skew-symmetry of its last two components.

To compute the global error of the Lie-Trotter method, we first assume that both vector fields f_1 and f_2 are one-sided Lipschitz with L_1 and L_2 as their respective one-sided Lipschitz constants. We will also use a result of [8, pg. 37], which states that for a first-order ODE that has two solutions $\mathbf{y}_1(h)$ and $\mathbf{y}_2(h)$, their difference is bounded by the inequality

$$||\mathbf{y}_1(h) - \mathbf{y}_2(h)|| \le \mathrm{e}^{Lh} ||\mathbf{y}_1(0) - \mathbf{y}_2(0)||, \tag{2.4.8}$$

for one-sided Lipschitz constant L. The global error at time $t = t_{n+1}$ is

$$\begin{aligned}\boldsymbol{e}_{n+1}^{[LT]} =& \mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}\\ =& \mathbf{\Phi}_h^{[LT]}(\mathbf{y}(t_n)) + \boldsymbol{\delta}^{[LT]}(\mathbf{y}(t_n)) - \mathbf{\Phi}_h^{[LT]}(\mathbf{y}_n),\end{aligned} \tag{2.4.9}$$

which is computed by decomposing $\mathbf{\Phi}_h^{[LT]} = \varphi_h^{[1]} \circ \varphi_h^{[2]}$ into its flow operators as follows

$$\begin{aligned}||\mathbf{\Phi}_h^{[LT]}(\mathbf{y}(t_n)) - \mathbf{\Phi}_h^{[LT]}(\mathbf{y}_n)|| &= ||\varphi_h^{[1]} \circ \varphi_h^{[2]}\mathbf{y}(t_n) - \varphi_h^{[1]} \circ \varphi_h^{[2]}\mathbf{y}_n||\\ &\le \mathrm{e}^{L_1 h} ||\varphi_h^{[2]}\mathbf{y}(t_n) - \varphi_h^{[2]}\mathbf{y}_n||,\\ &\le \mathrm{e}^{(L_1+L_2)h} ||e_n^{[LT]}||,\end{aligned} \tag{2.4.10}$$

where we have used inequality (2.4.8) twice. If we then assume that the local error is bounded by $\varepsilon h^2 d \ge ||\boldsymbol{\delta}^{[LT]}(y(t))||$, $\forall t \in [0, T]$ for some constant d and for sufficiently small h, then

$$||\boldsymbol{e}_{n+1}^{[LT]}|| \le \mathrm{e}^{(L_1+L_2)h} ||\boldsymbol{e}_n^{[LT]}|| + ||\varepsilon h^2 d||, \tag{2.4.11}$$

this implies that the global error is bounded as follows

$$||\boldsymbol{e}_{n+1}^{[LT]}|| \le \varepsilon h^2 ||d \sum_{i=0}^{n} \left(\mathrm{e}^{h(L_1+L_2)}\right)^i ||, \tag{2.4.12}$$

where $n = T/h$. Taylor expanding the exponential shows the sum is $\mathcal{O}(1/h)$. We can therefore conclude that the global error magnitude is $||\boldsymbol{e}_{n+1}^{[LT]}|| \sim \mathcal{O}(\varepsilon h)$.

The same argument of calculating the local error can be applied to the Strang method and although straightforward, involves the computation of nested commutator brackets. The results, however, are analogous and the local error $\boldsymbol{\delta}^{[S]}(\mathbf{y})$ is presented in appendix 2.A. We find that the local error for the Strang splitting is $||\boldsymbol{\delta}^{[S]}(\mathbf{y})|| \sim \mathcal{O}(\varepsilon h^3)$ at leading order and terms proportional to $\mathcal{O}(\varepsilon^2 h^3)$ can be ignored for $\varepsilon < h$. It then follows that the global error of the Strang method is $|\boldsymbol{e}_{n+1}^{[S]}| \sim \mathcal{O}(\varepsilon h^2)$.

For conventional one-step or multistep methods, such as the Adams-Bashforth two-step method, the perturbed and non-perturbed parts of the vector field are treated together, which means that the method does not see any error advantages due to the small parameter ε. As such, the global error is independent of ε. Using Taylor series it can be shown [8, chapt. III] that the global error of the Adams-Bashforth two-step method is

$$||\boldsymbol{e}_{n+1}^{[AB]}|| \sim \frac{5h^2}{12}||f''(\mathbf{y}_n)|| + \mathcal{O}(h^3), \tag{2.4.13}$$

which is $\mathcal{O}(h^2)$ as opposed to the Strang splitting method which is $\mathcal{O}(\varepsilon h^2)$.

2.4.2 Stiff case

In this section we will examine the error of the splitting method when the vector field $f_2(\mathbf{y})$ is stiff (i.e., when $\gamma^i >> 1$). The differential equation can then be represented by $\dot{\mathbf{y}} = f_1(\mathbf{y}) + \frac{1}{\varepsilon}\tilde{f}_2(\mathbf{y})$. A classical error analysis can be used in the non-stiff regime $h < \varepsilon$, and this shows that the global error behaves according to $\mathcal{O}(h^2/\varepsilon)$. However, in practise, one would like to use a step size $h > \varepsilon$ and in this situation, the flow operator $\varphi_h^{[2]}$ becomes somewhat more difficult to analyse because $||\frac{1}{\varepsilon}\tilde{f}_2(\mathbf{y})|| \geq 1$ and we cannot expand the flow of f_2 in its Taylor series, hence the classical error analysis fails when taking a Taylor expansion about the initial point of this flow operator. Many authors have studied the local error of various first- and second-order splitting methods in this situation using other means, such as singular perturbation theory [2,32] or Lie series [32]. In these studies, it is shown that in the regime $h < \varepsilon$ the local error behaves according to the classical theory; however, for $h > \varepsilon$ different order reduction phenomena are observed depending on the splitting operator. These studies were performed in the context of designing robust splitting methods that use step size control based on local error estimates; however, we are primarily interested in the behaviour of the global error. There has been somewhat less research into how the global error behaves in the stiff case or how the order reduction in the local error evolves when measuring the global error of ODEs. Here, we present numerical experiments relating the local and global error to the step size h and stiffness parameter ε. The results are presented in the next section.

2.5 Numerical Results

Numerical tests were performed for a perturbed and stiff fluid-particle system in the 3D flow field described by equations (2.2.26), (2.2.27) and (2.2.28). Numerical solutions are calculated using the second-order splitting method (SP2) and the second-order Adams-Bashforth two-step method (AB2) for comparison. The perturbed system uses the values $\lambda = 0.1$, $St = 100$, and the maximum eigenvalue of the dissipation matrix A is $\gamma_{max} \approx 0.0806$. The stiff system uses the values $\lambda = 10$, $St = 1$, and $\gamma_{max} \approx 24{,}062$. Both systems use gravity and 3D fluid terms of $g = 0.99$, $\alpha_f = 2\pi$ and $\beta_f = \pi$. The initial conditions for both experiments are $\mathbf{p}_0 = (1,1,1)^{\mathrm{T}}$, $\mathbf{m}_0 = (1,1,1)^{\mathrm{T}}$, $\mathbf{x}_0 = (0,0,0)^{\mathrm{T}}$ and $q_0 = (1/\sqrt{2}, 0, 1/\sqrt{2}, 0)^{\mathrm{T}}$ is the initial rotation quaternion. The error presented in the following figures is

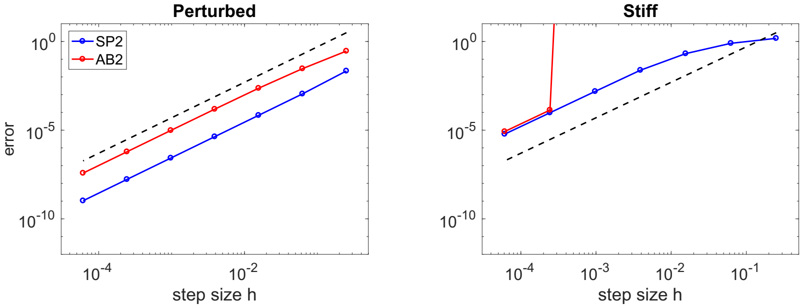
$$\text{error} = \frac{||\mathbf{y}_n - \mathbf{y}(t_n)||}{||\mathbf{y}(t_n)||}, \tag{2.5.1}$$

where $\mathbf{y}(t_n)$ is a reference solution calculated using the classical Runge-Kutta fourth-order method with a comparatively small time-step (e.g., $h = 2^{-14}$).

Figure 2.2 shows the second-order convergence of the SP2 solution compared to the AB2 solution for step sizes $h = 2^{-n}$ for $n = 2,4,6,8,10,12,14$. We observe that both methods achieve the correct order of convergence, however the error of the SP2 solution is significantly lower in the perturbed case compared to the AB2 solution. In the stiff case, the SP2 solution achieves the correct order of convergence for low time-steps and reduced order for larger time-steps. For large time-steps the AB2 solution becomes unstable as denoted by the nearly vertical line.

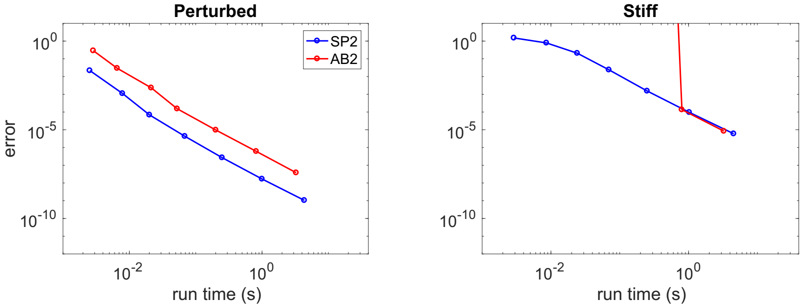
Figure 2.3 shows the relative computational cost of the two methods measured in simulation wall-clock time for MATLAB serial code implementation. We observe that the SP2 method yields numerical solutions that have over an order of magnitude less error for the same computational cost over the one second interval for the perturbed case.

Figure 2.4a shows the local error $||\boldsymbol{\delta}^{[ST]}||$ for varying stiffness parameters ε that are calculated via $\varepsilon = 1/\bar{\gamma}$, where $\bar{\gamma} = ||(\gamma^1, \gamma^2, \ldots, \gamma^{18})||/18$ for eigenvalues γ^i



(a) Perturbed case. **(b)** Stiff case.

Figure 2.2: Second-order convergence of the splitting method (blue line) and the AB2 method (red line).



(a) Perturbed case. **(b)** Stiff case.

Figure 2.3: Simulation wall-clock time of the splitting method (Blue line) and the AB2 method (red line.

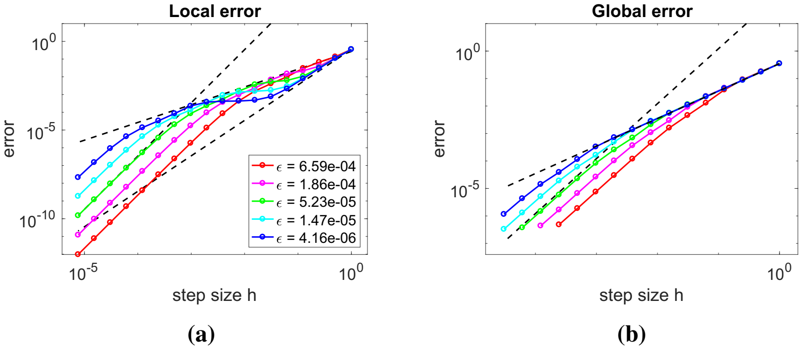


Figure 2.4: Convergence plots for varying stiffness parameters for the local error (a) and global error (b). Order-two and order-one reference lines are plotted on both figures as well as order-three for (a).

of A. Here, we observe the order reduction phenomenon sometimes referred to as the "hump" [7, p. 113] where we see no increase in error when the step size is increased. This usually occurs in the region $\varepsilon < h < \sqrt{\varepsilon}$ as was observed in [32] for the Van der Pol oscillator when the Strang splitting operator used contains the non-stiff flow operator in the middle. In the non-stiff regime, the local error behaves according to classical theory: it is order-three and proportional to $1/\varepsilon$. In the stiff regime, we observe various order reduction phenomena including convergence to an ε-independent low-order line. In addition to the predictions made by [32], we observe that the order is also reduced to about 1.5 in the region just below the "hump". This is most clearly observed by the blue line of figure 2.4a and is again emphasised in figure 2.5. Figure 2.4b presents the corresponding global errors. As expected, we observe that the solutions are of order two and proportional to $1/\varepsilon$ in the non-stiff regime. As the time-step is increased the order converges to an ε-independent order-one line. Although we perform no rigorous error analysis to explain this, our experiments suggest that there is some ε-independent upper bound of the form $u \leq hc(\mathbf{y}_0)$ for some value c that can depend on the initial conditions $\mathbf{y}_0$. This is highlighted by the dashed order-one reference line.

Figure 2.5 presents the orders of the lines in figure 2.4 and the corresponding values of ε by vertical dotted lines. We observe in figure 2.5a that for $h < \varepsilon$, the method has local order three and as the step size increases, we see some strange ε dependent order reduction phenomena. The global order of figure 2.5b shows a similar phenomenon in the transition region, where the lines go from order two to one in the stiff regime.

The reference energy H and dissipation $\dot{H}$ are calculated from equations (2.3.7)

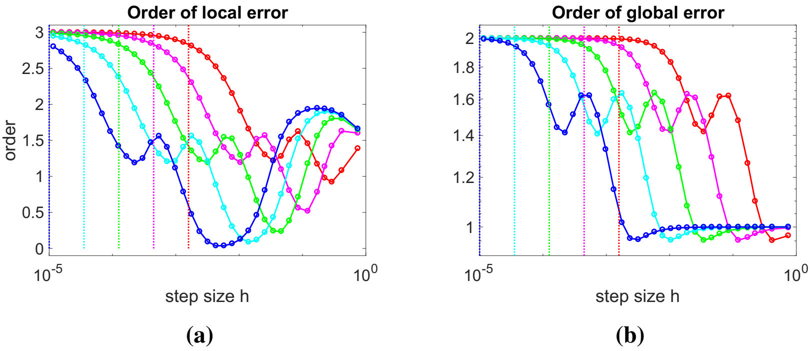
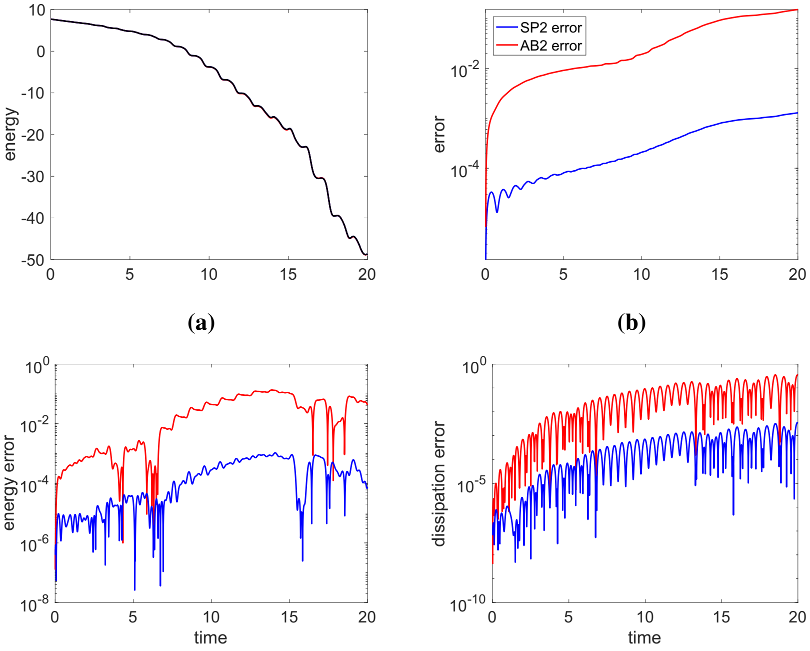


Figure 2.5: The global order for the stiff equation for varying values of ε (same as figure 2.4), which are displayed as vertical dotted lines.

and (2.3.14) using the reference solution and is compared against the numerical energy and dissipation from the SP2 and AB2 solutions in an oscillating shear flow $\mathbf{u}_S$, described in section 2.2.3, over a 20 second time interval with time-step $h = 0.001$. The system uses the same input parameters as the perturbed case in the previous experiment. Figure 2.6a presents the energy of the particle as its dynamics evolves over the 20 second interval. The solution errors are displayed in figure 2.6b, the energy errors are displayed in figure 2.6c and the dissipation errors are displayed in figure 2.6d, in all cases, the SP2 solution errors are approximately two orders of magnitude lower than those of the AB2.



(a) **(b)**

(c) **(d)**

Figure 2.6: The particle energy (a), solution error (b), energy error (c) and dissipation error (d) as functions of time for the splitting solution (blue line) and Adams-Bashforth solution (red line) compared to the reference solution (black line) for a perturbed system.

2.6 Conclusion

We have proposed a splitting method for particle dynamics in viscous flows, obtained by splitting the vector fields of the forced rigid-body dynamics equations into a conservative vector field and a vector field that accounts for the fluid forces. Using backward error analysis, we have shown for perturbed systems, the global error is proportional to $\mathcal{O}(\varepsilon h^2)$ which is an order ε lower than conventional methods. For the stiff case, the splitting method produces solutions that are stable in the unstable regime of the conventional method and retains stability for all $h \leq 1$. Via numerical experiment, we confirm results from the literature [32], on the local error order reduction phenomena for the splitting method. In the non-stiff regime, the global error is observed to behave according to $\mathcal{O}(h^2/\varepsilon)$ and transitions to $\mathcal{O}(h)$ in the stiff regime.

2.7 Acknowledgements

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (No. 691070) as well as the SPIRIT project (No. 231632) under the Research Council of Norway FRIPRO funding scheme. Part of this work was done while visiting the University of Cambridge, UK and La Trobe University, Melbourne, Australia.

Bibliography

[1] A. Tornberg and K. Gustavsson, *A numerical method for simulations of rigid fiber suspensions*, J. Comput. Phys., 215 (2006), p. 172–196.

[2] B. Sportisse, *An analysis of operator splitting techniques in the stiff case*, J. Comput. Phys., 161 (2000), p. 140–168.

[3] H. Brenner, *The Stokes resistance of an arbitrary particle IV: Arbitrary fields of flow*, Chem. Eng Sci., 19 (1964), pp. 703–727.

[4] M. M. W. S. C. Siewert, R.P.J. Kunnen, *Orientation statistics and settling velocity of ellipsoids in decaying turbulence*, Atmospheric research, 142 (2014), pp. 45–56.

[5] E. Celledoni, F. Fassó, N. Säfström, A. Zanna, *The Exact Computation of the Free Rigid Body Motion and Its Use in Splitting Methods*, SIAM J. Sci. Comput., 30(4) (2007), p. 2084–2112.

[6] E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration, Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, second ed., 2006.

[7] E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin Heidelberg, second ed., 1996.

[8] E. Hairer, S.P. Nørsett, G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems*, Springer-Verlag, Berlin Heidelberg, second ed., 1993.

[9] C. Etheir and D. Steinman, *Exact fully 3D Navier-Stokes solutions for benchmarking*, International journal for numerical methods in fluids, 19 (1994), pp. 369–375.

[10] F. G. Fan and G. Ahmadi, *Dispersion of ellipsoidal particle in an isotropic pseudo-turbulent flow field*, ASME J. Fluids Eng., 117 (1995), pp. 154–161.

[11] P. A. F. LUNDELL, L.D. SOĎERBERG, *Fluid mechanics of papermaking*, Annu. Rev. Fluid Mech, 43 (2011), p. 195–217.

[12] G. A. VOTH AND A. SOLDATI, *Anisotropic Particles in Turbulence*, Annu. Rev. Fluid Mech., 49 (2017), p. 249–76.

[13] G. STRANG, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), p. 506–517.

[14] H. BRENNER AND R. COX, *The resistance to a particle of arbitrary shape in translational motion at small Reynolds numbers*, J. Fluid Mech., 17 (1963), p. 561–595.

[15] H. GOLDSTEIN, C. P. POOLE AND J. L. SAFKO, *Classical Mechanics Differential*, Addison-Wesley, second ed., 2001.

[16] H. YOSHIDA, *Construction of higher order symplectic integrators*, Phys. Lett. A., 150 (1990), p. 262–268.

[17] H. ZHANG, G. AHMADI, F. G. FAN, AND J. B. MCLAUGHLIN, *Ellipsoidal particles transport and deposition in turbulent channel flows*, Int. J. Multiphase Flow, 27 (2001), pp. 971–1009.

[18] A. HEYMSFIELD, *Precipitation development in stratiform ice clouds: a microphysical and dynamical study*, J. Atmos. Sci., 34 (1977), p. 67–81.

[19] H.F. TROTTER, *On the product of semi-groups of operators*, Proc. Am. Math. Soc., 10 (1959), p. 545–551.

[20] J.E. MARSDEN, T.S. RATIU, *Introduction to Mechanics and Symmetry. A Basic Exposition of Classical Mechanical Systems*, Springer-Verlag, New York, second ed., 1999.

[21] G. JEFFERY, *The Motion of Ellipsoidal Particles Immersed in a Viscous Fluid*, Proceedings of the Royal Society of London. Series A., 102 (1922), pp. 161–179.

[22] R. V. H. L. SABBAN, *Measurements of pollen grain dispersal in still air and stationary, near homogeneous, isotropic turbulence*, J. Aerosol Sci, 42 (2011), p. 67–82.

[23] M. SHIN AND D. L. KOCH, *Rotational and translational dispersion of fibres in isotropic turbulent flows*, J. Fluid Mech., 540 (2005), p. 143–74.

[24] M. SUZUKI, *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations*, Phys. Lett. A., 146 (1990), p. 319–323.

[25] N. R. CHALLABOTLA, C. NILSEN AND H. I. ANDERSSON, *On rotational dynamics of inertial disks in creeping shear flow*, Phys. Lett. A., 379 (2015), pp. 157–162.

[26] A. OBERBECK, *Ueber stationare flussigkeitsbeweegungen mit berucksichtigung der inneren reibung*, Crelle's J, 81 (1876).

[27] I. M. E. W. P. F. P. ERNI, C. CRAMER, *Continuous flow structuring of anisotropic biopolymer particles*, Adv. Colloid Interface Sci, 150 (2009), pp. 16–26.

[28] P.H. MORTENSEN, H.I. ANDERSSON, J.J.J. GILLISSEN AND B.J. BOERSMA, *Dynamics of prolate ellipsoidal particles in a turbulent channel flow*, Phys. Fluids, 20 (2008).

[29] R. COX, *The steady motion of a particle of arbitrary shape at small Reynolds numbers*, J. Fluid Mech., 23 (1965), p. 625–643.

[30] R. E. KHAYAT AND R. G. COX, *Inertia effects on the motion of long slender bodies*, J. Fluid Mech., 209 (1989), pp. 435–62.

[31] R. I. MCLACHLAN AND G. R. W. QUISPEL, *Splitting methods*, Acta Numer., 11 (2002), pp. 341–434.

[32] R. KOZLOV, A. KVÆRNØ, B. OWREN, *The behaviour of the local error in splitting methods applied to stiff problems* , J. Comput. Phys., 195 (2003), p. 576–593.

[33] R. VAN ZON, J. SCHOFIELD, *Numerical implementation of the exact dynamics of free rigid bodies*, J. Comput. Phys, 225 (2007), p. 145–164.

[34] K. P. R.C. MOFFETT, *In-situ measurements of the mixing state and optical properties of soot with implications for radiative forcing estimates*, PNAS, 106 (2009), pp. 72–77.

[35] J. K. T.J. PEDLEY, *Hydrodynamic phenomena in suspensions of swimming microorganisms*, Annu. Rev. Fluid Mech, 24 (1992), p. 13–58.

Appendix

2.A Local error for Strang splitting

The local error for the Strang operator is given by

$$\begin{aligned}\boldsymbol{\delta}^{[S]}(\mathbf{y}_0) =& \varphi_h(\mathbf{y}_0) - \Phi_h^{[S]}(\mathbf{y}_0)\\ =& h^3(\frac{1}{12}[f_1,[f_1,f_2]] - \frac{1}{24}[f_2,[f_2,f_1]])\mathbf{y}_0 + \mathcal{O}(h^4),\end{aligned} \tag{2.A.1}$$

and can be computed explicitly by inserting equations (2.3.15) and (2.3.16)

$$\begin{aligned}\boldsymbol{\delta}^{[S]}(\mathbf{y}) = \frac{h^3}{12}\big(&-\nabla^2 A(S\nabla H - \frac{1}{2}A\mathbf{y}, S\nabla H, \mathbf{y}, \cdot) - \nabla A(\mathbf{c}_1, \mathbf{y}, \cdot) - \nabla A(2S\nabla H - \frac{1}{2}A\mathbf{y}, S\nabla H, \cdot)\\ &+ \nabla A(S\nabla H, \frac{1}{2}A\mathbf{y}, \cdot) + \nabla A(\mathbf{y}, S\nabla H - \frac{1}{2}A\mathbf{y}, (\cdot)S\nabla^2 H)\\ &- \frac{1}{2}\nabla A(S\nabla H, \mathbf{y}, (\cdot)A) + \nabla A(S\nabla H, \mathbf{y}, (\cdot)\nabla^2 HS)\\ &- \nabla S(S\nabla H - A\mathbf{y}, \nabla H, (\cdot)A) + \nabla S(\mathbf{c}_2, \nabla H, \cdot)\\ &+ \nabla S(S\nabla H - A\mathbf{y}, \nabla^2 HA\mathbf{y}, \cdot) + \nabla S(A\mathbf{y}, \nabla^2 HS\nabla H, \cdot)\\ &- \nabla S(A\mathbf{y}, \nabla H, (\cdot)\nabla SH^2) + AS\nabla^2 H(S\nabla H - A\mathbf{y}) + 2S\nabla^2 HAS\nabla H\\ &- S\nabla^2 HS\nabla^2 HA\mathbf{y} - \frac{1}{2}(S\nabla^2 HA^2\mathbf{y} + A^2 S\nabla H)\big) + \mathcal{O}(h^4),\end{aligned} \tag{2.A.2}$$

where we have used the fact that the matrix S is linear in $\mathbf{y}$ and vectors $\mathbf{c}_1 = \nabla S(S\nabla H - A\mathbf{y}, \nabla H, \cdot) + S\nabla^H(S\nabla H - A\mathbf{y}) + AS\nabla H + \nabla A(S\nabla H, \mathbf{y}, \cdot)$ and $\mathbf{c}_2 = \nabla A(S\nabla H - \frac{1}{2}A\mathbf{y}, \mathbf{y}, \cdot) + A(S\nabla H - \frac{1}{2}A\mathbf{y}) + \frac{2}{h^2}\boldsymbol{\delta}^{[LT]}(\mathbf{y})$ and $A = \varepsilon\tilde{A}$ for the perturbed case.

Computational geometric methods for preferential clustering of particle suspensions

Benjamin K Tapley, Helge I Andersson, Elena Celledoni, Brynjulf Owren

Submitted

Computational geometric methods for preferential clustering of particle suspensions

Abstract. A geometric numerical method for simulating suspensions of spherical and non-spherical particles with Stokes drag is proposed. The method combines divergence-free matrix-valued radial basis function interpolation of the fluid velocity field with a splitting method integrator that preserves the sum of the Lyapunov spectrum while mimicking the centrifuge effect of the exact solution. We discuss how breaking the divergence-free condition in the interpolation step can erroneously affect how the volume of the particulate phase evolves under numerical methods. The methods are tested on suspensions of 10^4 particles evolving in discrete cellular flow field. The results are that the proposed geometric methods generate more accurate and cost-effective particle distributions compared to conventional methods.

3.1 Introduction

Since the influential work of Maxey and Riley [32] in deriving the equations of motion of an inertial spherical particle immersed in viscous flow, there have been a multitude of studies exploring the collective behavior of suspensions of particles. In particular, the remarkable phenomenon of preferential concentration of inertial particles in turbulence has attracted the attention of many authors. This phenomenon, sometimes referred to as the “centrifuge effect", is also attributed to Maxey [30] who showed that particles disperse in regions where the fluid velocity strain rate is low compared to the vorticity. The theoretical mechanisms for particle clustering has since been further explored by means of Lyapunov exponent analysis [5], caustics [52] and perturbative methods to name a few. Sophisticated numerical simulations [44] have also advanced and verified our understanding of this phenomenon for a variety of flows and extended such observations to non-spherical particles [35]. As the need for large-scale simulations increase, the demand for cost-effective numerical methods is growing. However, despite the fact that numerical simulations are so well documented, there have been few studies that explore the extent to which the numerical methods used in simulations accurately reproduce the geometric properties that explain the preferential clustering of particles. In this paper we discuss some features of the equations of motion that influence the preferential concentration of particles and determine to what extent these features can be replicated by well designed numerical methods. In doing so, we propose an efficient numerical algorithm that is designed to replicate these features. The method combines matrix-valued radial basis functions for the divergence-free interpolation of the discrete fluid field with a splitting method that is designed specifically for the equations of motion under study.

Interpolation methods are necessary for simulating suspensions of particles as the flow field is usually generated by a direct numerical simulation of the Navier-Stokes equations and is therefore only available at discrete points in space, meaning that it must be approximated at the location of the particle. To achieve this in a simple and efficient manner many authors use a variant of a tri-polynomial interpolant, for example [7, 12, 14, 37–39, 44, 47, 48, 50]. Previous studies [4, 53] have explored the extent to which these interpolation methods accurately reproduce statistical properties of the turbulent flow field. However, all the interpolation methods considered in the aforementioned references are based on polynomials that create an approximation to the fluid velocity field that is not divergence-free. One major consequence is that the hydrodynamic Stokes force that determines the particle path lines is instead calculated from a non-conservative fluid velocity field. These non-conservative interpolation methods are still used in practice today despite the fact that the theoretical mechanisms that explain the preferential concentration are derived with the assumption of incompresibility. Furthermore, it is often argued (e.g., [47]) that interpolation errors are “averaged out" and it is concluded that one can acheive statistically similar results using a fast low-order interpolation method. This claim is supported by the fact that linear interpolation produces similar statistics to simulations using cubic interpolation [53]. However, neither linear nor cubic interpolation preserves the divergence-free condition of the fluid field and therefore it is not truly understood whether or not errors to the divergence of the fluid field are averaged out in the same way that standard truncation errors are. The implications of these divergence-errors have not been studied in detail, however there is numerical evidence suggesting that breaking this condition can lead to erroneous clustering in PDF methods, first presented in [34] and also in [19]. Divergence free interpolation has been used to good effect in particle-laden flow simulations [17, 18] as well as in other particle simulation problems, such as in geodynamic modelling [49] and magnetospheric physics [29], for example. One of the goals of this study is to explain, from a numerical analysis point of view, the consequences of breaking the divergence-free condition in the flow field. We also show the benefit of divergence-free interpolation using in simulations of suspensions of inertial particles as well as show how these errors affect the numerical time integration.

In addition, we study some numerical integration methods and how their errors affect the preferential concentration of particles. Two popular classes of methods are explicit Runge-Kutta [7, 39] and Adams-Bashforth methods

[12,37,38,48,50]. As the accuracy of the time stepping algorithm is limited by the interpolation error we consider only explicit order one and two methods. Such methods often do a reasonably good job at integrating the ODEs under study as they are efficient and easy to implement. However, the exact solution to the ODEs that govern the dynamics of particles with Stokes drag possess a number of physical features that can be exploited to increase the accuracy of the time stepping methods without increasing its order or cost. Such features include constant contractivity of phase space volume, the centrifuge effect, rigid body motion, linear dissipation and, in some cases, perturbative forces. These features are able to be exploited by a carefully designed splitting method. In this work we propose, as an alternative to Runge-Kutta and Adams-Bashforth methods, a splitting method that is especially designed to reduce the error in the centrifuge effect, which combined with divergence free interpolation techniques allow us to obtain a higher lever of accuracy in the distribution of particles in viscous flows.

3.1.1 Main contributions and summary of paper

We now highlight the main contributions and give a brief outline of the paper. We begin by outlining the equations of motion and the centrifuge effect in section 3.2. In section 3.3 we develop and analyze a contractive splitting method whose flow preserves the sum of the Lyapunov spectrum of the exact solution and show that conventional methods cannot do this. The splitting method is then applied to the equations of motion for spherical particles and the so-called “centrifuge-preserving" methods are presented, which are constructed to minimize the error of the centrifuge effect.

Section 3.4 presents the use and implementation of matrix-valued radial basis function interpolation to construct a divergence-free interpolation of the discrete flow field. We show that a vector field approximated by matrix-valued radial basis functions are compatible with the Stokes equations due to the fact they they are identical to the method of regularized Stokeslets. This results in a more physically realistic approximation to the underlying Navier-Stokes equations.

In section 3.5 we focus our attention to how physical volume of the particle phase Ψ evolves over a small time h. Upon expanding Ψ in h under the exact solution, we recover the centrifuge effect at $O(h^4)$. When expanding Ψ under the numerical solution, we find that errors to the divergence of the fluid velocity field appear at $O(h^2)$, overshadowing the centrifuge effect. However, when a divergence-free interpolation method is used, all the numerical methods under consideration replicate the *qualitative* behavior of the centrifuge effect. That is, physical volumes of particles will contract in regions where the vorticity is lower than the strain rate and vice versa, however, they do so at a slightly erroneous rate. To account for this error, we show that the centrifuge-preserving methods contract physical volume at the same rate as the exact solution to leading order in h, hence also preserving the *quantitative* behavior of the centrifuge-effect.

Section 3.6 is dedicated to simulations of particle suspensions evolving in a discrete cellular flow field where we compare the proposed geometric methods against conventional methods. What we observe is that a computationally inexpensive combination of divergence-free interpolation and centrifuge-preserving splitting methods yield far more accurate spatial distributions of particles compared to standard methods of higher cost. We present many examples where our geometric algorithm produced distributions of particles that are more similar to the “exact" distribution despite having higher error per particle than distributions produced by slow conventional methods. The main conclusion here is that numerical solutions that preserve the sum of the Lyapunov spectrum, the contractivity of phase space volume, the divergence-free condition and the centrifuge effect in simulations is of great benefit.

Section 3.7 is dedicated to conclusions.

3.2 The equations of motion

The translational dynamics of a small particle immersed in a viscous fluid is governed by the rigid body equations with a Stokes force term

$$\dot{\mathbf{v}} = \alpha K(\mathbf{u}(\mathbf{x}) - \mathbf{v}) \tag{3.2.1}$$

$$\dot{\mathbf{x}} = \mathbf{v} \tag{3.2.2}$$

where $\mathbf{u}(\mathbf{x})$ is the fluid velocity at the particle’s location $\mathbf{x}$, $\mathbf{v}$ the velocity, K is a positive definite resistance tensor and $\alpha = 1/St$ is the inverse particle Stokes number, which is a dimensionless measure of particle inertia. Note that unless mentioned we will assume that $\mathbf{u}(\mathbf{x})$ does not explicitly depend on t. Doing so improves the readability and presentation of the paper and does not affect the forthcoming results.

For spherical particles, $K = I$ is the identity, the rotational variables are constant and the above ODEs uniquely specify the dynamics of each particle. For non-spherical particles, the resistance tensor $K = Q^T K_b Q$, where K_b is the diagonal positive definite body frame resistance tensor and $Q \in SO(3)$ is a rotation matrix that transforms a vector in the body frame to one in the inertial frame. The angular velocity $\boldsymbol{\omega}$ evolves via

$$J\dot{\boldsymbol{\omega}} = J\boldsymbol{\omega} \times \boldsymbol{\omega} - \mathbf{T}, \tag{3.2.3}$$

where J is the diagonal body frame moment of inertia tensor and $\mathbf{T}$ is the hydrodynamic torque. The rotation matrix Q is calculated by solving the matrix ODE

$$\dot{Q} = Q\widehat{\boldsymbol{\omega}}, \tag{3.2.4}$$

where $\widehat{\cdot} : \mathbb{R}^3 \to \mathfrak{so}(3)$ is defined by

$$\begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \mapsto \widehat{\boldsymbol{\omega}} = \begin{pmatrix} 0 & -\omega_1 & \omega_2 \\ \omega_1 & 0 & -\omega_3 \\ -\omega_2 & \omega_3 & 0 \end{pmatrix}, \tag{3.2.5}$$

such that $\widehat{\boldsymbol{\omega}}\mathbf{v} = \boldsymbol{\omega} \times \mathbf{v}$. The expressions for K_b and $\mathbf{T}$ for spheroidal particles are given in 5.3.

3.2.1 The centrifuge effect

Here we will outline the centrifuge effect of the particle equations of motion, which is one of the mechanisms for particle clustering that is referred to throughout the paper. In [30], Maxey assumes $\alpha \gg 1$ and expands the the spherical particle ODEs (3.2.1) and (3.2.2) in powers of α^{-1} to derive a first-order ODE expression for $\mathbf{x}$

$$\dot{\mathbf{x}} = \mathbf{u}(\mathbf{x}) - \alpha^{-1}\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u}\right) + O(\alpha^{-2}) \tag{3.2.6}$$

where we have ignored the effect of gravity. Taking the divergence gives

$$\nabla \cdot \mathbf{v} = \frac{\partial u_i}{\partial x_i} - \frac{1}{\alpha}\left(\frac{\partial}{\partial t}\frac{\partial u_i}{\partial x_i} + \frac{\partial u_i}{\partial x_j}\frac{\partial u_j}{\partial x_i} + u_i \frac{\partial}{\partial x_i}\frac{\partial u_j}{\partial x_j}\right) + O(\alpha^{-2}) \tag{3.2.7}$$

where there is an implied summation over repeated indices, which is the convention that is assumed throughout the paper. Assuming that the fluid field is divergence-free, we arrive at the familiar relationship between the fluid field rate of strain, rate of rotation and the divergence of the particle velocity field

$$\nabla \cdot \mathbf{v} = -\frac{1}{\alpha}\frac{\partial u_i}{\partial x_j}\frac{\partial u_j}{\partial x_i} = -\frac{1}{\alpha}\left(\|S\|_F^2 - \|\Omega\|_F^2\right) + O(\alpha^{-2}) \tag{3.2.8}$$

where the rate of strain and rotation tensors S and Ω are given by

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \quad \text{and} \quad \Omega_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}\right) \tag{3.2.9}$$

and $\|\cdot\|_F$ is the Frobenius matrix norm. In other words, the divergence of the particle velocity field $\nabla\cdot\mathbf{v}$ is positive when the vorticity is large compared to the strain rate tensor meaning that the particulate phase disperses in these regions. Conversely, particles concentrate in regions where the strain rate is large compared to the vorticity. This phenomenon is the “centrifuge effect" of the exact solution to (3.2.1) and (3.2.2).

Finally, we remark that while the centrifuge effect was derived for spherical particles, one can make similar observations for non-spherical particles. In this scenario, the resistance tensor can be decomposed into a spherical part and a non-spherical part, e.g., for a spheroidal particle with rotational symmetry (see 5.3) we can write $K_b = aI + b\boldsymbol{e}_z\boldsymbol{e}_z^T$, where $\boldsymbol{e}_z = (0,0,1)^T$ and $b \to a$ in the spherical limit. In other words, the centrifuge effect still plays a central role in the preferential clustering of non-spherical particles in addition to the non-spherical effects due to $\boldsymbol{b}\boldsymbol{e}_z\boldsymbol{e}_z^T$ term in the resistance tensor.

3.3 Numerical integration of dissipative vector fields

The dynamics of small inertial particles (both spherical and non-spherical) can be modeled as the flow of a vector field with linear dissipation. Such vector fields arise due to the fact that for low Reynolds number flow the drag forces are linear in the slip velocity, for example the Stokes drag force for small ellipsoids, spheres or rigid slender particles [3]. We begin this section with a discussion of such linearly dissipative vector fields and their contractive properties of phase space volume. We then discuss the application of some conventional explicit methods for integrating such ODEs. In particular, we show that conventional methods cannot preserve the contractivity of phase space volume. A splitting scheme is then shown to preserve the exact contractivity of phase space volume. The section concludes with the application of the splitting scheme to spherical particles.

3.3.1 Linearly dissipative vector fields and contractivity of phase space volume

A linearly dissipative vector field in n dimensions is given in general by

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) - A\mathbf{y} \tag{3.3.1}$$

where A is a positive definite matrix and $\mathbf{f}(\mathbf{y})$ is volume preserving, that is, it satisfies $\nabla \cdot \mathbf{f}(\mathbf{y}) = 0$ (e.g., any Hamiltonian vector field). Note that the ODEs of both non-spherical and spherical particles can be cast in this form, where $\mathbf{f}(\mathbf{y})$ represents the free rigid-body vector field plus the conservative part of the Stokes force and $-A\mathbf{y}$ represents the dissipative part of the Stokes force.

The quantitative behavior of particle clustering can be explained in part by analyzing the Lyapunov exponents λ_i of the ODE, see for example [5]. It is therefore desirable that the numerical solution of the ODE reproduces similar Lyapunov exponent characteristics. Whilst there do not currently exist numerical methods that preserve individual Lyapunov exponents we can however construct a numerical method that preserves the sum of the Lyapunov spectrum $\sum_{i=1}^{n} \lambda_i$. From a backward error analysis point of view, a numerical method that preserves the Lyapunov spectrum is one that is the exact solution to an ODE with the same Lyapunov spectrum sum as the ODE being solved. For the equations of motion for spherical particles (equations (3.2.1) and (3.2.2)), the sum of the first three Lyapunov exponents characterizes the divergence of the velocity field and the sum of the spatial Lyapunov exponents characterizes the rate at which particle clouds contract or expand [17]. Generally speaking, the sum of the Lyapunov spectrum describes the rate at which phase space volume exponentially contracts or expands [21]. That is, by letting $\mathbf{y}(t)$ denote the exact solution of (3.3.1) with initial conditions $\mathbf{y}(0) = \mathbf{y}_0$ and

$$Y = \det\left(\frac{\partial \mathbf{y}(t)}{\partial \mathbf{y}_0}\right) \tag{3.3.2}$$

the n dimensional phase space volume, then

$$Y = \prod_{i=1}^{n} e^{t\lambda_i}. \tag{3.3.3}$$

It is also known that linearly dissipative systems contract phase space volume at a constant rate, as we will now show. Taking the Jacobian of $\mathbf{y}(t)$ with respect to $\mathbf{y}_0$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathbf{y}(t)}{\partial \mathbf{y}_0} = (\mathbf{f}' - A)\frac{\partial \mathbf{y}(t)}{\partial \mathbf{y}_0}. \tag{3.3.4}$$

We now recall Jacobi's formula, which relates the derivative of the determinant of a square matrix $M(t)$ by the following

$$\frac{\mathrm{d}}{\mathrm{d}t}\det(M(t)) = \det(M(t))\,\mathrm{Tr}\left(M(t)^{-1}\frac{\mathrm{d}}{\mathrm{d}t}M(t)\right). \tag{3.3.5}$$

Differentiating (3.3.2) with respect to time and applying (3.3.5) gives

$$\frac{\mathrm{d}}{\mathrm{d}t}Y = -Y\,\mathrm{Tr}\big(A\big), \tag{3.3.6}$$

as $\mathbf{f}'$ has zero trace. This is solved by

$$Y = e^{-t\,\mathrm{Tr}(A)}. \tag{3.3.7}$$

By equating this with (3.3.3) we obtain the relation

$$\sum_{i=1}^{n}\lambda_i = -\mathrm{Tr}(A). \tag{3.3.8}$$

As the trace of A is by definition positive, the phase space volume Y is strictly monotonically contracting in time. Equation (3.3.8) implies that a numerical integration method that preserves phase space volume Y also preserves the sum of the Lyapunov exponents of the underlying ODE. It is therefore logical to imply that a numerical flow of (3.3.1) that preserves the contractivity of phase space volume will better reproduce the clustering properties of the exact solution than one that doesn't. The rest of this section is dedicated to analysing to what extent some common numerical integration methods for particle dynamics can preserve this constant contractivity of phase space volume.

3.3.2 Preservation of the contractivity of phase space volume by numerical methods

Denote by Φ_h a numerical method for solving (3.3.1) such that $\Phi_h(\mathbf{y}_0) \approx \mathbf{y}(h)$ for time step $h \ll 1$. For Φ_h to be called *contractivity preserving* when applied to (3.3.1), we require that $\det\left(\frac{\partial\Phi_h(y)}{\partial y}\right) = e^{-h\mathrm{Tr}(A)}$ [20]. It is known that no standard methods (e.g., one with a B-series [15]) can preserve phase space volume for divergence free vector fields [23]. The same is also expected when it comes to preserving contractivity of phase space volume for dissipative vector fields [33]. Instead, a weaker requirement is that they contract phase space volume when the exact solution does so, that is, $\det\left(\frac{\partial\Phi_h(y)}{\partial y}\right) < 1$ when applied to (3.3.1). Such a numerical method that possesses this property is called *contractive*.

We now consider some popular numerical methods for particle dynamics. We consider explicit methods used in the literature, namely, explicit Runge-Kutta methods, Adams-Bashforth methods and a splitting method scheme.

Runge-Kutta methods and phase volume contractivity

Take an order-p Runge-Kutta method $\Phi_h^{[RK]}(\mathbf{y}_0)$ with stability function $R(z)$ applied to an ODE of the form (3.3.1) with linear $\mathbf{f}(\mathbf{y})$, say

$$\dot{\mathbf{y}} = B\mathbf{y} - A\mathbf{y}, \tag{3.3.9}$$

where B is a square and traceless matrix. Then the numerical solution by the Runge-Kutta method is given by

$$\Phi_h^{[RK]}(\mathbf{y}_0) = R(h(B-A))\mathbf{y}_0. \tag{3.3.10}$$

Recall that $R(z)$ is an order-p Padé approximation to the exponential function. We therefore have

$$\Phi_h^{[RK]}(\mathbf{y}_0) = \exp\big(h(B-A)\big)\mathbf{y}_0 + O(h^{p+1}) \tag{3.3.11}$$

which means that over one time-step, the phase space volume contracts via

$$\det\left(\frac{\partial\Phi_h^{[RK]}(\mathbf{y}_0)}{\partial\mathbf{y}_0}\right) = e^{-h\mathrm{Tr}(A)} + O(h^{p+1}). \tag{3.3.12}$$

That is, for such a linear system, a Runge-Kutta method will only preserve the phase volume contractivity up to the order of the method. So for a non-linear dissipative ODE of the form (3.3.1), one can hardly expect a Runge-Kutta method to preserve phase space volume exactly. In fact, due to this error explicit Runge-Kutta methods usually have a time-step restriction on h to even be contractive at all [20]. This is illustrated by the following examples of some low order explicit Runge-Kutta methods applied to equations (3.2.1) and (3.2.2).

Example 1. We apply the forward Euler method $\Phi_h^{[FE]}$ to the ODEs (3.2.1) and (3.2.2). Note that these ODEs are non-linear due to $\mathbf{u}(\mathbf{x})$. Setting $\mathbf{y} := (\mathbf{v},\mathbf{x})$, we have for the contractivity of phase space volume under the forward Euler method

$$\det\left(\frac{\partial\Phi_h^{[FE]}(\mathbf{y}_0)}{\partial\mathbf{y}_0}\right) = 1 - 3\alpha h + \alpha h^2\left(3\alpha - \frac{\partial u_i}{\partial x_i}\right) + O(h^3), \tag{3.3.13}$$

which is an order one approximation to the exact contractivity

$$\det\left(\frac{\partial \mathbf{y}(h)}{\partial \mathbf{y}_0}\right) = e^{-3\alpha h}. \tag{3.3.14}$$

The Forward Euler method must therefore satisfy the following time-step restriction for it to be contractive

$$h \lessapprox \frac{3}{3\alpha - \frac{\partial u_i}{\partial x_i}}. \tag{3.3.15}$$

Violating this restriction means that the forward Euler method will expand phase space volume despite the ODE dictating that it is always contracting. Furthermore, it can be seen that large values of $|\frac{\partial u_i}{\partial x_i}|$ will place further restrictions on the size of h.

▶

Example 2. Consider the following second order explicit Runge-Kutta method

$$\Phi_h^{[RK]}(\mathbf{y}_0) = \mathbf{y}_0 + h\big((1 - \tfrac{1}{2\theta})\mathbf{f}(\mathbf{y}_0) + \tfrac{1}{2\theta}\mathbf{f}(\mathbf{y}_0 + \theta h \mathbf{f}(\mathbf{y}_0)\big), \tag{3.3.16}$$

where $\theta = \frac{1}{2}, \frac{2}{3}$ and 1 correspond to the explicit midpoint method, Ralston's method and Heun's method, respectively. We apply this to the ODEs (3.2.1) and (3.2.2). Setting $\mathbf{y} := (\mathbf{v}, \mathbf{x})$, we have for the contractivity of phase space volume under $\Phi_h^{[RK]}$

$$\det\left(\frac{\partial \Phi_h^{[RK]}(\mathbf{y}_0)}{\partial \mathbf{y}_0}\right) = 1 - 3\alpha h + 9\alpha^2 \frac{h^2}{2!} \tag{3.3.17}$$

$$+ \left(3\alpha(\theta - 1)\frac{\partial^2 u_i}{\partial x_i \partial x_j} v_j + 3\alpha^2 \frac{\partial u_i}{\partial x_i} - 24\alpha^3\right)\frac{h^3}{3!} + O(h^4) \tag{3.3.18}$$

which is an order two approximation to the exact contractivity (3.3.14), which is expected for an order two method. The time-step h must be chosen small enough such that the $O(h^3)$ error term does not violate the contractivity condition. Violating this restriction means that the method will expand phase space volume despite the ODE dictating that it is always contracting. Furthermore, it can be seen that large values of $|\frac{\partial u_i}{\partial x_i}|$ and v_i will place further restrictions on the size of h. ▶

We remark that we can make similar observations for the above methods applied to the ODEs for non-spherical particles. That is, the phase space volume is conserved only to the order of the method.

Multi-step methods and phase volume contractivity

Another popular numerical method used for particle dynamics are multi-step methods. Consider the explicit k-step Adams-Bashforth methods. Such methods are of global order-k and can be seen as a map $\Phi_h^{[AB]} : (\mathbf{y}_0, ..., \mathbf{y}_{k-1}) \rightarrow (\mathbf{y}_1, ..., \mathbf{y}_k)$ such that

$$\mathbf{y}_k = \mathbf{y}_{k-1} + h \sum_{i=0}^{k-1} b_i \mathbf{f}(\mathbf{y}_i), \tag{3.3.19}$$

$$\mathbf{y}_i = \mathbf{y}_{i-1}, \quad \text{for} \quad i = 1, ..., k-1 \tag{3.3.20}$$

where the coefficients b_i satisfy $\sum_{i=0}^{k-1} b_i = 1$. That is, $\Phi_h^{[AB]}$ takes a point in a kn dimensional phase space to another point in the same space. Due to this and the fact that the initial vectors in the domain $\mathbf{y}_i$ for $i = 0, ..., k-1$ are independent of one another, it's less clear how to define the notion of numerical phase space volume that relates to that of the underlying ODE. However, in practice these initial vectors $\mathbf{y}_k$ for $i = 0, ..., k-1$, are usually computed by an order-k one step method, for example a Runge-Kutta method. Therefore, the vectors $\mathbf{y}_i$ depend on the vectors $\mathbf{y}_j$ for $j < i$. This implies that each $\mathbf{y}_i$ has the same series expansion as the exact solution up to $O(h^k)$. While a detailed analysis of multi-step methods and the preservation of phase volume lie outside the scope of the paper, we can illustrate this concept by the following example, which considers the phase volume properties of a second order Adams-Bashforth using a second order Runge-Kutta method to compute the initial vectors.

Example 3. Consider the second-order Adams-Bashforth method $\Phi_h^{[AB]} : (\mathbf{y}_1, \mathbf{y}_0) \rightarrow (\mathbf{y}_2, \mathbf{y}_1)$ where

$$\mathbf{y}_2 = \mathbf{y}_1 + \frac{h}{2}\big(3\mathbf{f}(\mathbf{y}_1) - \mathbf{f}(\mathbf{y}_0)\big) \tag{3.3.21}$$

Now define $\mathbf{y}_1 = \Phi_h^{[RK]}(\mathbf{y}_0)$ in the domain by the second order Runge-Kutta method (3.3.16). Applying $\Phi_h^{[AB]}$ to the ODEs (3.2.1) and (3.2.2) and taking the Jacobian determinant of $\mathbf{y}_2$ with respect to $\mathbf{y}_0$ gives

$$\det\left(\frac{\partial \mathbf{y}_2}{\partial \mathbf{y}_0}\right) = 1 - 6\alpha h + 36\alpha^2 \frac{h^2}{2!} \tag{3.3.22}$$

$$+ \left(\alpha\left(3\theta - \frac{21}{2}\right)\frac{\partial^2 u_i}{\partial x_i \partial x_j} v_j + \frac{21\alpha^2}{2}\frac{\partial u_i}{\partial x_i} - 24\alpha^3\right)\frac{h^3}{3!} + O(h^4) \tag{3.3.23}$$

which is an $O(h^2)$ approximation to the exact contractivity (3.3.14). Note that here we have taken the the contractivity over two time-steps $2h$. Like the previous example, we observe that the contractivity is affected by non-zero values of $\frac{\partial u_i}{\partial x_i}$ and v_i. ▶

3.3.3 A splitting scheme that preserves the contractivity of phase space volume

We now analyze the splitting method based on the following splitting of equation (3.3.1)

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}) - \mathbf{b}(\mathbf{y}) \quad \text{and} \quad \dot{\mathbf{y}} = -A\mathbf{y} + \mathbf{b}(\mathbf{y}) \tag{3.3.24}$$

where $\mathbf{b}(\mathbf{y})$ is any vector that is constant along the flow of the second vector field. A similar splitting was proposed in [45] for non-spherical particle dynamics. Denote their exact flow operators by $\psi_h^{[1]}$ and $\psi_h^{[2]}$, respectively. In the context of small particles immersed in a viscous fluid, the first vector field represents the free rigid body equations and the second is due to the Stokes viscous drag forces. The free rigid body vector field can be solved exactly. That is, by a forward Euler step for the spherical case or otherwise using trigonometric or Jacobi elliptic functions depending whether or not the body is axially symmetric [11]. Due to the existence of an exact solution, we immediately have volume preservation

$$\left| \frac{\partial \psi_h^{[1]}(\mathbf{y}_0)}{\partial \mathbf{y}_0} \right| = 1. \tag{3.3.25}$$

The second vector field is solved by the variation of parameters formula

$$\psi_h^{[2]}(\mathbf{y}_0) = e^{-hA}(\mathbf{y}_0 + A^{-1}\mathbf{b}) + A^{-1}\mathbf{b}. \tag{3.3.26}$$

Taking the Jacobian determinant gives

$$\left| \frac{\partial \psi_h^{[2]}(\mathbf{y}_0)}{\partial \mathbf{y}_0} \right| = e^{-h\mathrm{Tr}(A)}, \tag{3.3.27}$$

which is consistent with the exact solution (3.3.7). As the Jacobian of the composition of two or more maps is the product of the Jacobians of the maps, any splitting method based on the alternating compositions of the flows $\psi_h^{[1]}$ and $\psi_h^{[2]}$ will be contractivity preserving.

In forthcoming numerical experiments, we will consider only order one and two methods including the order one Lie-Trotter method

$$\Phi_h^{[LT]} = \psi_h^{[1]} \circ \psi_h^{[2]}, \tag{3.3.28}$$

and the order two Strang method

$$\Phi_h^{[SS]} = \Phi_{\frac{h}{2}}^{[LT]} \circ \Phi_{\frac{h}{2}}^{[LT]*}. \tag{3.3.29}$$

Here, we denote by $\Phi_h^{[LT]*} = \psi_h^{[2]} \circ \psi_h^{[1]}$ the conjugate of the $\Phi_h^{[LT]}$.

3.3.4 Application to spherical particle dynamics and the centrifuge-preserving methods

To construct a contractivity preserving splitting method for spherical particles, we split the ODEs (3.2.1) and (3.2.2) in to the following two vector fields

$$\begin{pmatrix} \dot{\mathbf{v}} \\ \dot{\mathbf{x}} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{v} \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} \dot{\mathbf{v}} \\ \dot{\mathbf{x}} \end{pmatrix} = \begin{pmatrix} \alpha(\mathbf{u}(\mathbf{x}) - \mathbf{v}) \\ 0 \end{pmatrix}. \tag{3.3.30}$$

Their exact flow operators are

$$\psi_h^{[1]} \begin{pmatrix} \mathbf{v} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{x} + h\mathbf{v} \end{pmatrix} \quad \text{and} \quad \psi_h^{[2]} \begin{pmatrix} \mathbf{v} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} e^{-\alpha h}(\mathbf{v} - \mathbf{u}(\mathbf{x})) + \mathbf{u}(\mathbf{x}) \\ \mathbf{x} \end{pmatrix}. \tag{3.3.31}$$

Indeed, letting $\mathbf{y}_0 = (\mathbf{v}_0^T, \mathbf{x}_0^T)^T$ we see that

$$\left| \frac{\partial \psi_h^{[1]}(\mathbf{y}_0)}{\partial \mathbf{y}_0} \right| = 1 \quad \text{and} \quad \left| \frac{\partial \psi_h^{[1]}(\mathbf{y}_0)}{\partial \mathbf{y}_0} \right| = e^{-3\alpha h} \tag{3.3.32}$$

hence any composition of the above flows will preserve contractivity. For the construction of the splitting method for non-spherical particle dynamics, we refer the reader to [45].

A draw back of splitting methods is that composing methods of order higher than two requires the use of negative time steps. As the ODEs in question are dissipative, such higher order methods would therefore require strict time step restrictions, which is preferably avoided. The idea behind geometric numerical integrators is that preserving relevant properties of the exact solution upon discretisation can lead to better qualitative and long time numerical solutions. With this in mind, instead of improving the accuracy of the method in a conventional sense by increasing the order, we propose as an alternative the following composition methods

$$\Phi_h^{[CP_1]} = \Phi^{[LT]}_{(1-\frac{\sqrt{6}}{6})h} \circ \Phi^{[LT]*}_{\frac{\sqrt{6}}{6}h} \tag{3.3.33}$$

$$\Phi_h^{[CP_2]} = \Phi^{[LT]}_{\frac{3h}{12}} \circ \Phi^{[LT]*}_{\frac{5h}{12}} \circ \Phi^{[LT]}_{\frac{4h}{12}} \tag{3.3.34}$$

which are order one and order two methods, respectively. We propose that the above splitting methods are particularly well suited to calculation of particle dynamics as their numerical solution preserves the centrifuge effect of the exact solution when considering the contraction of physical volume of the particle field. We will therefore refer to the methods (3.3.33) and (3.3.34) as the “centrifuge-preserving" methods. This favorable property is discussed in more detail in section 3.5.2. In section 3.6.4 we show through numerical simulations that integrators possessing this property predict more accurately the spatial distribution of particles (both spherical and non-spherical) compared to methods without this property.

3.4 Divergence-free interpolation with matrix-valued radial basis functions

To construct a divergence-free approximation to the discrete fluid field, we propose using matrix-valued radial basis functions (MRBFs). In this section, we will give a brief outline on their use and implementation. We then further motivate their use by showing that the interpolated vector field generated by MRBFs is a solution to the Stokes equation.

The interpolation problem is as follows. Given a set of vector-valued data $\{\mathbf{u}_i, \mathbf{x}_i\}_{i=1}^{n^3}$ generated by an accurate direct numerical simulation to the Navier-Stokes equations, construct a divergence-free vector field that locally interpolates the data. In our context, $\mathbf{u}_i = \mathbf{u}(\mathbf{x}_i)$ is the fluid velocity vector at the grid node located at $\mathbf{x}_i = (x_i, y_i, z_i)^{\mathrm{T}}$. When implementing a polynomial interpolation method, one usually chooses the $n \times n \times n$ cube of data points neighboring the particle, where $n = 2, 3$ or 4. This is because polynomial interpolation of degree $n-1$ requires n data points in each dimension to specify a unique interpolating polynomial. MRBFs are not restricted by this particular choice of data points, however to keep the interpolation methods comparable we will adopt this convention. The MRBF interpolating vector field $\mathbf{s}(\mathbf{x})$ is then constructed by

$$\mathbf{s}(\mathbf{x}) = \sum_{i=1}^{n^3} \Theta_i(\mathbf{x})\mathbf{c}_i, \tag{3.4.1}$$

where

$$\Theta_i(\mathbf{x}) = (\nabla\nabla^T - \nabla^2 I)\theta(r_i(\mathbf{x})) \in \mathbb{R}^{3\times 3} \tag{3.4.2}$$

is called an MRBF, $\theta(r_i(\mathbf{x}))$ is a (scalar-valued) radial basis function, $r_i(\mathbf{x}) = \|\mathbf{x}_i - \mathbf{x}\|$ is the distance from the point $\mathbf{x}_i$ and I is the identity matrix in three dimensions. The n^3 vector-valued coefficients $\mathbf{c}_i \in \mathbb{R}^3$ are chosen such that $\mathbf{s}(\mathbf{x}_i) = \mathbf{u}(\mathbf{x}_i)$, which amounts to solving the following $3n^3$ dimensional linear system

$$\begin{pmatrix} \Theta_1(\mathbf{x}_1) & \cdots & \Theta_n(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \Theta_1(\mathbf{x}_n) & \cdots & \Theta_n(\mathbf{x}_n) \end{pmatrix} \begin{pmatrix} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_n \end{pmatrix} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} \in \mathbb{R}^{3n^3}. \tag{3.4.3}$$

The particular RBF we use in the forthcoming experiments is the Gaussian $\theta(r) = \exp(-\epsilon^2 r^2)$, where ϵ is some user defined parameter that controls the flatness of the RBF. In general, one should choose ϵ as small as possible as this leads to less interpolation error, although more ill conditioned systems.

It can be easily seen that $\mathbf{s}(\mathbf{x})$ is divergence free. Using the double curl identity in $\mathbb{R}^3$ we have

$$\nabla \cdot \mathbf{s} = \sum_{i=1}^{3n} \nabla \cdot \left((\nabla\nabla^T - \nabla^2 I)\theta(r_i)\mathbf{c}_i \right) = \sum_{i=1}^{3n} \nabla \cdot \left(\nabla \times \nabla \times \left(\theta(r_i)\mathbf{c}_i\right) \right) = 0. \tag{3.4.4}$$

Finally, we list some advantages of MRBF interpolation over standard tri-polynomial interpolation: (1) they work equally well on scattered data points, meaning that they are just as well suited to interpolate data generated by a direct numerical simulation involving complex geometries on unstructured grids; (2) they have faster convergence of their derivatives [9, 51], compared to tri-polynomial interpolation [10]; and (3), they are compatible with the Stokes equations, meaning that they construct a more physically realistic fluid field for fluid simulations. We will discuss point (3) in the next section.

3.4.1 MRBFs as regularised Stokeslet solutions to the Stokes equations

In addition to the fact that the underlying flow field should be divergence-free, we are given extra knowledge that can be exploited; namely that the data is a numerical solution to the incompressible Navier-Stokes equations

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} \right) - \mu \nabla^2 \mathbf{u} \;=\; -\nabla p + \mathbf{F} \quad \text{and} \quad \nabla \cdot \mathbf{u} \;=\; 0. \tag{3.4.5}$$

We are only interpolating in space and hence approximating steady-state solutions to (3.4.5) and as the grid-spacing Δx is comparable to the smallest length scales of the flow (e.g., the Kolmogorov scale for turbulent flows), the Reynolds number is small and the non-linear terms of equations (3.4.5) can be ignored. Under the above assumptions, a good approximation for the local flow in a grid-cell can be given by the steady Stokes equations, which reads

$$\mu \nabla^2 \mathbf{u} - \nabla p \;=\; -\mathbf{F} \quad \text{and} \quad \nabla \cdot \mathbf{u} \;=\; 0, \tag{3.4.6}$$

where we have set $\mu = 1$. Cortez [13] presents what's called the regularised Stokeslet solution to the Stokes equation, which is an approximation to Green's function of the Stokes equation for body force $\mathbf{F} = \phi_\epsilon(\mathbf{x})\,\mathbf{f}_0$, where $\mathbf{f}_0 \in \mathbb{R}^3$ is constant. Here, $\phi_\epsilon(\mathbf{x})$ is the so-called "blob" function, which is a radially symmetric smooth approximation of the Dirac delta function $\delta(\mathbf{x})$ that decays to zero at infinity whilst satisfying

$$\int \phi_\epsilon(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1 \quad \text{and} \quad \lim_{\epsilon\to 0}(\phi_\epsilon(\mathbf{x})) = \delta(\mathbf{x}). \tag{3.4.7}$$

Now define the functions $G_\epsilon(\mathbf{x})$ and $B_\epsilon(\mathbf{x})$ as the solutions to

$$\nabla^2 G_\epsilon(\mathbf{x}) = \phi_\epsilon(\mathbf{x}) \quad \text{and} \quad \nabla^2 B_\epsilon(\mathbf{x}) = G_\epsilon(\mathbf{x}), \tag{3.4.8}$$

which are smooth approximations to Green's function and the biharmonic equation $\nabla^4 B(\mathbf{x}) = \delta(\mathbf{x})$, respectively. Then Cortez's regularised Stokeslet solution reads

$$\mathbf{u}_\epsilon(\mathbf{x}) = (\mathbf{f}_0 \cdot \nabla)\nabla B_\epsilon(\mathbf{x}) - \mathbf{f}_0 G_\epsilon(\mathbf{x}) \tag{3.4.9}$$

with pressure term

$$p_\epsilon(\mathbf{x}) = \mathbf{f}_0 \cdot \nabla G_\epsilon(\mathbf{x}). \tag{3.4.10}$$

Using the definition for $G_\epsilon(\mathbf{x})$, we can rewrite the regularised Stokeslet (3.4.9) as

$$\mathbf{u}_\epsilon(\mathbf{x}) = (\nabla\nabla^T - \nabla^2 I)(B_\epsilon(\mathbf{x})\mathbf{f}_0), \tag{3.4.11}$$

which is identical to an MRBF element if we can identify $B_\epsilon(\mathbf{x})$ with a positive-definite RBF $\psi(||\mathbf{x}||)$ (e.g., the Gaussian $\psi(||\mathbf{x}||) = \exp\left(-\epsilon^2||\mathbf{x}||^2\right)$) and the force vectors are identified with the interpolation coefficient vectors $\mathbf{c}_i$ from equation (3.4.1). This means that a vector field that is constructed from a linear combination of MRBFs, (i.e., equation (3.4.1)) corresponds to a linear combination of regularised Stokeslet solutions, with force $\mathbf{f}_i = \mu\mathbf{c}_i$. This leads to the following solution to the Stokes equation, now written in terms of MRBFs

$$\mathbf{s}(\mathbf{x}) = \mathbf{u}_\epsilon(\mathbf{x}) = \sum_{i=0}^{N} \Theta_i(\mathbf{x})\mathbf{f}_i, \text{ and } p_\epsilon(\mathbf{x}) = \sum_{i=0}^{N} \mathbf{f}_i \cdot \nabla(\nabla^2\theta(r_i)). \tag{3.4.12}$$

One implication of this is that the interpolated background fluid field is related to the gradient of a scalar pressure field, when MRBF interpolation is used. The benefit of this can be illustrated by inserting equation (3.4.5) into equation (3.2.6) to derive an expression for $\nabla \cdot \mathbf{v}(\mathbf{x})$ in terms of the pressure field [16]

$$\nabla \cdot \mathbf{v}(\mathbf{x}) = \nabla \cdot \mathbf{u}(\mathbf{x}) + \alpha^{-1}(\nabla^2 p_\epsilon(\mathbf{x})) + O(\alpha^{-2}). \tag{3.4.13}$$

This equation tells us that the pressure field is also related to the preferential concentration of particles. Moreover, this suggests that particles cluster in regions of maximum pressure ($\nabla^2 p_\epsilon(\mathbf{x}) < 0$) [16]. Indeed, in [28], numerical evidence is found to support the correlation between the Laplacian of the pressure field $\nabla^2 p_\epsilon(\mathbf{x})$ and the spatial distribution of the particles. However, if the background fluid field is interpolated by a standard polynomial method, then there is no background scalar pressure field, which could erroneously influence the particle path lines.

3.5 Numerical errors and preferential concentration of spherical particles

In what follows we will consider how volumes of inertial spherical particles evolve under the flow of the ODEs (3.2.1) and (3.2.2). The goal of this section is to relate the numerical interpolation and integration errors to the clustering mechanisms of the exact solution. This is done by first expanding the exact solution into its elementary differentials. We then discuss the effects of integration errors and interpolation errors on the evolution of volumes of particles. In what follows, we will initially assume that $\mathbf{u}(\mathbf{x})$ is an arbitrary vector field that is not necessarily divergence-free until explicitly mentioned.

3.5.1 Expanding the exact solution

We start this section by defining the notion of volume of the particle suspension. Given an open and bounded set $D_t \subset \mathbb{R}^3$ at time t, then its volume at $t = 0$ is given by

$$\mathrm{vol}(D_0) := \int_{D_0} \mathrm{d}\mathbf{x}_0 \tag{3.5.1}$$

where $\mathbf{x}(0) = \mathbf{x}_0$. This can be thought of as the volume occupied by a suspension of inertial particles confined to the region D_0. The idea is to consider how this volume expands or contracts in time. Consider now the same volume after evolving under the ODEs (3.2.1) and (3.2.2) for time t

$$\mathrm{vol}(D_t) = \int_{D_t} \mathrm{d}\mathbf{x}(t) = \int_{D_t} \det\left(\frac{\partial \mathbf{x}(t)}{\partial \mathbf{x}_0}\right) \mathrm{d}\mathbf{x}_0. \tag{3.5.2}$$

Hence, the quantity

$$\Psi := \det\left(\frac{\partial \mathbf{x}(t)}{\partial \mathbf{x}_0}\right) \tag{3.5.3}$$

determines how volumes of particles contract or expand over time. That is, given a volume of particles, if $\Psi > 1$ the volume is expanding, $\Psi < 1$ the volume is contracting and $\Psi = 1$ the volume is preserved. These three cases correspond to the particulate phase dispersing, concentrating or remaining a constant density, respectively. Note that we will refer to Ψ as the *physical* volume, to distinguish between phase space volume.

To illustrate the connection between $\nabla \cdot \mathbf{v}$ and Ψ, we can take the Jacobian of equation (3.2.2) with respect $\mathbf{x}_0$ [22]

$$\frac{\partial \dot{\mathbf{x}}}{\partial \mathbf{x}_0} = \frac{\partial \mathbf{x}}{\partial \mathbf{x}_0} \frac{\partial \mathbf{v}}{\partial \mathbf{x}}. \tag{3.5.4}$$

Applying Jacobi's formula (3.3.5) yields a differential equation for Ψ

$$\frac{\partial}{\partial t} \Psi = (\nabla \cdot \mathbf{v})\, \Psi. \tag{3.5.5}$$

It is clear that if $\nabla \cdot \mathbf{v} < 1$, then Ψ is decreasing and if $\nabla \cdot \mathbf{v} > 1$ then Ψ is increasing.

We will now show this more concretely, by expanding the exact solution into its elementary differentials, which we now recall. Denote by $\mathbf{y}(t)$ the exact solution of an ODE

$$\dot{y}_i(t) = f_i(\mathbf{y}(t)), \quad \text{for} \quad i = 1, ..., n \tag{3.5.6}$$

For some small time $0 < h \ll 1$, $y_i(h)$ has the following elementary differential expansion [15]

$$y_i(h) = y_i(0) + h f_i\big|_{t=0} + \frac{h^2}{2} \left(\frac{\partial f_i}{\partial y_j} f_j \right)\bigg|_{t=0} \tag{3.5.7}$$

$$+ \frac{h^3}{3!} \left(\frac{\partial^2 f_i}{\partial y_j \partial y_k} f_j f_k + \frac{\partial f_i}{\partial y_j} \frac{\partial f_j}{\partial y_k} f_k \right)\bigg|_{t=0} + \frac{h^4}{4!} \left(\frac{\partial^3 f_i}{\partial y_j \partial y_k \partial y_l} f_j f_k f_l \right. \tag{3.5.8}$$

$$\left. + 3 \frac{\partial^2 f_i}{\partial y_j \partial y_k} \frac{\partial f_l}{\partial y_l} f_l f_k + \frac{\partial f_i}{\partial y_j} \frac{\partial^2 f_j}{\partial y_k \partial y_l} f_k f_l + \frac{\partial f_i}{\partial y_j} \frac{\partial f_j}{\partial y_k} \frac{\partial f_k}{\partial y_l} f_l \right)\bigg|_{t=0} + ... \tag{3.5.9}$$

for $i = 1, ..., n$. We note that the expansion is convergent if h is small compared to $\|\mathbf{f}\|$.

The elementary differentials of the ODEs (3.2.1) and (3.2.2) are calculated and the terms up to $O(h^3)$ are presented

$$v_i(h) = v_i + h\alpha(u_i - v_i) + \frac{h^2}{2}\left(-\alpha^2(u_i - v_i) + \alpha\frac{\partial u_i}{\partial x_j}v_j\right) \tag{3.5.10}$$

$$+\frac{h^3}{3!}\left(\alpha v_j v_k \frac{\partial^2 u_i}{\partial x_j \partial x_k} + \alpha^3(u_i - v_i) - \alpha^2\frac{\partial u_i}{\partial x_j}(u_j - 2v_j)\right) + ... \tag{3.5.11}$$

$$x_i(h) = x_i + hv_i + \frac{h^2}{2}\big(\alpha(u_i - v_i)\big) + \frac{h^3}{3!}\left(\alpha\frac{\partial u_i}{\partial x_j}v_j - \alpha^2(u_i - v_i)\right) + ... \tag{3.5.12}$$

where the variable appearing on the right hand side are evaluated at $t = 0$. Here we assumed nothing about the size of α, but instead take $h \ll \alpha$ such that the series converges. Taking the determinant of the Jacobian of $\mathbf{x}(h)$ from equation (3.5.12) with respect to $\mathbf{x}_0$ yields an expansion for Ψ with repect to h

$$\Psi = 1 + h\Psi_1 + h^2\Psi_2 + h^3\Psi_3 + h^4\Psi_4 + O(h^5) \tag{3.5.13}$$

where

$$\Psi_1 = 0, \quad \Psi_2 = \frac{\alpha}{2}\chi_3, \quad \Psi_3 = \frac{\alpha}{6}\left(\chi_5 - \alpha\chi_3\right) \tag{3.5.14}$$

$$\Psi_4 = \frac{\alpha}{24}\left(\chi_1 + \alpha\chi_2 + \alpha^2\chi_3 + 3\alpha\chi_3^2 - 2\alpha\chi_4 - 2\alpha\chi_5\right). \tag{3.5.15}$$

The elementary differentials χ_i are given by

$$\chi_1 = v_i v_j \frac{\partial^3 u_k}{\partial x_i \partial x_j \partial x_k}, \quad \chi_2 = u_i\frac{\partial^2 u_j}{\partial x_i \partial x_j}, \quad \chi_3 = \frac{\partial u_i}{\partial x_i} \tag{3.5.16}$$

$$\chi_4 = \frac{\partial u_j}{\partial x_i}\frac{\partial u_i}{\partial x_j} = \|S\|_F^2 - \|\Omega\|_F^2, \quad \chi_5 = v_j\frac{\partial^2 u_i}{\partial x_i \partial x_j}. \tag{3.5.17}$$

This can be verified by expansion of (3.5.5) into its Taylor series. If we insist that the fluid field is divergence-free then the Ψ_i and χ_i all vanish except for Ψ_4 and χ_4. We are then left with

$$\Psi\big|_{\nabla\cdot\mathbf{u}=0} = 1 - \frac{\alpha^2}{12}h^4\left(\|S\|_F^2 - \|\Omega\|_F^2\right) + O\left(h^5\right) \tag{3.5.18}$$

which relates the fluid rate of strain and rotation with the contractivity of physical volume in the same way as the centrifuge effect (3.2.8). That is, if the rate of vorticity is greater than the rate of strain, physical volumes of particles will contract and vice versa.

3.5.2 Expanding the numerical solution and numerical errors

In this section, we perform a similar analysis to that of section 3.5.1 but instead of the exact flow of the ODE, we consider now how physical volumes of particles evolve under the *numerical* flow. That is, we will look at how errors to the divergence of the fluid field affect the evolution of volumes of particles under the numerical solution to the equations of motion. We do so by expanding the numerical methods into their elementary differentials and comparing the expansions with the exact solution. We then look at how errors to the divergence of the fluid field affect the evolution of physical volume under the numerical flow. The results are that if a divergence-free interpolation method is used, the numerical methods preserve the same qualitative behavior of the centrifuge effect. Moreover, we show here that the centrifuge-preserving methods replicate the centrifuge-effect from equation (3.5.18) up to the accuracy of the interpolation method when the fluid field is divergence-free.

Consider the map $\Phi_h^{[n]} : (\mathbf{v}_0, \mathbf{x}_0) \to (\mathbf{v}_1, \mathbf{x}_1)$, where the superscript $[n]$ denotes the numerical method in consideration. We calculate $\Psi^{[n]} = \det\left(\frac{\partial \mathbf{x}_1}{\partial \mathbf{x}_0}\right)$ and expand the solution in h yielding an expression of the form

$$\Psi^{[n]} = 1 + h\Psi_1^{[n]} + h^2\Psi_2^{[n]} + h^3\Psi_3^{[n]} + h^4\Psi_4^{[n]} + O(h^5) \qquad (3.5.19)$$

The values of $\Psi_i^{[n]}$ for $i = 2, 3, 4$ for the Forward Euler (FE1), Lie-Trotter (LT1), order one centrifuge-preserving (CP1), Ralston (RK2), Adams-Bashforth two-step (AB2), order two centrifuge-preserving (CP2) and Strang splitting (SS2) methods are presented in table 3.5.1. Note that $\Psi_1^{[n]} = 0$ for all the above methods.

We make a number of observations from this table. First, the divergence of the fluid field affects $\Psi^{[n]}$ at $O(h^2)$ for each method. The one exception to this is FE1, which satisfies $\Psi^{[FE]} = 1$ and therefore erroneously preserves physical volume. When the divergence of the fluid field is zero, all the $\chi_i = 0$ except χ_4. For example, setting $\nabla \cdot \mathbf{u} = 0$ gives for the SS2 method

$$\Psi^{[SS]}\big|_{\nabla\cdot\mathbf{u}=0} = 1 - \frac{\alpha^2}{8} h^4 \left(\|S\|_F^2 - \|\Omega\|_F^2\right) + O\left(h^5\right). \qquad (3.5.20)$$

This means that the numerical solution generated by the Strang splitting method (3.3.29) reproduces the *qualitative* nature of the centrifuge effect, in the sense that $\Psi^{[SS]}\big|_{\nabla\cdot\mathbf{u}=0} > 1$ when $\|\Omega\|_F > \|S\|_F$. This qualitative centrifuge effect is seen by all the methods (other than FE1) by setting $\nabla \cdot \mathbf{u} = 0$ in table 3.5.1. However, we note here that the coefficient of the $O(h^4)$ term in equation (3.5.20) is different to that of the exact solution (3.5.18). Meaning that, while the method

Method	$\Psi_2^{[n]}$	$\Psi_3^{[n]}$	$\Psi_4^{[n]}$
Exact solution	$\frac{\alpha}{2}\chi_3$	$\frac{\alpha}{6}\left(\chi_5-\alpha\chi_3\right)$	$\frac{\alpha}{24}\left(\chi_1+\alpha\chi_2+\alpha^2\chi_3+3\alpha\chi_3^2\right.$ $\left.-2\alpha\chi_4-2\alpha\chi_5\right)$
FE1	0	0	0
LT1	$\alpha\chi_3$	$\frac{\alpha^2}{2}\chi_3$	$-\frac{\alpha^2}{6}\left(3\,\chi_4-\alpha\,\chi_3-3\,{\chi_3}^2\right)$
CP1	$\frac{\alpha\sqrt{6}}{6}\chi_3$	$\frac{\alpha\left(\sqrt{6}-1\right)}{6}\chi_5-\frac{\alpha^2\sqrt{6}}{12}\chi_3$	$\frac{\alpha}{36}((\alpha^2\chi_3-3\,\alpha\,\chi_5+\frac{7}{2}\chi_1)\sqrt{6}$ $+(3\,{\chi_3}^2-3\,\chi_4+3\,\chi_5)\alpha-6\,\chi_1)$
AB2	$\frac{3\alpha}{16\theta}\chi_3$	$\frac{3\alpha}{32}\left(\chi_5-\alpha\chi_3\right)$	$\frac{\alpha}{128}\left(3\theta\chi_1+16\alpha\chi_3^2-16\alpha\chi_4\right)$
RK2	$\frac{\alpha}{2}\chi_3$	0	$\frac{\alpha^2}{8}\left(\chi_3^2-\chi_4\right)$
CP2	$\frac{\alpha}{2}\chi_3$	$\frac{3\alpha^2}{16}\chi_3+\frac{\alpha}{6}\chi_5$	$\frac{\alpha}{576}(33\,\alpha^2\chi_3+72\,\alpha\,{\chi_3}^2+24\,\alpha\,\chi_2$ $-48\,\alpha\,\chi_4-60\,\alpha\,\chi_5+32\,\chi_1)$
SS2	$\frac{\alpha}{2}\chi_3$	$\frac{\alpha}{4}\left(\chi_5-\alpha\chi_3\right)$	$\frac{\alpha}{48}\left(3\chi_1+4\alpha^2\chi_3+6\alpha\chi_3^2-6\alpha\chi_4-6\alpha\chi_5\right)$

Table 3.5.1: The terms in the series expansion (3.5.19) for the physical volume $\Psi^{[n]}$ under various numerical methods. Note that $\Psi_1^{[n]}=0$ for all the methods.

contracts physical volume when the exact solution does, it does so at an erroneous rate. This issue is circumvented by the centrifuge-preserving methods (CP1 and CP2), where we have chosen the time-step coefficients in such a way such that they yield the exact same expansion as (3.5.18) up to $O(h^4)$ and hence contracts physical volume at the same rate as the exact solution to leading order.

We now discuss the effect of interpolation errors in simulations of spherical particles in numerically calculated flows. Say that $\mathbf{u}_e(\mathbf{x})$ is the true solution to the underlying Navier-Stokes equations that satisfies $\nabla\cdot\mathbf{u}_e(\mathbf{x})=0$. As this exact solution is generally not available, we consider the following three cases

1. Case (a): the fluid field has interpolation errors $\boldsymbol{\delta}(\mathbf{x})$ that are not divergence free $\mathbf{u}(\mathbf{x})=\mathbf{u}_e(\mathbf{x})+\boldsymbol{\delta}(\mathbf{x})$, where $\nabla\cdot\boldsymbol{\delta}(\mathbf{x})\neq 0$ (e.g., using standard polynomial interpolation)

2. Case (b): the fluid field has interpolation errors $\boldsymbol{\delta}(\mathbf{x})$ and is divergence free $\mathbf{u}(\mathbf{x})=\mathbf{u}_e(\mathbf{x})+\boldsymbol{\delta}(\mathbf{x})$, where $\nabla\cdot\boldsymbol{\delta}(\mathbf{x})=0$ (e.g., using MRBF interpolation)

3. Case (c): the fluid field is free of errors $\mathbf{u}(\mathbf{x})=\mathbf{u}_e(\mathbf{x})$ and $\nabla\cdot\mathbf{u}(\mathbf{x})=0$ (e.g., when the velocity field is available in closed form, also referred to as

“exact" interpolation.)

We pay particular attention to how errors resulting in $\nabla\cdot\boldsymbol{\delta}(\mathbf{x})\neq 0$ affect how the numerical methods evolve physical volume. To quantify this we define the physical volume error by

$$\Delta\Psi^{[n]}=\Psi-\Psi^{[n]}. \tag{3.5.21}$$

Here, Ψ is used to denote the physical volume over time h of the exact solution with fluid field corresponding to case (c), that is, the physical volume of the true solution in the absence of any errors, whereas $\Psi^{[n]}$ is the physical volume of the numerical solution with fluid field corresponding to one of the three cases given below. The results are presented in table 3.5.2. We see here that the physical volume errors in case (a) are $O(h^2)$ and proportional to $\nabla\cdot\boldsymbol{\delta}(\mathbf{x})$ for all the methods. In case (b), the physical volume errors are proportional to $O(h^4)$ except for the centrifuge-preserving methods, which have physical volume error proportional to $O(h^4\delta_4)$, where $\delta_4=\left|\chi_4(\mathbf{u})-\chi_4(\mathbf{u}+\boldsymbol{\delta})\right|=O(|\boldsymbol{\delta}|)\ll|\chi_4(\mathbf{u})|$ is the error of χ_4 from the interpolation method, which we assume is small. In case (c), $\delta_4=0$ and the centrifuge-preserving methods have physical volume error proportional to $O(h^5)$, while the other methods are $O(h^4)$. It is due to this behavior that we expect all the methods to more accurately evolve physical volume, when a divergence-free interpolation method is implemented such as with MRBFs. In this case, we expect the centrifuge-preserving methods to perform especially well due to table 3.5.2.

	$\lvert\Delta\Psi^{[n]}\rvert$		
Method	Case (a) $(\nabla\cdot\mathbf{u}(\mathbf{x})\neq 0)$	Case (b) $(\nabla\cdot\mathbf{u}(\mathbf{x})=0)$	Case (c) $(\mathbf{u}(\mathbf{x})=\mathbf{u}_e(\mathbf{x}))$
FE1	$\frac{\alpha}{2}h^2\lvert\chi_3\rvert$	$\frac{\alpha^2}{12}h^4\lvert\chi_4\rvert$	$\frac{\alpha^2}{12}h^4\lvert\chi_4\rvert$
LT1	$\alpha h^2\lvert\chi_3\rvert$	$\frac{5\alpha^2}{12}h^4\lvert\chi_4+\frac{\delta_4}{2}\rvert$	$\frac{5\alpha^2}{12}h^4\lvert\chi_4\rvert$
CP1	$\frac{\alpha\sqrt{6}}{6}h^2\lvert\chi_3\rvert$	$\frac{\alpha^2}{12}h^4\lvert\delta_4\rvert$	$O(h^5)$
AB2	$\frac{3\alpha}{16\theta}h^2\lvert\chi_3\rvert$	$\frac{\alpha^2}{24}h^4\lvert\chi_4+\frac{\delta_4}{8}\rvert$	$\frac{\alpha^2h^2}{24}h^4\lvert\chi_4\rvert$
RK2	$\frac{\alpha}{2}h^2\lvert\chi_3\rvert$	$\frac{\alpha^2}{24}h^4\lvert\chi_4+\frac{\delta_4}{8}\rvert$	$\frac{\alpha^2}{24}h^4\lvert\chi_4\rvert$
CP2	$\frac{\alpha}{2}h^2\lvert\chi_3\rvert$	$\frac{\alpha^2}{12}h^4\lvert\delta_4\rvert$	$O(h^5)$
SS2	$\frac{\alpha}{2}h^2\lvert\chi_3\rvert$	$\frac{\alpha^2}{24}h^4\lvert\chi_4+\frac{\delta_4}{8}\rvert$	$\frac{\alpha^2}{24}h^4\lvert\chi_4\rvert$

Table 3.5.2: The errors of the physical volume after one time step for the numerical methods under consideration.

In addition to the erroneous contraction of physical volume, we note from examples 1 - 3 that large divergence errors impose more stringent restrictions on the time step for the numerical methods to be contractive.

3.6 Numerical simulations

In this section we test our numerical methods for simulating suspensions of particles in viscous flows. The section begins by outlining the flow field and summarizing the methods and numerical parameters. We then outline the computational cost and verify the convergence of the methods. Next we simulate suspensions of 10^4 particles in Taylor-Green vortices. This is the most important part of the section and is comprised of three experiments. The first compares the integration methods with exact evaluation of the fluid field. The second compares the effect of different interpolation errors with the CP2 integration. The third and final experiment explores how a combination of the proposed interpolation and integration methods can be used to generate cost-effective accurate particle distributions compared to conventional methods.

3.6.1 Preliminaries

Here we will briefly outline the numerical methods that are under consideration in the forthcoming numerical experiments, the fluid field and finally the particle models.

The integration methods under consideration and their properties are summarized in table 3.6.1.

	FE1	LT1	CP1	AB2	RK2	SS2	CP2
Order	1	1	1	2	2	2	2
Contractivity-preserving	No	Yes	Yes	No	No	Yes	Yes
Centrifuge-preserving	No	No	Yes	No	No	No	Yes

Table 3.6.1: Summary of the properties of the integration methods under consideration

We will abbreviate the divergence-free MRBF interpolation with the nearest $(n+1) \times (n+1) \times (n+1)$ data points by MRBFn and the non-divergence-free order n tripolynomial interpolation by TPn. The MRBF shape parameters are set to $\epsilon_1 = 0.31$, $\epsilon_2 = 0.23$ and $\epsilon_3 = 0.16$ corresponding to the MRBF1, MRBF2 and MRBF3 schemes, respectively, and are chosen empirically. We will compare the methods against a reference solution that uses exact evaluation of the analytic fluid field and the classical fourth order Runge-Kutta method for time integration with a time step that is 10 times smaller then that of the other methods. Note that such a reference solution is only available in the case that the flow field is known in closed form.

The discrete fluid field is generated by evaluating a closed form solution to the Navier Stokes equation on a regularly spaced grid with uniform sampling in each direction $\Delta x = \Delta y = \Delta z = 1/10$. We use a stationary Taylor-Green vortex solution that was proposed in [46] and has been used by other authors to study the behaviour of particles in cellular flow fields [6, 24, 31, 40]. The particular Taylor-Green flow field used in the experiments is given by $\mathbf{u}(\mathbf{x}) = (u(\mathbf{x}), v(\mathbf{x}), w(\mathbf{x}))^T$ where

$$u(\mathbf{x}) = 2\cos(2\pi x)\sin(2\pi y)\sin(2\pi z), \tag{3.6.1}$$

$$v(\mathbf{x}) = -\sin(2\pi x)\cos(2\pi y)\sin(2\pi z), \tag{3.6.2}$$

$$w(\mathbf{x}) = -\sin(2\pi x)\sin(2\pi y)\cos(2\pi z). \tag{3.6.3}$$

We will perform experiments on both spherical and non-spherical particles. Denoting by λ the aspect ratio of the particle, then $\lambda = 1$ corresponds to spherical particles, $\lambda > 1$ corresponds to a prolate spheroid and $\lambda < 1$ corresponds to an oblate spheroid. For $\lambda = 1$, the equations of motion are given by equations (3.2.1) and (3.2.2), while for $\lambda \neq 1$ the equations of motion are (3.2.1), (3.2.2), (5.5.1) and (3.2.4). For details about the moment of inertia tensor J, torque term $\mathbf{T}$, resistance tensor K for the $\lambda \neq 1$ cases we refer to 5.3. Finally, we note that for all of the following experiments, the particles are given a random initial location within in a box of width 0.01 centered at the point $x_0 = (1/3, 1/5, 1/7)^T$ in the domain and a random initial orientation for non-spherical particles.

3.6.2 Computational cost

Here, we outline the main computational costs associated with the methods. The two main steps in the algorithm are the interpolation step and the time integration step, which we examine separately. The wall clock times T_w for 10^4 time steps of the considered integration methods using exact evaluation of the fluid field are measured and presented in table 3.6.2 and T_w for 10^4 time steps of the various interpolation methods using the FE1 method are presented in table 3.6.3.

We note that the centrifuge-preserving methods are slightly more costly due to extra evaluations of the $\Phi_{ah}^{[LT]}$ operator. However, we note that one could speed up many of these splitting methods by observing that they are conjugate to a lower stage faster method, for example

$$\left(\Phi_h^{[SS]}\right)^N = \psi_{\frac{h}{2}}^{[1]} \circ \left(\Phi_h^{[LT]*}\right)^N \circ \psi_{\frac{-h}{2}}^{[1]},$$

hence repeated evaluations of the operator $\Phi_h^{[SS]}$ when implemented in this way effectively has the same cost as $\Phi_h^{[LT]*}$. Similar observations are made for the centrifuge preserving methods.

For the interpolation step, there are two main calculations that contribute the most to the computational cost. The first being the solution of a linear system of size $3n^3 \times 3n^3$ to find the interpolation coefficients. Guassian elimination is used for this purpose due to simplicity and the fact that the systems are not so large (at most 192 × 192 for the MRBF3 and TP3 methods). However, we note the existence of the exact matrix inverses for the coefficient matrices of these linear systems. This can be found in [27] for the TP method and [2] for the MRBF method, the latter being due to the fact that the coefficient matrix has a block toeplitz structure for MRBF interpolation on Cartesian grids. The next most significant cost is evaluation of the sums of basis functions, that is, the sum in (3.4.1) and a similar equation for the TP methods. MRBF interpolation involves evaluation of more complex basis functions (i.e., matrix-vector products containing exponentials of polynomials), which is more costly than evaluating sums of monomials for the TP interpolation. This cost is more of a burden for the MRBF2 and MRBF3 methods as seen in table 3.6.3.

	FE1	LT1	CP1	AB2	RK2	CP2	SS2
$T_w\,(s)$	1.1817	1.3610	1.6577	2.0524	2.0376	2.5001	1.6455

Table 3.6.2: The wall clock times for 10^4 time steps using different integration methods and exact interpolation.

	MRBF1	MRBF2	MRBF3	TP1	TP2	TP3
$T_w\,(s)$	3.3796	5.0776	12.0333	3.7263	4.2226	5.7551

Table 3.6.3: The wall clock times for 10^4 time steps using the FE1 method and different integration methods.

We see here that the MRBF1 and TP1 methods are roughly equal in cost. The MRBF2 method is about double that of TP1 and MRBF1 and is more expensive than the TP2 method. The MRBF3 method is double the cost of the TP3 method. We recall that we are not constrained to these three choices of MRBF methods and one is free to use any number of data points to achieve an optimum balance of accuracy and cost. This freedom is due to the fact that MRBF interpolation was designed for interpolation on scattered data points [51]. This option is not available for the TPn methods, where $n+1$ grid points in each dimension are required to ensure the existence of a unique degree n interpolating polynomial.

3.6.3 Convergence

In this section, we will verify the convergence of the integration methods first with exact interpolation then with various combinations of the interpolation methods for spherical ($\lambda = 1$) and non-spherical ($\lambda = 10$). In these experiments, we set $St = 1$ and compute the particles' dynamics for time $T = 1$.

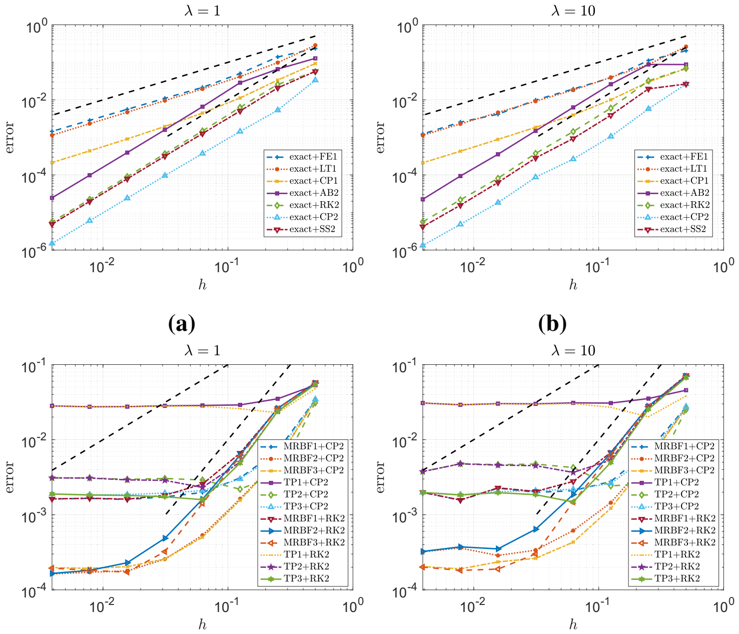
The convergence of the error, measured in the 2-norm, of the integration methods are presented in figures 3.6.1a and 3.6.1b. We observe here that the FE1 and LT1 methods have similar accuracy as do the RK2 and SS2 methods. It is noted that the benefits of preserving contractivity in the various splitting methods are expected to be seen after longer times. One remarkable observation here is that for this Stokes number the first order CP1 method is competitive with the second order RK2 and AB2 methods at large time steps, furthermore, the CP2 method is the most accurate by a factor of about 5 in both scenarios.

Figures 3.6.1c and 3.6.1d show the convergence of the CP2 and RK2 methods with different interpolation methods. We see that the methods initially converge at their expected order, but as h goes to zero we see that the integration error becomes overshadowed by the h-independent interpolation error. We observe that the MRBF solutions yield more accurate solutions than the TP solutions using the equal number of data points. However, the TP3 solution is expected to perform better for longer simulations where particles cross grid cells. This is because the piece-wise fluid field constructed from the TP3 method is globally $C^1(\mathbb{R}^3)$, meaning that the spatial derivatives of the fluid velocity are everywhere continuous. This is not true for the other methods.

Finally, we remark that the centrifuge-preserving methods perform equally well for non-spherical particles.

3.6.4 Simulating suspensions of particles

Up until now we have mainly focused on the average error in the positions of individual particles. It is well known that standard methods such as polynomial interpolation and Adams-Bashforth integration do a good job at minimizing this truncation error in some norm. However, while it is indeed important that



(a) **(b)**

(c) **(d)**

Figure 3.6.1: Figures (a) and (b) show the convergence of the numerical integration methods using exact interpolation for the $\lambda = 1$ and $\lambda = 10$ equations, respectively. Figures (c) and (d) show the convergence of the CP2 and RK2 methods with the six interpolation methods for the $\lambda = 1$ and $\lambda = 10$ equations, respectively. The dashed lines are $O(h)$ and $O(h^2)$

this conventional measure of error is kept at a minimum, in practice one is usually more interested in properties of distributions of many indistinguishable particles meaning that the individual error of each particle is less important. Due to this, it is more desirable that an algorithm reproduces accurately the spatial statistical properties of many particles rather than minimizing the absolute errors of each individual particle. With this in mind, the main goal here is to test to what extent the aforementioned errors affect suspensions of particles when viewed as a single discrete probability distribution. We will show that distributions of particles calculated by our proposed geometric methods will more closely resemble that of the exact solution, despite sometimes having higher average error per particle in the conventional sense.

In our context, a distribution $P = \{(\mathbf{x}_i, w_i)\}_{i=1}^{n_c}$ is a set of n_c non-empty equally sized cells, where $\mathbf{x}_i$ is the location of the cell center and w_i is a weight that is equal to the number of particles in that cell. We let P_n denote a distribution where the particle locations are calculated by a numerical method, P_{ref} refers to the distribution obtained by the reference solution and we use $300 \leq n_c \leq 400$ depending on the spread of particles. We will determine the accuracy of P_n using three measures which we now outline.

The first is the first Wasserstein distance, which is a natural metric to compare the distance between two discrete probability distributions of equal size (also known as the Earth Mover Distance). The first Wasserstein distance between two probability distributions is denoted by $W_1(P_1, P_2)$ and is a measure of the cost of transporting the distribution P_1 into P_2 in the cheapest way possible. The cost is measured as the distance between cell centers, measured in the 2-norm and weighted by the number of particles being transported. For mathematical details about the first Wasserstein distance, we refer the reader to [41] and the numerical computation of the first Wasserstein distances are computed using a publicly available MATLAB code [1]. We denote by $W_1(P_n) = W_1(P_n, P_{\text{ref}})$ the first Wasserstein distance between P_n and P_{ref}.

The second is the relative entropy (also known as the Kullback-Leibler divergence) [26], which is a measure of how much information is lost from a reference distribution P_2 when an approximate distribution P_1 is used. The relative entropy is calculated by

$$E(P_1, P_2) = \sum_{\mathbf{x}_i \in \Omega_P} P_1(\mathbf{x}_i) \log\left(\frac{P_1(\mathbf{x}_i)}{P_2(\mathbf{x}_i)}\right), \tag{3.6.4}$$

where $P(\mathbf{x}_i) = w_i$ is the number of particles in the cell at $\mathbf{x}_i$ and Ω_P is the support of the two distributions. If there is an empty cell in one distribution and not the other, say at $\mathbf{x} = \mathbf{x}_0$ we use $P(\mathbf{x}_0) = 10^{-1}$, to avoid singularities. This modestly penalizes the approximate solution for predicting a non-zero probability of having a particle in a cell that should have zero particles according to the reference distribution. We denote by $E(P_n) = E(P_n, P_{\text{ref}})/n_p$ the relative entropy between P_n and P_{ref} scaled by the number of particles $n_p = 10^4$.

Finally, the third means of determining the accuracy of the distribution is by the average error of the particle positions $\overline{\Delta \mathbf{x}_n}$. This conventional measure of error is calculated by taking the difference between the final position of the numerical and reference solution starting from the same initial conditions and averaging over all the $n_p = 10^4$ particles, that is

$$\overline{\Delta \mathbf{x}_n} = \frac{1}{n_p} \sum_{i=1}^{n_p} \|\mathbf{x}_{n,i} - \mathbf{x}_{\text{ref},i}\|_2, \tag{3.6.5}$$

where $\mathbf{x}_{n,i}$ is the ith particle calculated by the numerical method and $\mathbf{x}_{\text{ref},i}$ is the ith particle under the reference solution. As the rotational variables are strongly coupled with the translational variables, errors in the rotational dynamics will also influence the final positions of the particles, hence this is a reasonable measure of the error of the algorithms' overall accuracy in computing the dynamics of a single particle. We recall that this error is that which the conventional methods are designed to reduce when referring the the global order of accuracy of a method.

In the forthcoming experiments, we will use various combinations of integration and interpolation methods to compute the paths of 10^4 particles in the discrete Taylor-Green vortices starting with random positions and orientations within a cube of width $1/100$ centered at the point $(1/3, 1/5, 1/7)^T$. We perform three experiments. The first compares how the various numerical integration methods and their errors affect the spatial distribution of suspensions of particles in the absence of interpolation errors. The second experiment investigates how interpolation errors affect the spatial distribution of particles using the CP2 method. Finally, we look at how a combination of MRBF interpolation and centrifuge-preserving integration can be used to calculate fast and accurate suspensions of particles compared to the conventional AB2+TPn methods, similar to the methods used in [12, 37, 38, 48, 50], for example.

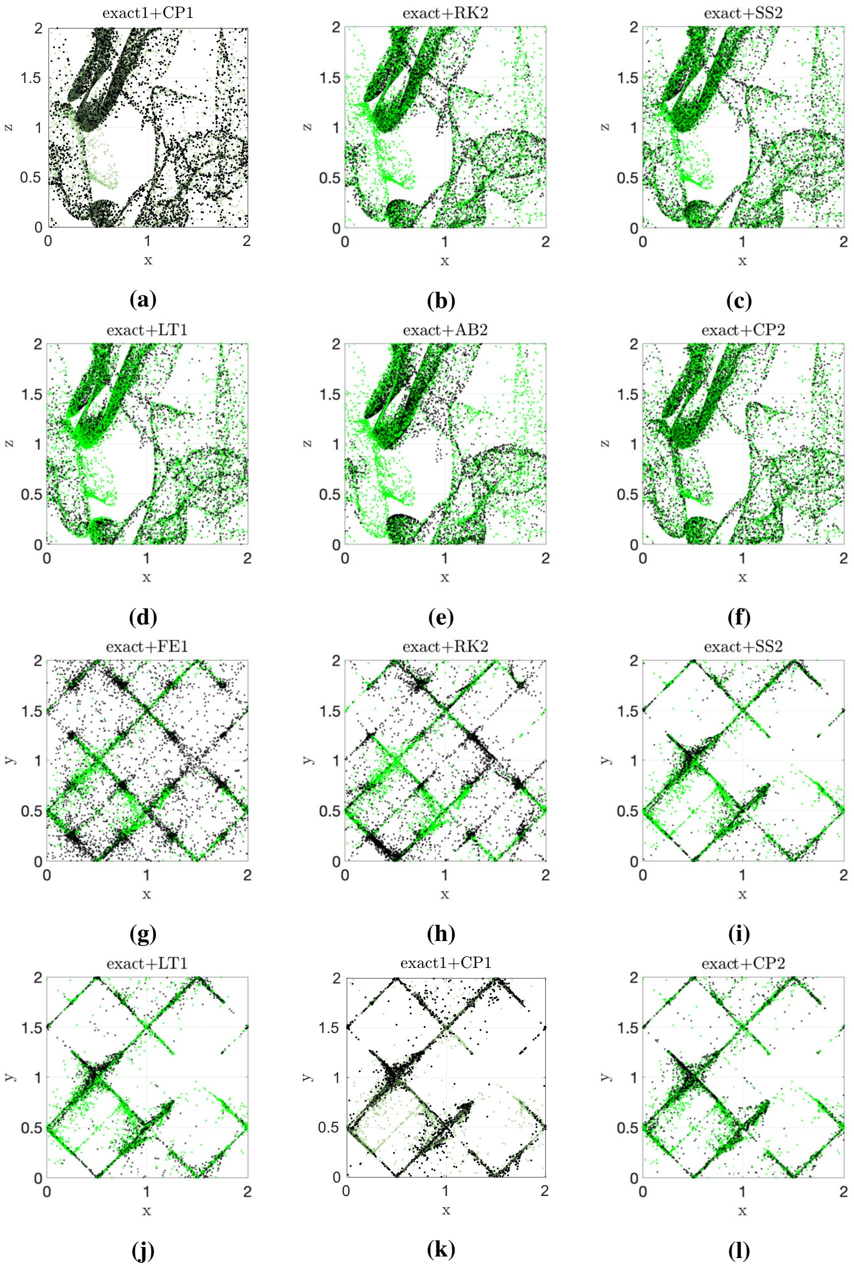
Comparison of integration methods

In this experiment we use the seven integration methods outlined in table 3.6.1 to simulate a suspension of particles evolving in the Taylor-Green flow with exact interpolation. The methods are each tested in six separate simulations, three with Stokes numbers of $St = 1/5, 1, 10$ for spherical particles ($\lambda = 1$) and three with the same Stokes numbers for non-spherical particles ($\lambda = 1/10$). At the end of the simulation the relative entropy $E(P_n)$, first Wasserstein distance $W(P_n)$ and the average spatial error $\overline{\Delta \mathbf{x}_n}$ between the numerical distribution and the reference distribution are calculated and presented in table 3.6.4. The time step h and total simulation time T are also presented in this table. We start by discussing some qualitative features of the final distributions, examples of which are given in figure 3.6.2. We then discuss the results of table 3.6.4 in detail.

Figure 3.6.2 depicts the final distribution of the particles for the various integration methods. The particle positions are plotted modulo 2 for presentation purposes and represented by black dots, while the reference solution is plotted using green dots. Figures 3.6.2a to 3.6.2f correspond to the $St = 10$, $\lambda = 1$ simulation and is viewed along the y direction. We see here that the CP2 solution is able to predict the correct clustering in all the regions that are predicted by the green reference solution. The LT1, CP1 and SS2 solutions are visually similar to each other, however do not correctly predict clustering of particles in some regions, given by regions of green dots that are void of black dots. The RK2 and AB2 solutions do a worse job as seen, again, by even more regions with a higher concentration of green dots compared to black dots. Similar observations are again seen in figures 3.6.2g to 3.6.2l, which correspond to the $St = 1$, $\lambda = 1/10$ simulation, viewed along the z direction. In this simulation, the particles more closely follow the streamlines of the fluid field and more quickly concentrate in regions of high strain as seen by the regions of dense green dots. In these figures, it is even more easily seen that the four contractivity preserving methods do a good job at correctly clustering particles in regions where the reference solution does, while we see with the FE1 and RK2 solutions multiple regions exhibiting an erroneous concentration of black dots that are void of green dots. The AB2 solution is unstable for these parameters.

To quantify the above observations, which have up until now been visual, we turn our attention to table 3.6.4. We start by outlining some general observations that are common to all six simulations. Looking first at the order one methods, we observe that the LT1 and CP1 methods, which are contractivity preserving, outperform the FE1 method in almost all measures in each simulation despite the fact that their order of accuracy is the same. What is striking here is that in most simulations the LT1 and CP1 methods generally have a lower $\overline{\Delta \mathbf{x}_n}$, $W(P_n)$ and $E(P_n)$ compared to the conventional RK2 and AB2 methods despite being of lower order and computational cost. Similar observations are made if we turn our attention towards the order two methods. That is, the SS2 method in most cases has lower $\overline{\Delta \mathbf{x}_n}$, $W(P_n)$ and $E(P_n)$ than the RK2 and AB2 methods, and better still is the CP2 method. The advantage of the CP2 method over the SS2 is more pronounced than the advantage of the CP1 method over the LT1. The CP2 method has the lowest $\overline{\Delta \mathbf{x}_n}$, $W(P_n)$ and $E(P_n)$ in all six simulations and is clearly the best method here in all three metrics.

For the $St = 1$, $\lambda = 1/10$ simulation, the CP1 solution has a larger $\overline{\Delta \mathbf{x}_n}$ than the SS2 solution, but a lower $W(P_n)$ and $E(P_n)$ which suggests that for these simulation parameters, the centrifuge-preserving property is more advantageous for producing more accurate distributions than simply reducing the accuracy of the method in the conventional sense. In this simulation, the CP1 method is the second best in all measures, the best being the CP2 method. It is also noteworthy that in other simulations the CP1 method has roughly equal, and sometimes lower $\overline{\Delta \mathbf{x}_n}$, $W(P_n)$ and $E(P_n)$ than the SS2 solution, which further suggests that the centrifuge-preserving property is advantageous.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 3.6.2: Figures (a) through (f) show the spatial distribution of the particles in the $x-z$ plane for the $St = 10$, $h = 1/5$, $T = 20$, $\lambda = 1$ simulation from table 3.6.4 (The exact+FE1 is not shown). Figures (g) through (l) show the spatial distribution of the particles in the $x-y$ plane for the $St = 1$, $h = 1/20$, $T = 8$, $\lambda = 1/10$ simulation (the exact+AB2 solution is not shown). The reference solution is plotted in green in all figures.

One of the most remarkable observations is made for the $St = 10$, $\lambda = 1/10$ experiment. Here, the values of $\overline{\Delta\mathbf{x}_n}$ are quite severe and roughly the same for all methods, due to the fact that the time step is quite large and the non-spherical ODEs are more stiff. Despite this, the contractivity preserving methods have a much lower $W(P_n)$ and $E(P_n)$ and the centrifuge preserving methods are better still. This highlights the fact that preserving the aforementioned physical features in the numerical solution results in spatial distributions that more closely resemble the reference solution, despite having the same $\overline{\Delta\mathbf{x}_n}$.

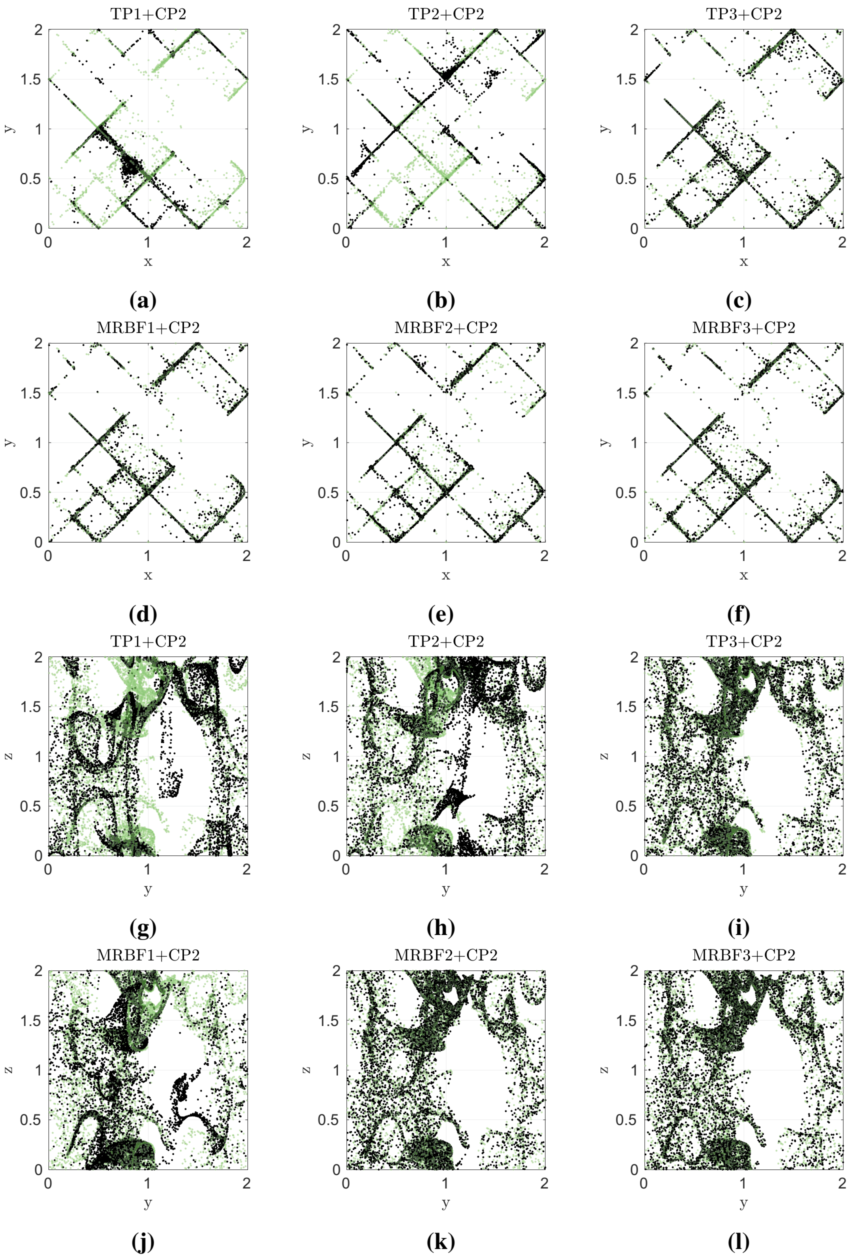
Finally, we mention that all the splitting schemes have better stability properties and are still able to produce accurate clusters of particles for low Stokes numbers and reasonably large time steps as noted by the $\lambda = 1/10$ simulations for $St = 1/5$ and $St = 1$ where we begin to see some of the conventional methods losing stability.

Comparison of interpolation methods

In this experiment we compare the MRBF and TP interpolation methods in combination with the CP2 method to simulate a suspension of particles evolving in the Taylor-Green flow. Six separate simulations are performed, three with Stokes numbers of $St = 1/10, 1, 10$ for spherical particles ($\lambda = 1$) and three with the same Stokes numbers for non-spherical particles ($\lambda = 5$). At the end of each simulation the average spatial error $\overline{\Delta\mathbf{x}_n}$, relative entropy $E(P_n)$ and the first Wasserstein distance $W(P_n)$ between the numerical distribution and the reference distribution are calculated and the results are presented in table 3.6.5. The time step h and total simulation time T are also presented in this table. Some spatial distributions produced by the different interpolation methods are presented in figure 3.6.3.

Directing our attention towards figures 3.6.3a to 3.6.3f, which show the final distribution of the $St = 1/10$, $\lambda = 5$ simulation looking down the z-axis. It can be seen here that the three MRBF solutions look visually very similar to the reference solution, as does the TP3 solution. If we look towards the corresponding part of table 3.6.5, we see that the TP3 solution has a $\overline{\Delta\mathbf{x}_n}$ of 0.3468, which is lower than the MRBF1 solution, which has a $\overline{\Delta\mathbf{x}_n}$ of 0.5389. Despite this, the MRBF1 solution, which we note performs exceptionally well here, has a lower $E(P_n)$ and $W(P_n)$ meaning that the final distribution is more similar to the reference distribution even though the $\overline{\Delta\mathbf{x}_n}$ is greater.

Figures 3.6.3g to 3.6.3l show the final distribution of the $St = 1$, $\lambda = 1$ simulation looking down the x-axis. Here we see that the TP1, TP2 and MRBF1



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 3.6.3: Figures (a) through (f) show the spatial distribution of the particles in the $x-y$ plane for the $St = 1/10$, $h = 1/100$, $T = 6$, $\lambda = 1$ simulation from table 3.6.5. Figures (g) through (l) show the spatial distribution of the particles in the $y-z$ plane for the $St = 1$, $h = 1/40$, $T = 12$, $\lambda = 1/10$ simulation. The reference solution is plotted in green in all figures.

solutions differ visually from the reference solution, whilst the TP3, MRBF2 and MRBF3 solutions look quite similar. From the corresponding section of table 3.6.5, we see that the MRBF2 solution's $\overline{\Delta \mathbf{x}_n}$ is 0.5004 compared to the TP3 solution, which is 0.3332, but both have a similar $E(P_n)$ and $W(P_n)$. Furthermore, there are many examples here of the MRBF solutions having higher $\overline{\Delta \mathbf{x}_n}$, but lower $E(P_n)$ and $W(P_n)$. This can be seen in all three $\lambda = 1$ simulations, where the MRBF2 solution has larger $\overline{\Delta \mathbf{x}_n}$ than the TP3 solution but similar or lower $E(P_n)$ and $W(P_n)$. For the $\lambda = 5$ simulations and for $St = 1$ and $St = 10$, the MRBF2 solution outperforms the TP3 solution in all three measures. Such examples indicate that preserving the divergence-free condition is important to acheive accurate spatial distributions.

We now make some general observations about table 3.6.5. We see that in all but one simulation, the MRBF1 solutions outperform the TP2 solution in all three measures. It is noteworthy that the MRBF1 solution is as fast as TP1 interpolation where both require only eight data points for the interpolation as opposed to 27 data points for the TP2 interpolation, which is a slower method. Additionally, in all six simulations the MRBF3 interpolation method outperforms all the TP solutions in all measures.

To summarize, we have seen many examples of the MRBF solutions producing distributions that are more similar to the reference solution than the TP solutions, despite having worse $\overline{\Delta \mathbf{x}_n}$. These observations are consistent with the fact that the CP2 method, among others, loses accuracy in $\Delta\Psi^{[n]}$ when evolving particles in a non-divergence-free flow field. That is, the physical volume $\Psi^{[n]}$ is more strongly affected when the divergence-free condition is broken, despite the fact that the order of accuracy of the method remains unaffected.

Comparison of interpolation and integration methods

Our final experiment explores the benefit that is gained by combining MRBF interpolation with the centrifuge- and contractivity-preserving methods compared to the standard methods used in the literature. We will compare the methods TP1+FE1, MRBF1+CP1, TP2+AB2, TP3+AB2, MRBF2+CP2 and TP2+CP2. The first two methods are the cheapest and are of roughly equal cost. The method TP2+ABn are used in, for example [12, 37, 38, 48, 50] and subsequent studies. We include TP3+AB2 to test whether increasing the interpolation accuracy is worthwhile use of computational resources. We also consider the MRBF2+CP2 solution, which is an accurate and economical combination of our proposed geometric methods. Finally, the TP2+CP2 method is considered to emphasize the negative implications of using a non-divergence-free interpolation method with the CP2 method.

Six simulations are performed, three with Stokes numbers of $St = 1/10, 1, 10$ for spherical particles ($\lambda = 1$) and three with the same Stokes numbers for non-spherical particles ($\lambda = 10$). At the end of the simulation, the average spatial error $\overline{\Delta\mathbf{x}_n}$, relative entropy $E(P_n)$ and the first Wasserstein distance $W(P_n)$ between the numerical distribution and the reference distribution are computed and presented in table 3.6.5 along with the time step and total simulation times used.

Figures 3.6.4a to 3.6.4f show the final distribution of the $St = 1/10$, $\lambda = 10$ simulation looking down the z-axis. We see that the only methods that look similar to the green reference solution are the MRBF1+CP1, MRBF2+CP2 and TP3+AB2 solutions. That is, both of our geometric methods and the most costly conventional method. It is worth noting that the TP2+CP2 and the TP2+AB2 solutions look very similar, despite one being generated by the CP2 method. This is likely due to the fact that the CP2 method loses its centrifuge-preserving properties when the fluid field is not divergence-free as seen in table 3.5.2 as well as the interpolation method being the dominant source of error. Turning our attention towards the corresponding part of table 3.6.6 (i.e., the top-right) we make a few remarks. The most striking one here is that the MRBF1+CP1 solution has a larger $\overline{\Delta\mathbf{x}_n}$ than the TP3+AB2 solution, but its $E(P_n)$ and $W(P_n)$ are both far lower. This is in agreement with the fact that the TP3+AB2 method is better in reducing error in the conventional sense, (i.e., the average 2-norm of the particle position errors $\overline{\Delta\mathbf{x}_n}$), but does a poorer job at reproducing the mechanisms that are responsible for the preferential distribution of particles (i.e., the centrifuge effect and the sum of the Lyapunov spectrum). We can make similar remarks for the $St = 1/10$, $\lambda = 1$ experiment where, despite having higher $\overline{\Delta\mathbf{x}_n}$, the MRBF1+CP1 method has a lower $E(P_n)$ and $W(P_n)$ than the TP3+AB2 method. These significant observations are prevalent in both $St < 1$ experiments, which corresponds to a more stiff vector field with greater relative influence of fluid inertia.

We finish with some general observations. First, the MRBF1+CP2 solutions outperform the TP1+FE1, TP2+AB2 and TP2+CP2 methods in all three measures. Second, the TP2+AB2 and TP2+CP2 methods perform about the same in each simulation, suggesting there is little advantage from using the CP2 method in conjunction with the TP2 solution. Finally, we note that the MRBF2+CP2 method is more accurate than the TP2+AB2 and the TP3+AB3 in all cases. The only exception is the $St = 1$, $\lambda = 1$ simulation where the MRBF2+CP2 method has worse $\overline{\Delta\mathbf{x}_n}$ than the TP3+AB3 solution.

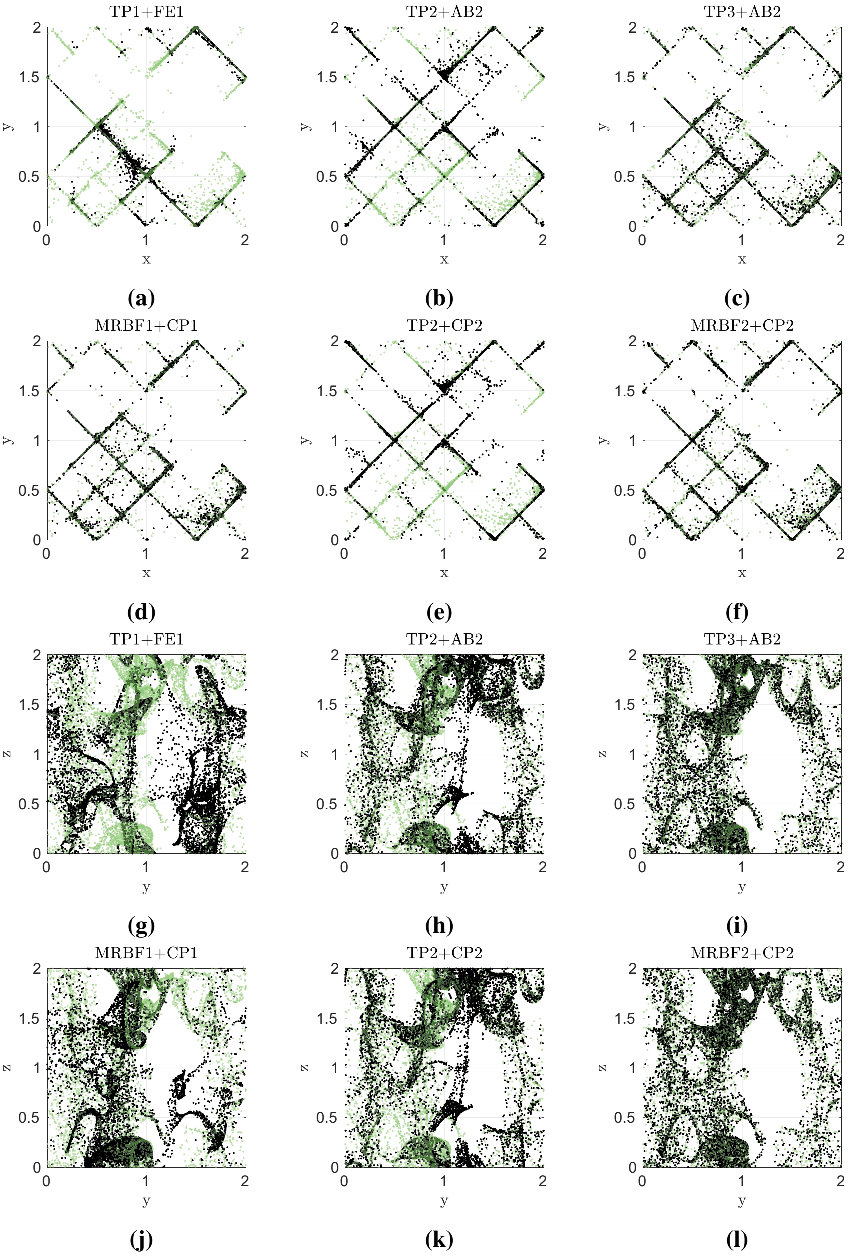


Figure 3.6.4: Figures (a) through (f) show the spatial distribution of the particles in the $x-y$ plane for the $St = 1/10$, $h = 1/100$, $T = 6$, $\lambda = 10$ simulation from table 3.6.6. Figures (g) through (l) show the spatial distribution of the particles in the $z-y$ plane for the $St = 1$, $h = 1/40$, $T = 8$, $\lambda = 1$ simulation. The reference solution is plotted in green in all figures.

3.7 Conclusions

A novel combination of geometric numerical methods for calculating accurate distributions of inertial particles in viscous flows is proposed. The algorithm consists of MRBFs to construct a divergence-free approximation of the background flow field and a geometric splitting method for the time integration.

The splitting method is shown to preserve the sum of the Lyapunov spectrum and hence the contractivity of phase space volume. By expanding the exact solution we derive an expression for how a physical volume of particles change over a small time step h, with which we recover the centrifuge effect at $O(h^4)$. We show that when a divergence-free interpolation method is used, one can implement a so-called centrifuge-preserving splitting method that preserves not only the qualitative but also the quantitative behavior of this centrifuge effect. Moreover, it is shown that errors to the divergence of the fluid field can overshadow this effect when a conventional polynomial interpolation method is used, for example.

It is shown through numerical experiments that MRBF interpolation yields particle distributions that are more similar to the exact solution than standard TP interpolation. In many examples, this is observed even when the MRBF solution has higher error per particle. This is, in part, explained by the fact that: (1) MRBF interpolation is divergence-free meaning that the numerical time integration methods mimic the qualitative centrifuge effect; and (2) MRBF interpolation produces a vector field that solves the Stokes equations, meaning that the background flow field more physically resembles that of the exact solution (e.g., the flow field is related to the gradient of a scalar pressure function).

Furthermore, we see that the proposed centrifuge-preserving methods are superior to the standard methods in terms of error per particle and how closely the particle distribution resembles the exact distribution. This is true, remarkably so, even when comparing the order-one CP1 method to the order-two RK2 and AB2 methods. In particular, for experiments with low particle inertia, the MRBF1+CP1 method produces more accurate distributions of particles than the expensive TP3+AB2 solutions, despite having slightly worse error and being an order one integration method that uses far less data points for the interpolation step.

These observations strongly suggest that preserving certain physical features of

ODEs under study in the numerical solution is of importance when simulating inertial particles in discrete flow fields. Of particular interest for future studies would be to implement the proposed methods in a physically realistic flow fields generated by a direct numerical simulation of homogeneous isotropic turbulence or turbulent channel flow, for example.

3.8 Acknowledgments

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (No. 691070). B. K. Tapley, E. Celledoni and B. Owren would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme *Geometry, compatibility and structure preservation in computational differential equations* (2019) where part of the work on this paper was undertaken.

		$\lambda = 1$			$\lambda = 1/10$		
	P_n	$E(P_n)$	$W(P_n)$	$\overline{\Delta \mathbf{x}_n}$	$E(P_n)$	$W(P_n)$	$\overline{\Delta \mathbf{x}_n}$
$St = \frac{1}{5}$ $h = \frac{1}{50}$ $T = 4$	exact+FE1	7.0143	0.4466	0.5534	–	–	–
	exact+LT1	0.4596	0.0070	0.0098	2.3929	0.1793	0.2086
	exact+CP1	0.1212	0.0049	0.0065	2.3100	0.1750	0.2048
	exact+AB2	4.0952	0.0697	0.0585	–	–	–
	exact+RK2	1.6791	0.0234	0.0215	–	–	–
	exact+CP2	0.0581	0.0022	0.0027	0.5339	0.0666	0.0983
	exact+SS2	0.1169	0.0048	0.0062	2.2956	0.1739	0.2039
$St = 1$ $h = \frac{1}{20}$ $T = 8$	exact+FE1	6.9916	1.6321	2.5224	0.9086	0.6445	1.7383
	exact+LT1	0.1538	0.0929	1.0000	0.6645	0.4788	1.2536
	exact+CP1	0.1061	0.1034	1.0084	0.5716	0.4293	1.2554
	exact+AB2	0.4833	0.2715	1.4977	–	–	–
	exact+RK2	0.2696	0.1566	1.3083	1.3583	0.4631	1.4790
	exact+CP2	0.0786	0.0558	0.5512	0.2505	0.2892	1.0857
	exact+SS2	0.1435	0.0890	1.0044	0.6374	0.4858	1.2535
$St = 10$ $h = \frac{1}{5}$ $T = 20$	exact+FE1	6.8070	2.7326	2.8528	1.3145	1.9214	4.8392
	exact+LT1	0.3531	0.1626	0.6588	0.1002	0.5384	5.1653
	exact+CP1	0.3014	0.1619	0.6479	0.0658	0.5721	5.1883
	exact+AB2	0.7651	0.3095	0.8671	0.8826	1.9070	4.9994
	exact+RK2	0.5804	0.2522	0.9074	0.1593	0.7891	5.2612
	exact+CP2	0.0733	0.0667	0.2919	0.0711	0.2761	4.9678
	exact+SS2	0.3157	0.1567	0.6700	0.0992	0.5433	5.2206

Table 3.6.4: The relative entropy $E(P_n)$, first Wasserstein distance $W(P_n)$ and $\overline{\Delta \mathbf{x}_n}$ between the numerical distribution P_n and the reference distribution. The numerical distributions are calculated by various integration methods that use exact interpolation as shown in the second column. The first column contains the Stokes number St, time step h and simulation time T used in the six simulations. The first row contains the aspect ratio λ of the particle shape. Values with a $-$ mean that the numerical solution is unstable.

		$\lambda = 1$			$\lambda = 5$		
	P_n	$E(P_n)$	$W(P_n)$	$\overline{\Delta \mathbf{x}_n}$	$E(P_n)$	$W(P_n)$	$\overline{\Delta \mathbf{x}_n}$
$St = \frac{1}{10}$	MRBF1+CP2	0.1285	0.0860	0.3063	0.0514	0.0478	0.5389
	MRBF2+CP2	0.1002	0.0388	0.1256	0.0967	0.0657	0.3578
$h = \frac{1}{100}$	MRBF3+CP2	0.0380	0.0164	0.0614	0.0337	0.0242	0.1732
	TP1+CP2	7.9636	0.7427	0.7957	3.1871	0.5657	1.1261
$T = 4$	TP2+CP2	3.7166	0.2975	0.3959	1.5779	0.3409	0.8653
	TP3+CP2	0.3349	0.0625	0.1197	0.0778	0.0704	0.3468
$St = 1$	MRBF1+CP2	0.9861	0.5802	1.4746	0.0402	0.0965	1.3596
	MRBF2+CP2	0.0444	0.0439	0.5004	0.0351	0.0698	0.7334
$h = \frac{1}{40}$	MRBF3+CP2	0.0367	0.0353	0.2442	0.0258	0.0564	0.3755
	TP1+CP2	1.6501	0.7003	1.5487	0.4281	0.3222	1.8162
$T = 8$	TP2+CP2	1.5784	1.2164	1.8284	0.0585	0.1871	1.7269
	TP3+CP2	0.0404	0.0440	0.3332	0.0310	0.0714	0.8375
$St = 10$	MRBF1+CP2	1.5452	0.0979	0.1223	0.1401	0.0726	0.1998
	MRBF2+CP2	0.1071	0.0109	0.0154	0.0178	0.0183	0.0548
$h = \frac{1}{10}$	MRBF3+CP2	0.0416	0.0041	0.0065	0.0112	0.0084	0.0216
	TP1+CP2	5.8812	0.1545	0.1674	0.6419	0.1957	0.5617
$T = 12$	TP2+CP2	4.0693	0.1540	0.2427	0.0955	0.0537	0.2222
	TP3+CP2	0.7464	0.0135	0.0125	0.0339	0.0242	0.0946

Table 3.6.5: The relative entropy $E(P_n)$, first Wasserstein distance $W(P_n)$ and average error per particle $\overline{\Delta \mathbf{x}_n}$ between the numerical distribution P_n and the reference distribution. The numerical distributions are calculated by various interpolation methods that use CP2 integration as shown in the second column. The first column contains the Stokes number St, time step h and simulation time T used in the six simulations. The first row contains the aspect ratio λ of the particle shape.

		$\lambda = 1$			$\lambda = 10$		
	P_n	$E(P_n)$	$W(P_n)$	$\overline{\Delta \mathbf{x}_n}$	$E(P_n)$	$W(P_n)$	$\overline{\Delta \mathbf{x}_n}$
$St = \frac{1}{10}$	MRBF1+CP1	0.1305	0.1050	0.3351	0.0423	0.0431	0.5453
	TP1+FE1	7.9359	0.6340	0.7866	3.2075	0.7961	1.2645
$h = \frac{1}{100}$	MRBF2+CP2	0.1249	0.0414	0.1242	0.0706	0.0456	0.3513
	TP2+CP2	2.7166	0.2998	0.3990	1.2931	0.3131	0.8785
$T = 6$	TP2+AB2	2.6957	0.2836	0.3805	1.2585	0.2852	0.8833
	TP3+AB2	2.9179	0.1339	0.2512	0.1326	0.0804	0.4506
$St = 1$	MRBF1+CP1	0.9463	0.5888	1.4858	0.0395	0.1000	1.3961
	TP1+FE1	4.3703	1.4721	2.1060	0.4507	0.5446	1.9263
$h = \frac{1}{40}$	MRBF2+CP2	0.0507	0.0525	0.5055	0.0306	0.0753	0.7845
	TP2+CP2	1.6212	1.2123	1.8251	0.0534	0.1630	1.7856
$T = 8$	TP2+AB2	1.1982	1.0943	1.7759	0.0532	0.1595	1.7693
	TP3+AB2	0.0589	0.0532	0.4793	0.0376	0.0865	1.0213
$St = 10$	MRBF1+CP1	1.2748	0.3379	0.5122	0.0727	0.1190	0.7322
	TP1+FE1	6.1882	0.7767	0.8750	1.9416	0.6882	1.5169
$h = \frac{1}{10}$	MRBF2+CP2	0.0979	0.0348	0.0663	0.0262	0.0438	0.2688
	TP2+CP2	3.2931	0.4742	0.7083	0.0754	0.1176	0.8004
$T = 16$	TP2+AB2	3.2219	0.4063	0.6304	0.0774	0.1194	0.7933
	TP3+AB2	0.2631	0.0575	0.0865	0.0397	0.0594	0.4864

Table 3.6.6: The relative entropy $E(P_n)$, first Wasserstein distance $W(P_n)$ and average error per particle $\overline{\Delta \mathbf{x}_n}$ between the numerical distribution P_n and the reference distribution. The numerical distributions are calculated by various combinations of integration and interpolation methods as shown in the second column. The first column contains the Stokes number St, time step h and simulation time T used in the six simulations. The first row contains the aspect ratio λ of the particle shape.

Bibliography

[1] *The Earth Mover's Distance.* `https://se.mathworks.com/matlabcentral/fileexchange/22962-the-earth-mover-s-distance`. Accessed: 01-02-2019.

[2] H. AKAIKE, *Block toeplitz matrix inversion*, SIAM Journal on Applied Mathematics, 24 (1973), pp. 234–241.

[3] H. I. ANDERSSON, E. CELLEDONI, L. OHM, B. OWREN, AND B. K. TAPLEY, *An integral model based on slender body theory, with applications to curved rigid fibers*, arXiv preprint arXiv:2012.11561, (2020).

[4] S. BALACHANDAR AND M. MAXEY, *Methods for evaluating fluid velocities in spectral simulations of turbulence*, Journal of Computational Physics, 83 (1989), pp. 96 – 125.

[5] J. BEC, *Fractal clustering of inertial particles in random flows*, Physics of fluids, 15 (2003), pp. L81–L84.

[6] L. BERGOUGNOUX, G. BOUCHET, D. LOPEZ, AND E. GUAZZELLI, *The motion of solid spherical particles falling in a cellular flow field at low stokes number*, Physics of Fluids, 26 (2014), p. 093302.

[7] P. S. BERNARD, M. F. ASHMAWEY, AND R. A. HANDLER, *An analysis of particle trajectories in computer-simulated turbulent channel flow*, Physics of Fluids A: Fluid Dynamics, 1 (1989), pp. 1532–1540.

[8] H. BRENNER, *The stokes resistance of an arbitrary particle*, Chemical Engineering Science, 18 (1963), pp. 1–25.

[9] M. D. BUHMANN, *Radial basis functions: theory and implementations*, vol. 12, Cambridge university press, 2003.

[10] R. CARLSON AND C. HALL, *Error bounds for bicubic spline interpolation*, Journal of Approximation Theory, 7 (1973), pp. 41–47.

[11] E. Celledoni, F. Fassò, N. Säfström, and A. Zanna, *The exact computation of the free rigid body motion and its use in splitting methods*, SIAM Journal on Scientific Computing, 30 (2008), pp. 2084–2112.

[12] N. R. Challabotla, L. Zhao, and H. I. Andersson, *Orientation and rotation of inertial disk particles in wall turbulence*, Journal of Fluid Mechanics, 766 (2015).

[13] R. Cortez, *The method of regularized stokeslets*, SIAM Journal on Scientific Computing, 23 (2001), pp. 1204–1225.

[14] J. W. Deardorff and R. L. Peskin, *Lagrangian statistics from numerically integrated turbulent shear flow*, The Physics of Fluids, 13 (1970), pp. 584–595.

[15] E. Hairer, C. Lubich, G. Wanner, *Geometric Numerical Integration, Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, second ed., 2006.

[16] T. Elperin, N. Kleeorin, and I. Rogachevskii, *Turbulent thermal diffusion of small inertial particles*, Physical review letters, 76 (1996), p. 224.

[17] M. Esmaily-Moghadam and A. Mani, *Analysis of the clustering of inertial particles in turbulent flows*, Phys. Rev. Fluids, 1 (2016), p. 084202.

[18] C. Gobert and M. Manhart, *Numerical experiments for quantification of small-scale effects in particle-laden turbulent flow*, in High Performance Computing in Science and Engineering, Garching/Munich 2009, Springer, 2010, pp. 77–88.

[19] C. Gobert, F. Schwertfirm, and M. Manhart, *Lagrangian scalar tracking for laminar micromixing at high schmidt numbers*, in ASME Fluids Engineering Division Summer Meeting, vol. 47500, 2006, pp. 1053–1062.

[20] V. Grimm and G. Quispel, *Geometric integration methods that unconditionally contract volume*, Applied numerical mathematics, 58 (2008), pp. 1103–1112.

[21] W. G. Hoover, C. G. Tull, and H. A. Posch, *Negative lyapunov exponents for dissipative systems*, Physics Letters A, 131 (1988), pp. 211–215.

[22] R. H. A. IJZERMANS, E. MENEGUZ, AND M. W. REEKS, *Segregation of particles in incompressible random flows: singularities, intermittency and random uncorrelated motion*, Journal of Fluid Mechanics, 653 (2010), p. 99–136.

[23] A. ISERLES, G. QUISPEL, AND P. TSE, *B-series methods cannot be volume-preserving*, BIT Numerical Mathematics, 47 (2007), pp. 351–378.

[24] R. JAYARAM, Y. JIE, L. ZHAO, AND H. I. ANDERSSON, *Clustering of inertial spheres in evolving taylor–green vortex flow*, Physics of Fluids, 32 (2020), p. 043306.

[25] G. B. JEFFERY, *The motion of ellipsoidal particles immersed in a viscous fluid*, Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character, 102 (1922), pp. 161–179.

[26] S. KULLBACK AND R. A. LEIBLER, *On information and sufficiency*, The annals of mathematical statistics, 22 (1951), pp. 79–86.

[27] F. LEKIEN AND J. MARSDEN, *Tricubic interpolation in three dimensions*, International Journal for Numerical Methods in Engineering, 63 (2005), pp. 455–471.

[28] K. LUO, J. FAN, AND K. CEN, *Pressure-correlated dispersion of inertial particles in free shear flows*, Physical Review E, 75 (2007), p. 046309.

[29] F. MACKAY, R. MARCHAND, AND K. KABIN, *Divergence-free magnetic field interpolation and charged particle trajectory integration*, Journal of Geophysical Research: Space Physics, 111 (2006).

[30] M. R. MAXEY, *The gravitational settling of aerosol particles in homogeneous turbulence and random flow fields*, Journal of Fluid Mechanics, 174 (1987), p. 441–465.

[31] M. R. MAXEY, *The motion of small spherical particles in a cellular flow field*, The Physics of Fluids, 30 (1987), pp. 1915–1928.

[32] M. R. MAXEY AND J. J. RILEY, *Equation of motion for a small rigid sphere in a nonuniform flow*, The Physics of Fluids, 26 (1983), pp. 883–889.

[33] R. I. MCLACHLAN AND G. QUISPEL, *Numerical integrators that contract volume*, Applied numerical mathematics, 34 (2000), pp. 253–260.

[34] D. MEYER AND P. JENNY, *Conservative velocity interpolation for pdf methods*, in PAMM: Proceedings in Applied Mathematics and Mechanics, vol. 4, Wiley Online Library, 2004, pp. 466–467.

[35] P. MORTENSEN, H. ANDERSSON, J. GILLISSEN, AND B. BOERSMA, *Dynamics of prolate ellipsoidal particles in a turbulent channel flow*, Physics of Fluids, 20 (2008), p. 093302.

[36] A. OBERBECK, *Uber stationare flussigkeitsbewegungen mit berucksichtigung der inner reibung*, J. reine angew. Math., 81 (1876), pp. 62–80.

[37] Y. PAN AND S. BANERJEE, *Numerical simulation of particle interactions with wall turbulence*, Physics of Fluids, 8 (1996), pp. 2733–2755.

[38] L. M. PORTELA AND R. V. A. OLIEMANS, *Eulerian-Lagrangian DNS/LES of particle-turbulence interactions in wall-bounded flows*, International Journal for Numerical Methods in Fluids, 43, pp. 1045–1065.

[39] D. W. ROUSON AND J. K. EATON, *On the preferential concentration of solid particles in turbulent channel flow*, Journal of Fluid Mechanics, 428 (2001), p. 149.

[40] X. RUAN, S. CHEN, AND S. LI, *Structural evolution and breakage of dense agglomerates in shear flow and taylor-green vortex*, Chemical Engineering Science, 211 (2020), p. 115261.

[41] Y. RUBNER, C. TOMASI, AND L. J. GUIBAS, *The earth mover's distance as a metric for image retrieval*, International Journal of Computer Vision, 40 (2000), pp. 99–121.

[42] M. SHAPIRO AND M. GOLDENBERG, *Deposition of glass fiber particles from turbulent air flow in a pipe*, Journal of aerosol science, 24 (1993), pp. 65–87.

[43] C. SIEWERT, R. KUNNEN, M. MEINKE, AND W. SCHRÖDER, *Orientation statistics and settling velocity of ellipsoids in decaying turbulence*, Atmospheric research, 142 (2014), pp. 45–56.

[44] K. D. SQUIRES AND J. K. EATON, *Preferential concentration of particles by turbulence*, Physics of Fluids A: Fluid Dynamics, 3 (1991), pp. 1169–1178.

[45] B. TAPLEY, E. CELLEDONI, B. OWREN, AND H. I. ANDERSSON, *A novel approach to rigid spheroid models in viscous flows using operator splitting methods*, Numerical Algorithms, (2019), pp. 1–19.

[46] G. I. TAYLOR AND A. E. GREEN, *Mechanism of the production of small eddies from large ones*, Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences, 158 (1937), pp. 499–521.

[47] W. Uijttewaal and R. Oliemans, *Particle dispersion and deposition in direct numerical and large eddy simulations of vertical pipe flows*, Physics of Fluids, 8 (1996), pp. 2590–2604.

[48] B. A. van Haarlem, *The dynamics of particles and droplets in atmospheric turbulence-A numerical study*, PhD thesis, 2000.

[49] H. Wang, R. Agrusta, and J. van Hunen, *Advantages of a conservative velocity interpolation (cvi) scheme for particle-in-cell methods with application in geodynamic modeling*, Geochemistry, Geophysics, Geosystems, 16 (2015), pp. 2015–2023.

[50] Q. Wang and K. D. Squires, *Large eddy simulation of particle-laden turbulent channel flow*, Physics of Fluids, 8 (1996), pp. 1207–1223.

[51] H. Wendland, *Scattered data approximation*, vol. 17, Cambridge university press, 2004.

[52] M. Wilkinson and B. Mehlig, *Caustics in turbulent aerosols*, EPL (Europhysics Letters), 71 (2005), p. 186.

[53] P. K. Yeung and S. B. Pope, *Lagrangian statistics from direct numerical simulations of isotropic turbulence*, Journal of Fluid Mechanics, 207 (1989), pp. 531–586.

[54] L. Zhao, N. R. Challabotla, H. I. Andersson, and E. A. Variano, *Rotation of nonspherical particles in turbulent channel flow*, Physical review letters, 115 (2015), p. 244501.

Appendix

3.A Non-spherical particle model

Here, we give details of the specific rigid spheroid model that is used in the numerical experiments. The surface of a spheroid is defined by the equation

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} + \frac{z^2}{c^2} = 1, \tag{3.A.1}$$

where a and c are the distinct semi-axis lengths. The particle shape is characterised by the dimensionless aspect ratio $\lambda = c/a > 0$, which distinguishes between spherical ($\lambda = 1$), prolate ($\lambda > 1$) and oblate ($\lambda < 1$) particles (the latter two shapes are also called as rods and disks).

An inertial particle immersed in a fluid will experience forces on its surface that have magnitude governed by many parameters such as the particles density ρ_p, length a, fluid density ρ_f, kinematic viscosity ν and fluid time scale τ_f. The particle Stokes number is formally defined as the ratio of the particle and fluid time scales $St = \tau_p/\tau_f$. For a spherical particle the Stokes number is

$$St_0 = \frac{2Da^2}{9\nu\tau_f}, \tag{3.A.2}$$

where $D = \rho_p/\rho_f$ is the particle-fluid density ratio. Note that this definition only depends on the particle size and inertia. For spheroidal particles, the following shape dependent Stokes numbers are used, which are derived by Shapiro and Goldenberg [42] and Zhao, et al. [54]

$$St = \begin{cases} St_0\,\lambda\log(\lambda + \sqrt{\lambda^2 - 1})/\sqrt{\lambda^2 - 1} & \text{for} \quad \lambda > 1 \\ St_0\,(\pi - k_0)/(2\sqrt{1 - \lambda^2}) & \text{for} \quad \lambda < 1 \end{cases} \tag{3.A.3}$$

where $k_0 = \log((\lambda - \sqrt{\lambda^2 - 1})/(\lambda + \sqrt{\lambda^2 - 1}))$. Note that $St \to St_0$ as $\lambda \to 1$ from above or below. All the following equations are implemented in their non-dimensional form and all parameters have dimension equal to 1.

The particle experiences a hydrodynamic drag force due to Brenner [8],

$$\mathbf{F} = QK_bQ^{\mathrm{T}}(\mathbf{u} - \mathbf{v}), \tag{3.A.4}$$

where $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ is the fluid velocity evaluated at the particle center of mass $\mathbf{x}$ and $\mathbf{v} = \mathbf{p}/m$ is the particle velocity. The body frame resistance tensor K_b was calculated by Oberbeck [36], is diagonal, positive definite and given by

$$K_b = 16\pi\lambda \ \mathrm{diag}\left(\frac{1}{\chi_0 + \alpha_0}, \frac{1}{\chi_0 + \beta_0}, \frac{1}{\chi_0 + \lambda^2\gamma_0}\right) \tag{3.A.5}$$

where the constants χ_0, α_0, β_0 and γ_0 were calculated for ellipsoidal particles by Siewert et al. [43] and are presented in table 3.A.1

	$\lambda < 1$	$\lambda = 1$	$\lambda > 1$
χ_0	$\frac{\lambda^2(\pi-\kappa_0)}{\sqrt{1-\lambda^2}}$	2	$\frac{-\kappa_0\lambda}{\sqrt{\lambda^2-1}}$
$\alpha_0 = \beta_0$	$\frac{-\lambda\left(\kappa_0-\pi+2\lambda\sqrt{1-\lambda^2}\right)}{2(1-\lambda^2)^{3/2}}$	$\frac{2}{3}$	$\frac{\lambda^2}{\lambda^2-1} + \frac{\lambda\kappa_0}{2(\lambda^2-1)^{3/2}}$
γ_0	$\frac{\left(\lambda(\kappa_0-\pi)+2\sqrt{1-\lambda^2}\right)}{(1-\lambda^2)^{3/2}}$	$\frac{2}{3}$	$\frac{-2}{\lambda^2-1} - \frac{\lambda\kappa_0}{(\lambda^2-1)^{3/2}}$
κ_0	$2\arctan\left(\frac{\lambda}{\sqrt{1-\lambda^2}}\right)$	1	$\ln\left(\frac{\lambda-\sqrt{\lambda^2-1}}{\lambda+\sqrt{\lambda^2-1}}\right)$

Table 3.A.1: The expressions for the constants χ_0, α_0, β_0 and γ_0 for $\lambda < 1$, $\lambda = 1$ and $\lambda > 1$.

The torque vector $\mathbf{T}$ depends on the particle shape and the local fluid velocity derivatives, and is given in non-dimensional form by [25]

$$T_x = \frac{16\pi\lambda}{3(\beta_0 + \lambda^2\gamma_0)}\left[(1-\lambda^2)S_{yz} + (1+\lambda^2)(\Omega_x - \omega_y)\right], \tag{3.A.6}$$

$$T_y = \frac{16\pi\lambda}{3(\alpha_0 + \lambda^2\gamma_0)}\left[(\lambda^2-1)S_{zx} + (1+\lambda^2)(\Omega_y - \omega_z)\right], \tag{3.A.7}$$

$$T_z = \frac{32\pi\lambda}{3(\alpha_0 + \beta_0)}(\Omega_z - \omega_z). \tag{3.A.8}$$

A slender body model for thin rigid fibers: validation and comparison

Laurel Ohm, Benjamin K Tapley, Helge I Andersson, Elena Celledoni, Bryjulf Owren

Proceedings of MEKiT'19, 10th National Conference on Computational Mechanics Mech., 2019, pp 343 - 366

A slender body model for thin rigid fibers: validation and comparisons

Abstract. In this paper we consider a computational model for the motion of thin, rigid fibers in viscous flows based on slender body theory. Slender body theory approximates the fluid velocity field about the fiber as the flow due to a distribution of singular solutions to the Stokes equations along the fiber centerline. The velocity of the fiber itself is often approximated by an asymptotic limit of this expression. Here we investigate the efficacy of simply evaluating the slender body velocity expression on a curve along the surface of the actual 3D fiber, rather than limiting to the fiber centerline. Doing so may yield an expression better suited for numerical simulation. We validate this model for two simple geometries, namely, thin ellipsoids and thin rings, and we compare the model to results in the literature for constant and shear flow. In the case of a fiber with straight centerline, the model coincides with the prolate spheroid model of Jeffery. For the thin torus, the computed force agrees with the asymptotically accurate values of Johnson and Wu and gives qualitatively similar dynamics to oblate spheroids of similar size and inertia

4.1 Introduction

Understanding the dynamics of particles immersed in viscous fluids is of importance in many areas of nature and industry. The first problem one encounters when simulating the dynamics of particles with complicated shapes is determining an appropriate model. As the forces and torques of arbitrarily shaped particles are not known in general, one must make a number of assumptions on the particle size and shape to accurately and cheaply specify the forces and torques on the particle. If the particle length scale is small (for example, smaller than the Kolmogorov scale in turbulent flows), the local fluid velocity can be accurately approximated by creeping Stokes flow and then the problem is amenable to a number of mathematical techniques that are available in the literature. One popular technique involves implementing slender body theories to model long and thin particles. An advantage of using slender body models is that they have the freedom to model flexible and arbitrarily shaped particles (with free ends or closed loops) provided that the particle is thin and the parametrization of centerline is known. The theoretical assumptions on which slender body models are based are also valid for long particles whose centerline lengths are comparable or extend beyond the limiting length scales of the fluid field. In particular, slender body theory has the potential to model particles that are longer than the Kolmogorov scale, where conventional models such as the Jeffery model for ellipsoids are not valid. This is a major advantage over current state-of-the-art particle simulations in, for example, [23, 30]. We also refer to [28] and references therein for a review of other available models and methodologies for treating anisotropic particles in turbulent flows.

In this article, we will consider a model based on slender body theory for rigid fibers that have either free ends or are closed loops. The purpose of this paper is primarily to provide a numerical validation of the proposed slender body model. For this reason, we will primarily focus on two simple geometries: long ellipsoids and thin rings (also referred to as thin tori). These geometries are chosen as there are verified ellipsoid and torus models available in the literature with well-studied dynamics, see for example [23, 29] for prolate ellipsoids and [15] for thin torus models. This will serve as grounding for future work that will focus on more interesting and complex particle shapes (e.g., helical particles, complex closed loops or very long particles) in more complex flows (e.g., 3D numerical turbulence) that can be approached with more advanced numerical methods [25, 26]. Such studies could impact our understanding of the transport and deposition of microplastics in the ocean, since a large percentage of these microplastics are thin fibers [20].

The slender body approximation expresses the fluid velocity away from the fiber centerline as an integral of singular solutions to the Stokes equations along the fiber centerline. As such, the approximation itself is singular along the fiber centerline, and there exist various methods to obtain a limiting integral expression for the velocity of the slender body itself [7, 16, 17]. For the purposes of particle simulations, we are primarily interested in solving for the forces and torques on the particle given a flow about the body. In the case of slender body theory, this involves inverting the limiting integral expression for the fiber velocity to find the force per unit length. Thus we need to be careful that the limiting expression is suitable for numerical inversion. In particular, we hope to avoid the high wavenumber instabilities that arise in some of the existing centerline expressions which require additional regularization to overcome. Often the methods for regularization lack a physical justification.

Here we consider approximating the fiber velocity by simply evaluating the slender body fluid velocity expression on a curve along the actual slender body surface, away from the fiber centerline. Numerical evidence suggests that this method does not require further regularization to yield an invertible matrix equation for any discretization level or fiber centerline shape. We also show that our model agrees well with exact or asymptotically accurate expressions for the forces and torques on fibers with simple geometries in simple flows.

The next section presents the mathematical theory for the slender body formalism, as well as a brief review of rigid body mechanics and spheroidal particle models. Section 4.3 is dedicated to numerical experiments, and the final section is for conclusions.

4.2 Particle modeling

We begin by reviewing the rigid body dynamics that are relevant to particle modeling. The theoretical basis for the slender body model is then presented for rigid free ended fibers and rigid closed loops. Finally, we present the Jeffery model for torques on an ellipsoid, which is used for comparison purposes.

4.2.1 Dynamics

The angular momentum $\boldsymbol{m}$ of a rigid particle with torque $\boldsymbol{N}$ is governed by the ordinary differential equation

$$\dot{\boldsymbol{m}} = \boldsymbol{m} \times \boldsymbol{\omega} + \boldsymbol{N}, \tag{4.2.1}$$

where $\boldsymbol{\omega} = J^{-1}\boldsymbol{m}$ is the angular velocity and J is the diagonal moment of inertia tensor. All the above quantities are given in the particle frame of reference. The particle orientation (with respect to a fixed inertial frame of reference) is specified using Euler parameters $q \in \mathbb{R}^4$ which satisfy the constraint $||q||_2 = 1$ and are determined by solving the ODE

$$\dot{q} = \frac{1}{2} q \cdot w, \tag{4.2.2}$$

where $w = (0, \boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^4$ and $\cdot$ here denotes the Hamilton product of two quaternions [10]. A vector in the particle reference frame $\mathbf{x}_p$ can be rotated to a vector in an inertial co-translating reference frame $\mathbf{x}_T = Q\mathbf{x}_p$ where Q is the rotation matrix that is the image of q under the Euler-Rodriguez map. We refer the reader to [10] for details on quaternion algebra and rigid body mechanics.

4.2.2 Slender body theory

We begin by describing the slender body geometries that will be considered in the free end and closed loop settings. To condense notation, we will use $\mathscr{I}$ to denote the interval $[-1/2, 1/2]$ in the free end setting and the unit circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ in the closed loop setting. We take $\boldsymbol{X} : \mathscr{I} \to \mathbb{R}^3$ to be the coordinates of an open or closed non-self-intersecting C^2 curve in $\mathbb{R}^3$, parameterized by arclength s. We let $\boldsymbol{e}_s(s) = \frac{d\boldsymbol{X}}{ds}$ denote the unit tangent vector to $\boldsymbol{X}(s)$. The curve $\boldsymbol{X}(s)$ will be the centerline of the slender body, and we assume that all cross sections of the slender body are circular.

Let $0 < \epsilon \ll 1$. In the closed loop setting, we consider fibers with uniform radius ϵ on each cross section. In the free end setting, we consider the actual endpoints of the fiber to be $\pm\sqrt{1/4+\epsilon^2}$ rather than $\pm 1/2$, and define a *radius function* $r \in C^2(-\sqrt{1/4+\epsilon^2}, \sqrt{1/4+\epsilon^2})$ such that $0 < r(s) \leq 1$ for each $s \in [-1/2, 1/2]$, and $r(s)$ decays smoothly to zero at the fiber endpoints $\pm\sqrt{1/4+\epsilon^2}$. We will mostly be concerned with the prolate spheroid, for which we have

$$r(s) = \frac{1}{(\frac{1}{4}+\epsilon^2)^{1/2}}\left(\frac{1}{4}+\epsilon^2-s^2\right)^{1/2}. \tag{4.2.3}$$

Notice that the interval $[-1/2, 1/2]$ extends from focus to focus of this prolate spheroid, and that $r = \mathcal{O}(\epsilon)$ at $s = \pm\frac{1}{2}$ (see figure 4.2.1). In numerical applications, we will also briefly consider the case of a free end fiber with uniform radius (except for hemispherical caps at the fiber endpoints – see section 4.3.2), but we note that the slender body approximation is better suited for the prolate spheroid. Throughout this paper, for the sake of conciseness, we will often write one expression to encompass both the free end and closed loop settings, in which case we note that in the closed loop setting we define $r(s) = 1$ for each $s \in \mathbb{T}$.

The idea behind slender body theory is to approximate the fluid velocity about the fiber as the Stokes flow due to a one-dimensional curve of point forces in $\mathbb{R}^3$. The basic theory originated with Hancock [12], Cox [8], and Batchelor [2] with later improvements by Keller and Rubinow [16] and Johnson [14]. Here we will consider specifically the slender body theory of Johnson, which was further studied by Götz [11] and Tornberg and Shelley [27]. Let $\mathbf{u}_0(\mathbf{x}, t)$ denote the (known) velocity of the fluid in the absence of the fiber at time t, and let μ denote the viscosity of the fluid. The classical slender body approximation $\mathbf{u}^{\mathrm{SB}}(\mathbf{x}, t)$ to the fluid velocity at any point $\mathbf{x}$ away from the fiber centerline $\boldsymbol{X}(s, t)$ is then given by

$$8\pi\mu\big(\mathbf{u}^{\mathrm{SB}}(\mathbf{x},t) - \mathbf{u}_0(\mathbf{x},t)\big) = -\int_{\mathcal{I}}\left(\mathscr{S}(\boldsymbol{R}) + \frac{\epsilon^2 r^2(s')}{2}\mathscr{D}(\boldsymbol{R})\right)\boldsymbol{f}(s',t)\,ds', \quad \boldsymbol{R} = \boldsymbol{x} - \boldsymbol{X}(s',t); \tag{4.2.4}$$

$$\mathscr{S}(\boldsymbol{R}) = \frac{\mathbf{I}}{|\boldsymbol{R}|} + \frac{\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}}}{|\boldsymbol{R}|^3}, \; \mathscr{D}(\boldsymbol{R}) = \frac{\mathbf{I}}{|\boldsymbol{R}|^3} - \frac{3\boldsymbol{R}\boldsymbol{R}^{\mathrm{T}}}{|\boldsymbol{R}|^5}. \tag{4.2.5}$$

Here $\frac{1}{8\pi\mu}\mathscr{S}(\boldsymbol{R})$ is the Stokeslet, the free space Green's function for the Stokes equations in $\mathbb{R}^3$, and $\frac{1}{8\pi\mu}\mathscr{D}(\boldsymbol{R}) = \frac{1}{16\pi\mu}\Delta\mathscr{S}(\boldsymbol{R})$ is the doublet, a higher order

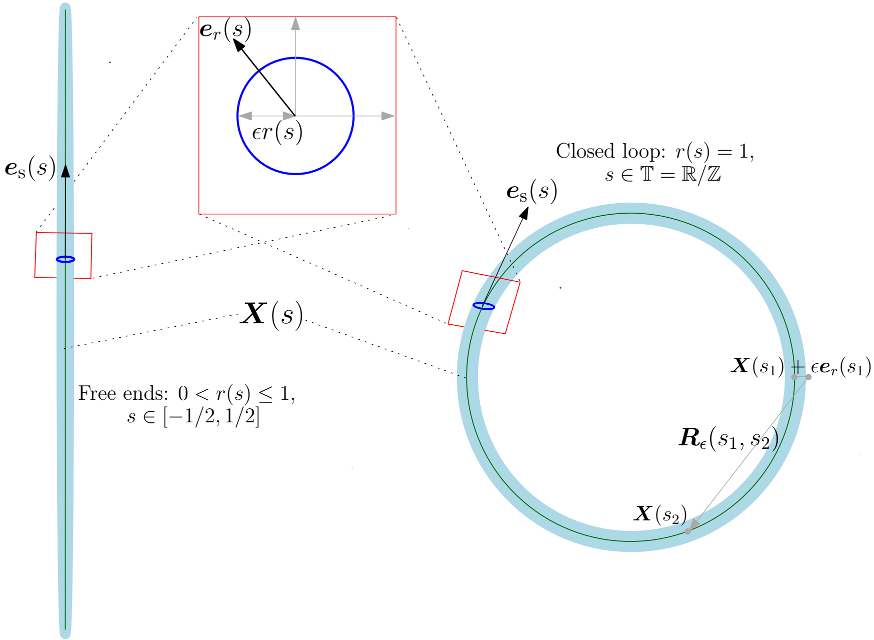


Figure 4.2.1: A depiction of the geometries under consideration in the free end and closed loop settings.

correction to the velocity approximation. The force density $\boldsymbol{f}(s,t)$ is here considered as the force per unit length exerted by the fluid on the body. The sign convention is opposite if we instead consider $\boldsymbol{f}$ to be the force exerted by the body on the fluid. Note that in the free end case, this force density is only distributed between the generalized foci of the slender body ($s = \pm 1/2$) rather than between the actual endpoints of the fiber.

In the stationary setting, Mori et al. in [21] (closed loop case) and [22] (free end case) prove a rigorous error bound for the difference between the velocity field given by (5.3.1) and the velocity field around a three-dimensional flexible rod satisfying a well-posed *slender body PDE*. In particular, for the closed loop, given a force density $\boldsymbol{f} \in C^1(\mathbb{T})$, the difference between $\mathbf{u}^{\mathrm{SB}}$ and the PDE solution exterior to the slender body is bounded by an expression proportional to $\epsilon|\log\epsilon|$. In the free end case, given a force density $\boldsymbol{f} \in C^1(-1/2, 1/2)$ which decays like a spheroid at the fiber endpoints ($\boldsymbol{f}(s) \sim \sqrt{1/4 - s^2}$ as $s \to \pm 1/2$), the difference between the free end slender body approximation $\mathbf{u}^{\mathrm{SB}}$ and the well-posed PDE solution of [22] is similarly bounded by an expression proportional to $\epsilon|\log\epsilon|$. Thus the Stokeslet/doublet expression (5.3.1) is quantitatively a good approximation of the flow field around a slender body.

To approximate the velocity of the slender body itself, we would like to use (5.3.1) to obtain an expression for the relative velocity of the fiber centerline $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$ depending only on the arclength parameter s and time t. In the case of a rigid fiber, given the velocity $\frac{\partial \boldsymbol{X}(s,t)}{\partial t} = \mathbf{v} + \boldsymbol{\omega} \times \boldsymbol{X}(s,t)$, $\mathbf{v}, \boldsymbol{\omega} \in \mathbb{R}^3$, of the filament centerline, we would like to then be able to invert the centerline velocity expression to solve for the force density $\boldsymbol{f}(s,t)$ along the fiber. We use this $\boldsymbol{f}(s,t)$ to compute the total force $\boldsymbol{F}(t)$ and torque $\boldsymbol{N}(t)$ exerted on the body as

$$\int_{\mathcal{I}} \boldsymbol{f}(s,t)\, ds = \boldsymbol{F}(t), \qquad \int_{\mathcal{I}} \boldsymbol{X}(s,t) \times \boldsymbol{f}(s,t)\, ds = \boldsymbol{N}(t). \tag{4.2.6}$$

Since the expression (5.3.1) is singular at $\mathbf{x} = \boldsymbol{X}(s,t)$, deriving a limiting expression for the fiber centerline must be done carefully. There are various ways to use (5.3.1) to obtain a centerline expression depending on s only, including the methods of Lighthill [17], Keller and Rubinow [16], and the method of regularized Stokeslets [3,6,7]. Each method expresses the velocity of the slender body centerline $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$ as an integral operator acting on the force density $\boldsymbol{f}(s,t)$. A brief overview of these methods is given in appendix 4.A.

Because solving for the force density $\boldsymbol{f}(s,t)$ given $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$ involves inverting an integral operator at each time step, we need to take particular care that the operator – at least when discretized – is suitable for inversion. In particular, we need to avoid the high wavenumber instabilities that limit discretization of the integral operator and hinder some of the asymptotic methods described in appendix 4.A. At the same time, we would like the centerline expression to have a clear physical meaning and connection to the Stokeslet/doublet expression (5.3.1).

Thus we will use the following expression to approximate the velocity $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$ of the slender body itself. Taking $\boldsymbol{e}_r(s,t)$ to be a particular unit vector normal to $\boldsymbol{X}(s,t)$ (we will discuss the choice of $\boldsymbol{e}_r$ later), we essentially evaluate (5.3.1) at $\mathbf{x} = \boldsymbol{X}(s,t) + \epsilon r(s)\boldsymbol{e}_r(s,t)$, a curve along the actual surface of the slender body. For $\mathcal{S}$, $\mathcal{D}$ as in (5.3.1), we have

$$8\pi\mu\left(\frac{\partial \boldsymbol{X}}{\partial t} - \mathbf{u}_0(\boldsymbol{X}(s,t),t)\right) = -\int_{\mathcal{I}} \left(\mathcal{S}_\epsilon(s,s',t) + \frac{\epsilon^2 r^2(s')}{2}\mathcal{D}_\epsilon(s,s',t)\right)\boldsymbol{f}(s',t)\, ds'; \tag{4.2.7}$$

$$\mathcal{S}_\epsilon = \mathcal{S}(\boldsymbol{R}_\epsilon(s,s',t)) - \frac{\epsilon^2 r^2 \boldsymbol{e}_r \boldsymbol{e}_r^{\mathrm{T}}}{|\boldsymbol{R}_\epsilon(s,s',t)|^3}, \qquad \mathcal{D}_\epsilon = \mathcal{D}(\boldsymbol{R}_\epsilon(s,s',t)) + \frac{3\epsilon^2 r^2 \boldsymbol{e}_r \boldsymbol{e}_r^{\mathrm{T}}}{|\boldsymbol{R}_\epsilon(s,s',t)|^5}, \tag{4.2.8}$$

$$\boldsymbol{R}_\epsilon(s,s',t) = \boldsymbol{X}(s,t) - \boldsymbol{X}(s',t) + \epsilon r(s)\boldsymbol{e}_r(s,t). \tag{4.2.9}$$

Here we are relying on the fact that for any point $\mathbf{x}$ on the actual fiber surface, the expression (5.3.1) for $\mathbf{u}^{\mathrm{SB}}(\mathbf{x})$ is designed to depend only on arclength s to leading order in ϵ – in particular, on each cross section of the slender body, the angular dependence about the fiber centerline is only $\mathcal{O}(\epsilon \log \epsilon)$ (see [21], proposition 3.9, and [22], proposition 3.11). This is because the leading order angular-dependent terms (the $\epsilon^2 r^2 \boldsymbol{e}_r \boldsymbol{e}_r^{\mathrm{T}}$ term in both the Stokeslet and the doublet, which is $\mathcal{O}(1)$ at $s = s'$) cancel each other asymptotically to order $\epsilon \log(\epsilon)$ (see estimates 3.62 and 3.65 in [21] and estimates 3.40 and 3.43 in [22]). We therefore eliminate these two terms from the formulation (5.3.4), in part due to this cancellation and in part because their omission appears to improve the stability of the discretized integral operator (5.3.4) when n, the number of discretization points, is large. This apparent improvement in stability merits further study in future work.

Thus to approximate the velocity of the fiber centerline, we evaluate (5.3.1) on the actual slender body surface along a normal vector $\boldsymbol{e}_r(s,t) \in C^2(\mathscr{I})$ extending from $\boldsymbol{X}(s,t)$ but cancel the $\epsilon^2 r^2 \boldsymbol{e}_r \boldsymbol{e}_r^{\mathrm{T}}$ terms that would otherwise appear. Note that the choice of normal vector $\boldsymbol{e}_r$ is somewhat arbitrary, and does have an $\mathcal{O}(\epsilon \log \epsilon)$ effect on the resulting approximation. These effects can and should be studied further in future work. However, we use this normal vector as a physically meaningful means of avoiding the high wavenumber instabilities that appear in other asymptotic methods (see appendix 4.A). Numerical evidence suggests that the discretized centerline equation (5.3.4) yields a matrix equation that is solvable for $\boldsymbol{f}(s,t)$ given $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$, as all eigenvalues of the matrix are positive even for very large n. This is not necessarily the case for some of the other centerline equations (again, see appendix 4.A) unless additional regularizations are added, which may affect the physical meaning of the equations. The possibility of resolving very fine scales along the length of the fiber is desirable especially when dealing with turbulent flows.

4.2.3 Spheroid model

The above slender body model is valid for arbitrary parameterizations of the centerline $\boldsymbol{X}(s,t)$ and a wide choice of radius functions. However, to validate the model we will focus on a simple case where the centerline is a straight line and the radius function corresponds to an ellipsoid. In this case the torques have a known expression due to Jeffery [13] and the motion of such a particle in simple flows is well-known [4, 19] which makes this choice of geometry a perfect arena for model validation. We will now briefly review some theory related to spheroids immersed in viscous fluids.

An axisymmetric spheroid in the particle frame is given by

$$\frac{x^2}{a^2}+\frac{y^2}{a^2}+\frac{z^2}{b^2}=1, \tag{4.2.10}$$

where a and b are the distinct semi-axis lengths. The particle shape is characterized by the dimensionless aspect ratio $\lambda = b/a > 0$, which distinguishes between spherical ($\lambda = 1$), prolate ($\lambda > 1$) and oblate ($\lambda < 1$) particles (the latter two shapes are also called as rods and disks). In the case of a slender prolate spheroid, we take $a = \epsilon$. The axisymmetric moment of inertia tensor for a spheroid in the body frame is

$$J = ma^2 \text{diag}\left(\frac{(1+\lambda^2)}{5}, \frac{(1+\lambda^2)}{5}, \frac{2}{5}\right), \tag{4.2.11}$$

where $m = \frac{4}{3}\pi\lambda a^3 \rho_p$ is the particle mass and ρ_p is the particle density. Jeffery [13] calculated the torque $\mathbf{N}$ of an ellipsoid in creeping Stokes flow, which in the above axisymmetric case reads

$$N_x = \frac{16\pi\lambda\mu a^3}{3(\beta_0+\lambda^2\gamma_0)}\left[(1-\lambda^2)S_{yz}+(1+\lambda^2)(\Omega_x-\omega_x)\right], \tag{4.2.12}$$

$$N_y = \frac{16\pi\lambda\mu a^3}{3(\alpha_0+\lambda^2\gamma_0)}\left[(\lambda^2-1)S_{zx}+(1+\lambda^2)(\Omega_y-\omega_y)\right], \tag{4.2.13}$$

$$N_z = \frac{32\pi\lambda\mu a^3}{3(\alpha_0+\beta_0)}(\Omega_z-\omega_z), \tag{4.2.14}$$

where $S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)$ is the fluid shear tensor and $\mathbf{\Omega} = \frac{1}{2}\nabla\times\mathbf{u}$ is the fluid rotation, both taking constant values in shear flow. The values α_0, β_0 and γ_0 are λ-dependent parameters that were calculated in [9].

There are a number of distinctions to make between this model and the slender body model. First, Jeffery assumes that the particle is small enough that the fluid Jacobian $\nabla\mathbf{u}$ is constant across the volume of the spheroid. In shear flow, $\nabla\mathbf{u}$ is constant everywhere, hence this assumption is true and the model validity is independent of the size of the particle. However, in more complex flows such as turbulence, the Jeffery model is only valid for $a, b << \eta$ for Kolmogorov length η. On the other hand, the slender model requires only that the maximal cross sectional radius $\epsilon << \eta$ to be valid. Hence, the slender body model is valid for particles with lengths larger than η whilst satisfying the Stokes flow assumptions. Second, the Jeffery torque depends on the fluid velocity derivatives only, while the slender body model derives the torques from the velocity field along the centerline. Because of this, we cannot expect the models to coincide when the particle is aligned exactly in the shear plane (i.e., the plane where $\mathbf{u} = \mathbf{0}$ but $\frac{\partial u_j}{\partial x_i} \neq 0$).

4.3 Numerical experiments

This section presents numerical results for the slender body model and comparisons with other similar models. We begin with a validation of the slender body expression (5.3.4) by comparing the total force $\boldsymbol{F}$ given by inverting (5.3.4) for a stationary slender body velocity with the exact expression for the Stokes drag on a particular object (when available) or with an expression valid asymptotically as $\epsilon \to 0$. We consider the slender prolate spheroid (section 4.3.2; exact expression given by Chwang and Wu [5]), the straight, uniform cylinder with hemispherical endpoints (section 4.3.2; asymptotic expression given by Keller and Rubinow [16]), and the slender torus (section 4.3.3; asymptotic expression given by Johnson and Wu [15]). In each case we expect $\mathcal{O}(\epsilon \log \epsilon)$ agreement between the force $\boldsymbol{F}$ computed using (5.3.4) and the exact or asymptotically accurate expressions; however, we find that this trend is clearly visible only in the closed loop setting. We then examine the rotational dynamics of a prolate spheroid in shear flow using expression (5.3.4) and compare it with the Jeffery model for ellipsoids [13]. We look at the dynamics of the two models for a range of aspect ratios and orientations and then explore the effect of the discretization parameter on the periodic Jeffery orbits. We finally compare the dynamics of thin rings to oblate spheroids for a range of fluid viscosities.

4.3.1 Computational considerations

In many applications, one needs to simulate the dynamics of thousands or millions of particles; hence computational cost plays a role in determining the model choice. One thing to consider is that the slender body model involves inverting a $3n \times 3n$ matrix at each time step, where n is the user-defined discretization parameter that arises from discretizing the integral in equation (5.3.4). On the other hand, the Jeffery model requires an accurate approximation of the fluid Jacobian at the location of the particle center of mass, while the slender body model only requires the fluid velocity values at the n locations on its centerline. When the fluid velocity is defined at discrete locations in space, such as in direct numerical simulations of turbulent flows, the Jeffery model is faced with the problem of approximating the fluid Jacobian at the location of the particle center of mass, which is more costly than just interpolating the velocity field. In practice, however, one should use the Jeffery model when computing dynamics of small, thin ellipsoids when possible and the slender body model for more complicated shapes or longer particles. As the purpose of this article is focused on the theoretical and numerical validation of the slender body model, computational cost and numerical methods will be left for future work.

4.3.2 Free ended fibers in constant flow

We validate the free end formulation of (5.3.4) in the case of a slender body with straight centerline $\boldsymbol{X}(s) = s\boldsymbol{e}_x$, $s \in [-1/2, 1/2]$, aligned with the x-axis. Here we will consider both the slender prolate spheroid with radius function $r(s)$ as in (5.1.2) and a slender cylinder with hemispherical caps at the fiber endpoints. In both cases, we take the actual filament length to be $2\sqrt{1/4+\epsilon^2}$, but distribute the force density $\boldsymbol{f}(s)$ only along $[-1/2, 1/2]$. As in the closed loop setting, we use (5.3.4) to calculate the drag force $\boldsymbol{F}$ on the slender body as it translates with unit speed, and compare this $\boldsymbol{F}$ to either exact or asymptotically accurate expressions for the Stokes drag on a prolate spheroid or cylinder. In both cases we will use the unit normal vector $\boldsymbol{e}_r(s) = \cos(2\pi s)\boldsymbol{e}_y + \sin(2\pi s)\boldsymbol{e}_z$, which rotates once in the yz-plane perpendicular to $\boldsymbol{X}(s) = s\boldsymbol{e}_x$ for $s \in [-1/2, 1/2]$. This normal vector is chosen because it represents a sort of average normal direction along the length of the filament.

In the free end setting, we also need to make sure that the computed force density $\boldsymbol{f}(s)$ is decaying sufficiently rapidly at the fiber endpoints to ensure that the solution makes sense physically. The inclusion of the decaying radius function $r(s)$ in the slender body velocity expression (5.3.4) ensures this decay by making the integral kernel very large near the fiber endpoints.

In the case of a prolate spheroid, we can actually compare the total force $\boldsymbol{F}$ given by (5.3.4) to the analytical expression for Stokes drag on a spheroid calculated by Chwang and Wu [5] (see table 4.3.1). We consider the drag force on a slender prolate spheroid translating with unit speed in either the y-direction (perpendicular to the semi-major axis) or the x-direction (parallel to the semi-major axis). In all cases, the integral term of (5.3.4) is discretized using the trapezoidal rule with uniform discretization along the filament centerline. We use $n = 2/\epsilon$ discretization points.

We also look at a plot of the computed force per unit length $\boldsymbol{f}(s)$ along the filament (figure 4.3.1) to verify that the force density makes sense physically.

From figure 4.3.1, we can see that the force density $\boldsymbol{f}(s)$ decays rapidly as

	$\boldsymbol{F}\cdot\boldsymbol{e}_y$ for $\mathbf{u}=\boldsymbol{e}_y$		$\boldsymbol{F}\cdot\boldsymbol{e}_x$ for $\mathbf{u}=\boldsymbol{e}_x$		
ϵ	Expression (5.3.4)	Chwang-Wu	Expression (5.3.4)	Chwang-Wu	$\epsilon\lvert\log\epsilon\rvert$
0.01	-2.4498	-2.4618	-1.5245	-1.5302	0.0461
0.005	-2.1579	-2.1673	-1.3051	-1.3094	0.0265
0.0025	-1.9281	-1.9358	-1.1408	-1.1442	0.0150
0.00125	-1.7426	-1.7491	-1.0133	-1.0159	0.0084

Table 4.3.1: Comparison of the computed (via expression (5.3.4)) and exact (from Chwang and Wu [5]) Stokes drag force $\boldsymbol{F}$ on a slender prolate spheroid of length $2\sqrt{1/4+\epsilon^2}$ with semi-major axis aligned with the x-axis. Columns 2 and 3 compare the y-component of $\boldsymbol{F}$ for a spheroid translating with unit speed in the y-direction, while columns 4 and 5 compare the x-component of $\boldsymbol{F}$ for translation in the x-direction. Note that for both directions, the force difference decreases with ϵ, but not quite at the expected $\epsilon\log\epsilon$ rate.

$s \to \pm 1/2$, but does not vanish identically at $|s| = 1/2$. However, it should be noted that in [22], we are given the force density $\boldsymbol{f}(s)$, $s \in [-1/2, 1/2]$, and use it to solve for the corresponding slender body velocity. In that case, the force must vanish identically at $\pm 1/2$ to yield a unique velocity. Since in this case we are using the fiber velocity to solve for the force density, it appears that what we are doing instead here is ignoring a certain (small) amount of force contribution from the very ends of the fiber (between $1/2 \le |s| \le \sqrt{1/4+\epsilon^2}$). Whether or not this is a good approximation is unclear – it is possible that the same force density could result from flows that differ slightly at the actual fiber endpoints. However, it appears that because $\boldsymbol{f}(s)$ decays so rapidly at $s = \pm 1/2$, any force contribution beyond this would be negligible. This may indicate that sufficient decay in the slender body radius toward the endpoints of the fiber ensures that the endpoints (beyond $|s| = 1/2$) are not contributing a significant amount to the total force and thus can be safely ignored.

To test the formulation (5.3.4) for a different choice of radius function $r(s)$, we next consider the drag force on a straight cylinder with uniform radius everywhere along its length except for hemispherical caps at the fiber endpoints. In particular, we take the cylinder to be the same length as the prolate spheroid (actual fiber endpoints at $s = \pm\sqrt{1/4+\epsilon^2}$) with a radius that decays smoothly to zero at the endpoint via a hemispherical cap of radius ϵ centered at $d_\epsilon = \sqrt{1/4+\epsilon^2} - \epsilon$:

$$\epsilon r(s) = \begin{cases} \epsilon, & -d_\epsilon \le s \le d_\epsilon \\ \sqrt{\epsilon^2 - (s+d_\epsilon)^2}, & s < -d_\epsilon \\ \sqrt{\epsilon^2 - (s-d_\epsilon)^2}, & s > d_\epsilon \end{cases} \tag{4.3.1}$$

$$d_\epsilon := \sqrt{1/4+\epsilon^2} - \epsilon.$$

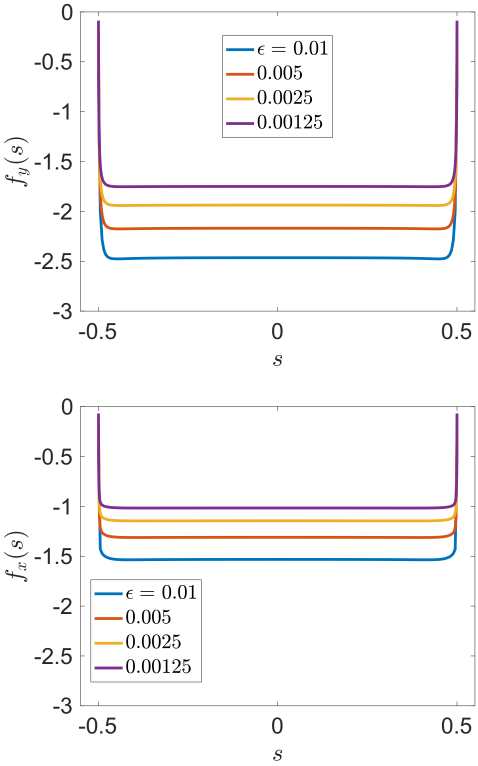


Figure 4.3.1: Force per unit length $\boldsymbol{f}(s)$, $s \in [-1/2, 1/2]$, along the prolate spheroid with semi-major axis aligned with the x-axis. The left figure shows the y-component of the force density for the cylinder translating with unit speed in the y-direction, while the right figure shows the x-component of the force density for the cylinder translating in the x-direction. Note that in both flows the force density $\boldsymbol{f}(s)$ decays to near zero at $s = \pm 1/2$, as expected.

As in the case of the prolate spheroid, we distribute the force density $\boldsymbol{f}(s)$ along the interval $[-1/2, 1/2]$. Using (5.3.4) to find $\boldsymbol{F}$ in the same way as in the case of the prolate spheroid, we compare the resulting drag force with the asymptotic expression derived by Keller and Rubinow [16] in table 4.3.2.

The computed drag force in table 4.3.2 agrees well with the asymptotic expression of Keller and Rubinow [16]; however, the computed force-per-unit-length $\boldsymbol{f}(s)$ is not as physically reasonable at the fiber endpoints. According to [22], in the case of a cylinder with hemispherical caps, we actually want a faster rate of decay in the force near the fiber endpoints – in particular, we need $\boldsymbol{f}(s)/(1/4 - s^2) \in C(-1/2, 1/2)$. However, as shown in figure 4.3.2, flow about the cylinder results in wild oscillations in $\boldsymbol{f}(s)$ near the fiber endpoints. Pos-

	$\boldsymbol{F}\cdot\boldsymbol{e}_y$ for $\mathbf{u}=\boldsymbol{e}_y$		$\boldsymbol{F}\cdot\boldsymbol{e}_x$ for $\mathbf{u}=\boldsymbol{e}_x$		
ϵ	Eqn (5.3.4)	Keller-Rubinow	Eqn (5.3.4)	Keller-Rubinow	$\epsilon\lvert\log\epsilon\rvert$
0.01	-2.6433	-2.6401	-1.6864	-1.6712	0.0461
0.005	-2.3085	-2.3024	-1.4216	-1.4094	0.0265
0.0025	-2.0472	-2.0417	-1.2274	-1.2189	0.0150
0.00125	-1.8384	-1.8342	-1.0796	-1.0738	0.0084

Table 4.3.2: Comparison of the computed (via expression (5.3.4)) and asymptotic (from Keller and Rubinow [16]) Stokes drag force $\boldsymbol{F}$ on a cylinder of length $2\sqrt{1/4+\epsilon^2}$ with hemispherical endpoints and with centerline along the x-axis. Columns 2 and 3 compare the y-component of $\boldsymbol{F}$ for a cylinder translating with unit speed in the y-direction, while columns 4 and 5 compare the x-component of $\boldsymbol{F}$ for translation in the x-direction. Here the expected $\epsilon\log\epsilon$ scaling of the difference between forces is less apparent, particularly in the y-direction. This may be due to endpoint effects (see figure 4.3.2).

sibly this indicates that this method (and likely others based on slender body theory) are really designed to treat prolate spheroids with sufficient decay in radius near the fiber endpoints.

4.3.3 Closed loops in constant flow

To validate the slender body approximation (5.3.4) in the closed loop setting ($\mathscr{I}=\mathbb{T}$), we compute the Stokes drag about a translating thin torus of length 1 with centerline in the xy-plane and axis of symmetry about the z-axis. We compare the computed drag force for various values of ϵ to the asymptotic expression of Johnson and Wu [15] (see table 4.3.3). Note that for the thin filaments that we consider here, the Johnson and Wu expression for the drag force corresponds well with the semianalytic expression for a torus translating in the z-direction, derived by Majumdar and O'Neill [18] with corrections by Amarakoon, et al. [1]. The Majumdar-O'Neill expression, consisting of an infinite sum of Legendre functions, holds for general values of s_0, where s_0 is defined to be the ratio of the outer radius of the torus (measured centerline to longitudinal axis) to the cross sectional radius. In [1], Amarakoon, et al. numerically verify the reported $\mathscr{O}(s_0^{-2})$ accuracy of the Johnson-Wu expression. In our case, we are mainly concerned with the parameter region $s_0=1/(2\pi\epsilon)>10$, so the Johnson-Wu expression agrees with the exact expression for Stokes drag in the z-direction to at least two digits.

Since the torus centerline $\boldsymbol{X}(s)$ is planar, we choose the normal vector $\cos(2\pi s)\boldsymbol{e}_x+\sin(2\pi s)\boldsymbol{e}_y$ to also lie in the xy-plane. The integral term in (5.3.4) is discretized

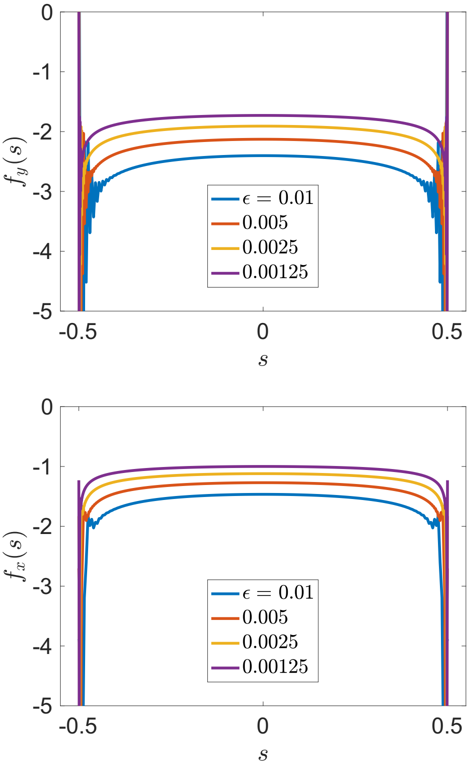


Figure 4.3.2: Force per unit length $\boldsymbol{f}(s)$, $s \in [-1/2, 1/2]$, along the uniform cylinder with hemispherical caps at the endpoints and centerline aligned with the x-axis. The left figure shows the y-component of the force density for the cylinder translating with unit speed in the y-direction, while the right figure shows the x-component of the force density for the cylinder translating in the x-direction. Comparing with figure 4.3.1, it is clear that the shape of the radius function $r(s)$ at the fiber endpoint has a large effect on $\boldsymbol{f}(s)$. In particular, despite the decay in $\boldsymbol{f}(s)$ at the very endpoint of the fiber, the oscillations leading up to the endpoint brings the physical validity of this force density into question.

using the trapezoidal rule, and the number of discretization points n along the fiber centerline is taken to be $n = 2/\epsilon$. Given zero background flow and uniform unit speed in the z-direction (columns 2 and 3, table 4.3.3) and y-direction (columns 4 and 5, table 4.3.3), the discretized operator (5.3.4) is inverted to find the force per unit length $\boldsymbol{f}(s)$, which is then summed over s to find the drag force $\boldsymbol{F}$. We plot the calculated $\boldsymbol{f}(s)$ in figure 4.3.3 to verify that the computed force density makes physical sense. For all computations, we take the viscosity $\mu = 1$.

Our method agrees quite well with the asymptotic expression of Johnson and

	$\boldsymbol{F}\cdot\boldsymbol{e}_z$ for $\mathbf{u}=\boldsymbol{e}_z$		$\boldsymbol{F}\cdot\boldsymbol{e}_y$ for $\mathbf{u}=\boldsymbol{e}_y$		
ϵ	Eqn (5.3.4)	Johnson-Wu	Eqn (5.3.4)	Johnson-Wu	$\epsilon\lvert\log\epsilon\rvert$
0.01	-2.4093	-2.3503	-1.8740	-1.8292	0.0461
0.005	-2.1076	-2.0806	-1.6309	-1.6103	0.0265
0.0025	-1.8788	-1.8664	-1.4484	-1.4389	0.0150
0.00125	-1.6979	-1.6922	-1.3051	-1.3007	0.0084

Table 4.3.3: We consider a translating slender torus of length 1 with centerline lying in the xy-plane, and compare the resulting Stokes drag force given by the slender body model (expression (5.3.4)) to the asymptotic expression calculated by Johnson and Wu [15]. Columns 2 and 3 compare the z-component of the drag force for a slender torus translating with speed 1 in the z-direction ("broadwise translation"), while columns 4 and 5 show the y-component of the drag for translation in the y-direction ("translation perpendicular to the longitudinal axis"). Here we can see an approximate $\epsilon\log\epsilon$ scaling in the difference between the two expressions.

Wu – as expected, table 4.3.3 shows roughly an $\mathcal{O}(\epsilon\log\epsilon)$ difference between the slender body approximation to the drag force and the asymptotic expression. This is encouraging since both (5.3.4) and the Johnson-Wu asymptotics are based on the Stokeslet/doublet expression (5.3.1). We have chosen these particular values of ϵ so that our method can also be compared with the regularized Stokeslet method of Cortez and Nicholas [7].

4.3.4 Free ended fibers in shear flow

In this section we calculate the angular momentum of a prolate spheroid with aspect ratio $\lambda = 1/\epsilon$ in the shear flow field $\mathbf{u}(z) = (z,0,0)^{\mathrm{T}}$. The torques are derived using both slender body theory (equation (5.3.4)) and the Jeffery model (equation (4.2.13)) for comparison. Figure 4.3.4a shows how the torque of the ellipsoid varies as a function of its orientation. Here, θ_2 is the second Euler angle and $\theta_2 = [-\pi/2,\pi/2]$ corresponds to a full revolution about the y-axis. We see that the torques agree at $\theta_2 = \pm\pi/2$ and the discrepancy between the two models increases as the orientation approaches alignment in the shear plane; in particular, the torque in the slender body model goes to zero but the Jeffery torque remains bounded away from zero. Since the fluid velocity is exactly zero along the particle centerline, the slender body model does not yield a torque on the particle. On the other hand, in the Jeffery model, the spheroid is aware of the non-zero fluid velocity gradient, and hence experiences a non-zero torque at this orientation.

Figure 4.3.4b shows the difference between the y-component of the torques due

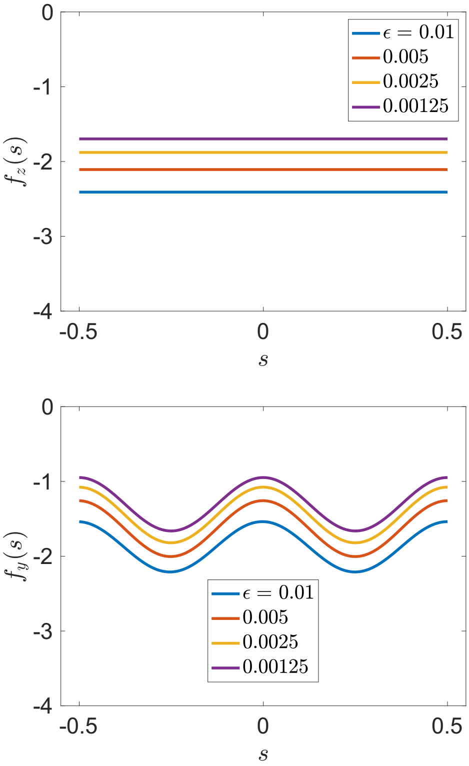
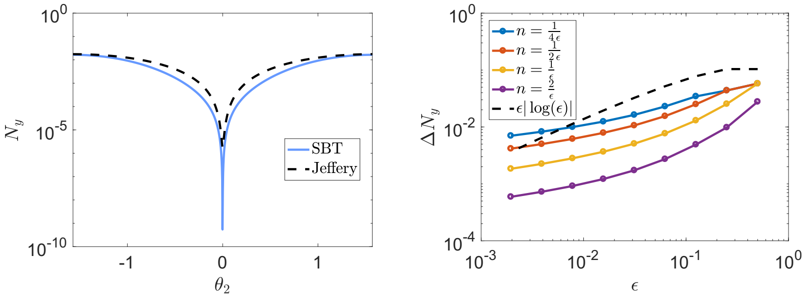


Figure 4.3.3: Force per unit length $\boldsymbol{f}(s)$, $s \in \mathbb{T}$, along the slender torus with centerline in the xy-plane. The left figure shows the z-component $\boldsymbol{f}(s) \cdot \boldsymbol{e}_z$ for a slender body translating with unit speed in the z-direction, while the right picture shows the y component $\boldsymbol{f}(s) \cdot \boldsymbol{e}_y$ for translation with unit speed in the y-direction.

to Jeffery and slender body theory as a function of ϵ for different values of n. The particles are oriented with $\theta_2 = \pi/2$, perpendicular to the shear plane. We see roughly $\mathcal{O}(\epsilon \log(\epsilon))$ convergence for the five largest values of ϵ. For smaller values of ϵ, the model converges at a slower rate. This is similar to the observed convergence in the force values (table 4.3.1), which are calculated for $\epsilon \leq 10^{-2}$. In addition, the two models show better agreement as the discretization parameter n is increased.

Figure 4.3.5 shows the the y-component from equation (5.5.1) of the torques due to slender body theory and Jeffery. The ODE for angular momentum is solved using one of MATLAB's built in functions such as `ode15s`. The particles are aligned as before with initial conditions $\boldsymbol{m}_0 = (0, 0.1, 0)^{\mathrm{T}}$ and Euler angles $(0, \pi/2, 0)^{\mathrm{T}}$; hence the only non-zero component of the angular momentum is m_y. We observe that for a relatively low aspect ratio (i.e., figure 4.3.5a)



(a) **(b)**

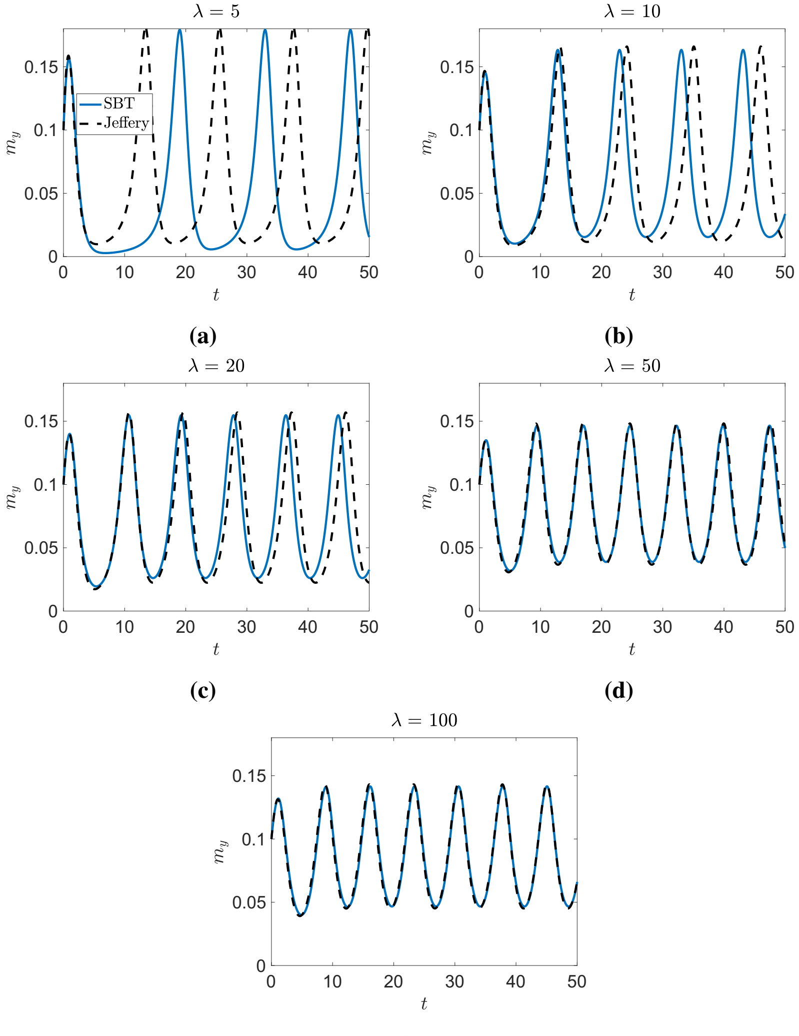
Figure 4.3.4: (a) The y-component of the torque for a prolate spheroid with $\lambda = 100$ for different orientations in shear flow. The values $\theta_2 = 0, \pm\pi/2$ correspond to alignment parallel and perpendicular to the shear plane, respectively. (b) The difference ΔN_y between the y-component of the torques due to Jeffery and slender body theory for a prolate spheroid of aspect ratio $\lambda = 1/\epsilon$ aligned in the z-direction in shear flow.

the models do not agree so well, however $\lambda = 5$ is not considered to be in the “slender" regime and we therefore do not expect good agreement here. As λ increases, the dynamics become almost indistinguishable.

We now turn our attention to figure 4.3.6, which displays how the choice of the discretization parameter n affects the solution quality. Figure 4.3.6a shows m_y for the slender body model for different numbers of discretization points n and figure 4.3.6b shows its 40 highest Fourier modes. The main observation here is that the model becomes more accurate as n increases. In particular, if n is chosen to be too low (here, too low corresponds to roughly less than $1/(2\epsilon)$) then the model does not resolve the low frequency modes, which can be seen by the spike at $k = 16$ in figure 4.3.6b, where only the $n = 50$ and 100 lines are able to reasonably capture this mode correctly.

4.3.5 Closed loops and oblate spheroids in shear flow

In this section we compare the rotational dynamics of a thin torus modeled by slender body theory to the rotational dynamics of an oblate disk of similar shape and mass. This comparison differs from the prolate spheroid comparisons in that here the particle shapes are different and we do not expect the two solutions to coincide. The slender torus experiences a force only along its centerline, whilst the oblate spheroid experiences a force all across its surface. In addition, the moment of inertia tensor for a torus of inner radius 2ϵ and of outer radius a



(a)

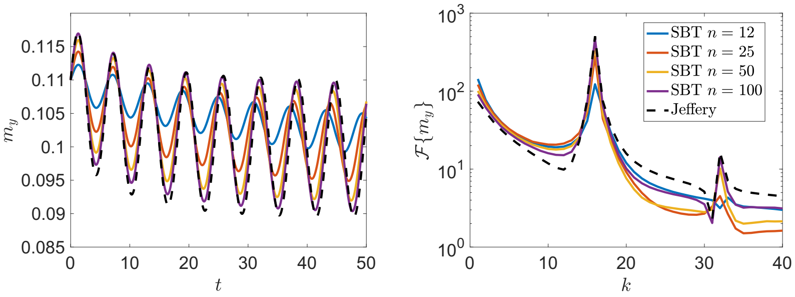
(b)

(c)

(d)

(e)

Figure 4.3.5: The y-component of the angular momentum of a particle in shear flow calculated from slender body theory (blue) and Jeffery (black, dashed). The aspect ratio takes different values in the range $\lambda \in [10, 100]$. The simulation parameters are $\mu = 0.06$, $n = \frac{2}{\epsilon}$



(a)

(b)

Figure 4.3.6: The y-component of the angular momentum of a particle in shear flow (a) and the first 40 Fourier modes (b). The colored lines are calculated from slender body theory with discretization parameter varying in the range $n \in [12, 100]$ and the dashed line is due to Jeffery. The simulation parameters are $\mu = 0.01$ and $\lambda = 50$ and $\boldsymbol{m}_0 = (0, 0.11, 0)^{\mathrm{T}}$.

(measured from the center of mass to the centerline) is given by

$$J_T = m_T \operatorname{diag}\left(\frac{4\,a^2 + 5\epsilon^2}{8}, \frac{4\,a^2 + 5\epsilon^2}{8}, \frac{4\,a^2 + 3\epsilon^2}{4}\right). \tag{4.3.2}$$

Setting the mass of the torus to $m_T = 2\,m_p/5$, where m_p is the mass of the spheroid, we have the relation $J - J_T = \mathcal{O}(\epsilon^2)$ for an oblate spheroid with semi minor axis length $b = \epsilon$. Due to the particle shape, the oblate spheroid experiences a much stronger torque; hence for the torques to be of the same magnitude, a viscosity of $\mu_T = 200\mu$ is chosen for the torus. The particles are placed at rest in the shear flow with the initial Euler angles $(0.01, 0.01, 0.01)$. We do this for two reasons: the first being that the Euler angles $(0, 0, 0)$ correspond to a neutrally stable orbit where the ellipsoid exhibits a tumbling motion forever. The second reason is that these angles correspond to exact alignment in the xy plane, where the slender model will not experience a force since the fluid velocity is exactly zero.

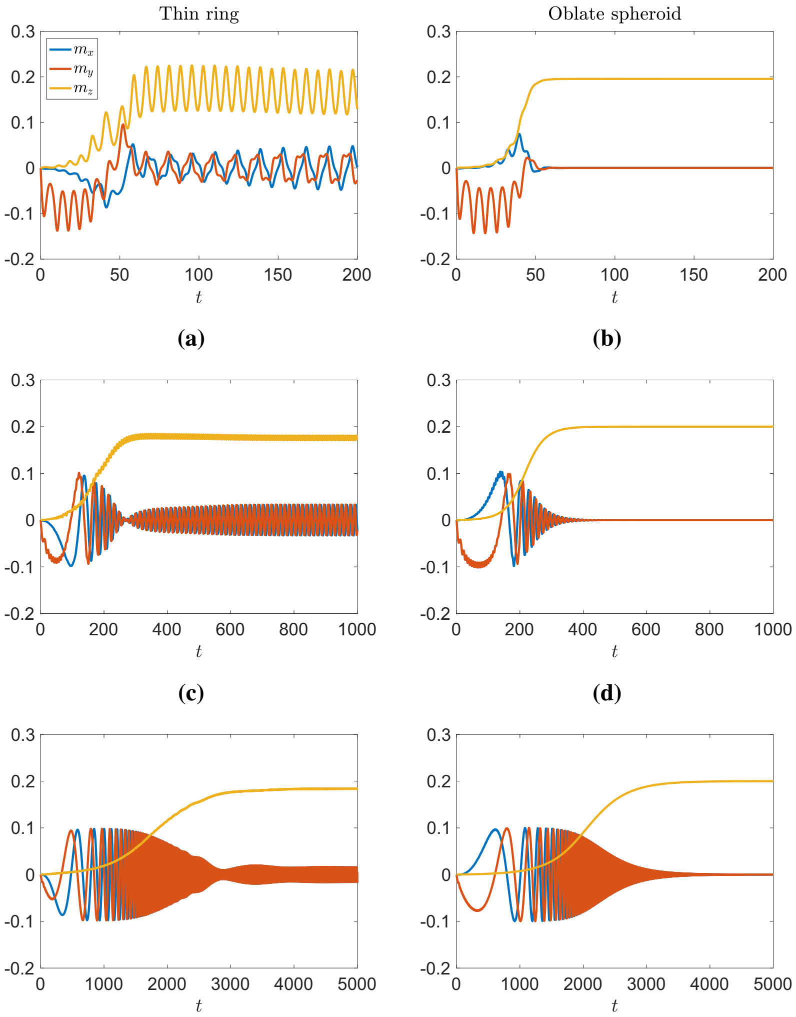
Challabotla et al. [4] conduct a similar experiment with oblate spheroids in shear flow and observe two phases of rotation: (1) an unstable wobbling phase of length proportional to the particle inertia, and (2) a stable rolling phase, where the spheroid aligns and rolls perfectly in the shear plane. Figure 4.3.7 shows $\boldsymbol{m}(t)$ for the thin ring with $\epsilon = 1/100$ and oblate spheroid with $\lambda = 1/100$ for three different values of μ (and the corresponding values of μ_T). For the spheroid model, we observe the temporary initial wobbling phase followed by the stable rolling phase where the particle rotates in the shear plane with a constant m_z component. In addition, as the relative particle inertia increases (that is, as the μ decreases), the wobbling phase is prolonged. These two observations are in agreement with the results in [4]. If we turn our attention to the thin ring, we observe some similarities: there is an initial wobbling phase followed by a somewhat different rolling phase. In the rolling phase, the particle's symmetry axis (the z-axis in the particle frame) precesses about the y-axis in the inertial frame. This is seen as oscillations in the m_x and m_y components about a mean zero value, which in turn affects the m_z component. A possible explanation for this precession is the fact that the slender ring does not experience a torque in the x or y directions (i.e., a restoring torque) when the axis of symmetry aligns perfectly with the y-axis in the inertial frame, since the gradient of the fluid velocity is not used in the calculation of the slender body torque. Hence the ring is susceptible to wobbling/precession at this orientation. This is in contrast with the spheroid, which experiences a non-zero torque in shear flow because of the positive fluid velocity gradient, regardless of the particle orientation. These discrepancies may not appear in more complex 3D flows and geometries.

4.4 Conclusion

In this paper we consider a model for thin, rigid fibers in viscous flows based on slender body theory. We investigate using the slender body approximation for the fluid field away from the fiber centerline as an approximation for the motion of the fiber itself by evaluating the expression on a curve along the slender body surface. Numerically, this yields a matrix equation for the force density along the length of the fiber that appears to be suitable for inversion even for very fine discretization of the fiber centerline.

For simple geometries and simple flows, we compare the slender body model to exact or asymptotically accurate expressions for the total force and torque acting on the particle. For the thin prolate spheroid, we compare the Stokes drag force predicted by slender body theory to the exact expression of Chwang and Wu [5]; for the cylinder, we compare with the asymptotic expression of Keller and Rubinow [16]; and for the thin torus, we compare with the asymptotic force expression of Johnson and Wu [15]. In the case of the prolate spheroid and the thin torus, we find essentially $\mathcal{O}(\epsilon \log \epsilon)$ agreement between our model and the exact or asymptotically accurate force values (tables 4.3.1 and 4.3.3), which is the accuracy predicted by rigorous error analyses [21, 22].

We also compared the torques on a thin prolate spheroid in shear flow for which the exact torques are given by Jeffery [13]. In the case of a thin torus, we qual-



(a) **(b)**

(c) **(d)**

(e) **(f)**

Figure 4.3.7: The angular momentum components of a thin ring (left column) and an oblate spheroid (right column) for $\mu_0 = 0.01$, 0.001 and 0.0001 (from top to bottom). The particle parameters are $\epsilon = \frac{1}{100}$, $\lambda = \frac{1}{100}$, $\mu_T = 200\mu$, $m_T = \frac{2}{5}m$, $m = 1$, $a = 1$, $n = \frac{1}{2\epsilon}$

itatively compared the dynamics of the torus with the Jeffery torques on an oblate spheroid of similar size. For the prolate spheroid, we found good agreement between our model and the Jeffery model, especially as the aspect ratio of the particle increases. In particular, in the slender body model, the dynamics appear to be better resolved for finer discretization of the filament (large n). For the thin torus, we observe somewhat similar results to those of Challabotla [4] for oblate spheroids; namely, we observe an initial “wobbling" phase followed by a steady “rolling" phase. The main difference is that in the rolling phase, the thin torus precesses about the directions perpendicular to the shear plane, while the spheroid maintains a constant angular momentum. This may be due to the fact that the slender model does not explicitly experience torque through the gradient, but only the values of the fluid velocity at the location of the centerline.

In the future, we aim to use this model to simulate elongated particles to determine the length scale at which the Jeffery model for prolate spheroids begins to lose validity in turbulent flows. We also aim to study the aggregation properties of many slender particles with more complicated shapes in turbulence (for example, helices or arbitrary closed loops). On the theoretical side, we would also like to obtain a more complete characterization of solvability conditions for the centerline equation. This would involve a spectral analysis of the equation (5.3.4) as well as the slender body PDE of [21, 22].

Bibliography

[1] A. Amarakoon, R. Hussey, B. J. Good, and E. G. Grimsal, *Drag measurements for axisymmetric motion of a torus at low Reynolds number*, Phys. Fluids, 25 (1982), pp. 1495–1501.

[2] G. Batchelor, *Slender-body theory for particles of arbitrary cross-section in Stokes flow*, J. Fluid Mech., 44 (1970), pp. 419–440.

[3] E. L. Bouzarth and M. L. Minion, *Modeling slender bodies with the method of regularized Stokeslets*, J. Comput. Phys., 230 (2011), pp. 3929–3947.

[4] N. R. Challabotla, C. Nilsen, and H. I. Andersson, *On rotational dynamics of inertial disks in creeping shear flow*, Phys. Lett. A, 379 (2015), pp. 157–162.

[5] A. T. Chwang and T. Y.-T. Wu, *Hydromechanics of low-Reynolds-number flow. Part 2: Singularity method for Stokes flows*, J. Fluid Mech., 67 (1975), pp. 787–815.

[6] R. Cortez, L. Fauci, and A. Medovikov, *The method of regularized Stokeslets in three dimensions: analysis, validation, and application to helical swimming*, Phys. Fluids, 17 (2005), p. 031504.

[7] R. Cortez and M. Nicholas, *Slender body theory for Stokes flows with regularized forces*, Commun. Appl. Math. Comput. Sci., 7 (2012), pp. 33–62.

[8] R. Cox, *The motion of long slender bodies in a viscous fluid part 1. general theory*, J. Fluid Mech., 44 (1970), pp. 791–810.

[9] I. Gallily and A.-H. Cohen, *On the orderly nature of the motion of nonspherical aerosol particles. ii. inertial collision between a spherical large droplet and an axially symmetrical elongated particle*, J. Colloid Interface Sci., 68 (1979), pp. 338–356.

[10] H. Goldstein, C. Poole, and J. Safko, *Classical mechanics*, 2002.

[11] T. GÖTZ, *Interactions of fibers and flow: asymptotics, theory and numerics*, Doctoral dissertation, University of Kaiserslautern, 2000.

[12] G. HANCOCK, *The self-propulsion of microscopic organisms through liquids*, Proc. R. Soc. Lond. A, 217 (1953), pp. 96–121.

[13] G. B. JEFFERY, *The motion of ellipsoidal particles immersed in a viscous fluid*, Proc. R. Soc. Lond. A, 102 (1922), pp. 161–179.

[14] R. E. JOHNSON, *An improved slender-body theory for Stokes flow*, J. Fluid Mech., 99 (1980), pp. 411–431.

[15] R. E. JOHNSON AND T. Y. WU, *Hydromechanics of low-Reynolds-number flow. Part 5: Motion of a slender torus*, J. Fluid Mech., 95 (1979), pp. 263–277.

[16] J. B. KELLER AND S. I. RUBINOW, *Slender-body theory for slow viscous flow*, J. Fluid Mech., 75 (1976), pp. 705–714.

[17] J. LIGHTHILL, *Flagellar hydrodynamics*, SIAM review, 18 (1976), pp. 161–230.

[18] S. MAJUMDAR AND M. O'NEILL, *On axisymmetric Stokes flow past a torus*, Z. Angew. Math. Phys., 28 (1977), pp. 541–550.

[19] W. MAO AND A. ALEXEEV, *Motion of spheroid particles in shear flow with inertia*, J. Fluid Mech., 749 (2014), pp. 145–166.

[20] J. MARTIN, A. LUSHER, R. C. THOMPSON, AND A. MORLEY, *The deposition and accumulation of microplastics in marine sediments and bottom water from the irish continental shelf*, Sci. Rep, 7 (2017), p. 10772.

[21] Y. MORI, L. OHM, AND D. SPIRN, *Theoretical justification and error analysis for slender body theory*, Comm. Pure Appl. Math, to appear, (2018).

[22] Y. MORI, L. OHM, AND D. SPIRN, *Theoretical justification and error analysis for slender body theory with free ends*, arXiv preprint arXiv:1901.11456, (2019).

[23] P. MORTENSEN, H. ANDERSSON, J. GILLISSEN, AND B. BOERSMA, *Dynamics of prolate ellipsoidal particles in a turbulent channel flow*, Phys. Fluids, 20 (2008), p. 093302.

[24] M. J. SHELLEY AND T. UEDA, *The Stokesian hydrodynamics of flexing, stretching filaments*, Phys. D, 146 (2000), pp. 221–245.

[25] B. Tapley, E. Celledoni, B. Owren, and H. I. Andersson, *A novel approach to rigid spheroid models in viscous flows using operator splitting methods*, Numer. Algorithms, (2019), pp. 1–19.

[26] B. K. Tapley, *Computing cost-effective particle trajectories in numerically calculated incompressible fluids using geometric methods*, arXiv preprint arXiv:1901.05236, (2019).

[27] A.-K. Tornberg and M. J. Shelley, *Simulating the dynamics and interactions of flexible fibers in Stokes flows*, J. Comput. Phys., 196 (2004), pp. 8–40.

[28] G. A. Voth and A. Soldati, *Anisotropic particles in turbulence*, Annu. Rev. Fluid Mech., 49 (2017), pp. 249–276.

[29] H. Zhang, G. Ahmadi, F.-G. Fan, and J. B. McLaughlin, *Ellipsoidal particles transport and deposition in turbulent channel flows*, Int. J. Multiph. Flow, 27 (2001), pp. 971–1009.

[30] L. Zhao, N. R. Challabotla, H. I. Andersson, and E. A. Variano, *Rotation of nonspherical particles in turbulent channel flow*, Phys. Rev. Lett., 115 (2015), p. 244501.

Appendix

4.A Other limiting slender body velocity expressions

Here we provide a brief overview of other methods used to obtain an expression for the motion of the fiber centerline $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$.

One such method is that of Lighthill [17] in which, away from $s = s'$, we simply plug $\mathbf{x} = \boldsymbol{X}(s)$ into the integral expression (5.3.1) (note that the doublet has negligible effect away from $s = s'$). Near $s = s'$, under the assumption that the centerline is essentially straight and the force density is approximately constant within this small region, the expression (5.3.1) can be evaluated exactly to obtain

$$8\pi\mu\big(\mathbf{u}^{\mathrm{L}}(s,t) - \mathbf{u}_0(\boldsymbol{X}(s,t),t)\big) = 2(\mathbf{I} - \boldsymbol{e}_s\boldsymbol{e}_s^{\mathrm{T}})\boldsymbol{f}(s,t) + \int_{|\boldsymbol{R}_0|>\delta}\left(\frac{\mathbf{I}}{|\boldsymbol{R}_0|} + \frac{\boldsymbol{R}_0\boldsymbol{R}_0^{\mathrm{T}}}{|\boldsymbol{R}_0|^3}\right)\boldsymbol{f}(s',t)\,ds';$$
$$\boldsymbol{R}_0(s,s',t) = \boldsymbol{X}(s,t) - \boldsymbol{X}(s',t), \quad \delta = \epsilon r(s)\sqrt{e}/2. \tag{4.A.1}$$

Here $\mathbf{u}^{\mathrm{L}}(s,t)$ approximates $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$, the actual motion of the fiber centerline, and $\mathbf{u}_0(\boldsymbol{X}(s,t),t)$ is the fluid flow at the spatial point $\mathbf{x} = \boldsymbol{X}(s,t)$ in the absence of the fiber.

Another popular method is that of Keller and Rubinow [16] in which the expression (5.3.1) is evaluated on the *actual* slender body surface (i.e. at a distance $\epsilon r(s)$ from $\boldsymbol{X}(s,t)$) and the method of matched asymptotics is used to obtain an expression for $\epsilon = 0$. In the far field (away from $s = s'$), (5.3.1) is simply Taylor expanded about $\epsilon = 0$. In the near field (near $s = s'$), the expression (5.3.1) is rewritten in terms of the rescaled variable $\xi = (s - s')/\epsilon$ and then expanded about $\epsilon = 0$. The far- and near-field expressions are then matched to create a centerline velocity expression that includes a local operator and a singular finite-part non-local operator:

$$8\pi\mu\big(\mathbf{u}^{\mathrm{KR}}(s,t) - \mathbf{u}_0(\boldsymbol{X}(s,t),t)\big) = -\boldsymbol{\Lambda}[\boldsymbol{f}](s,t) - \boldsymbol{K}[\boldsymbol{f}](s,t). \tag{4.A.2}$$

In the free end setting, the operators $\mathbf{\Lambda}$ and $\boldsymbol{K}$ are given by

$$\begin{aligned}\mathbf{\Lambda}[\boldsymbol{f}](s,t) &:= \left[(\mathbf{I}-3\boldsymbol{e}_{\rm s}\boldsymbol{e}_{\rm s}^{\rm T}) + (\mathbf{I}+\boldsymbol{e}_{\rm s}\boldsymbol{e}_{\rm s}^{\rm T})L(s)\right]\boldsymbol{f}(s,t)\\ \boldsymbol{K}[\boldsymbol{f}](s,t) &:= \int_{-1/2}^{1/2}\left[\left(\frac{\mathbf{I}}{|\boldsymbol{R}_0|}+\frac{\boldsymbol{R}_0\boldsymbol{R}_0^{\rm T}}{|\boldsymbol{R}_0|^3}\right)\boldsymbol{f}(s',t) - \frac{\mathbf{I}+\boldsymbol{e}_{\rm s}(s)\boldsymbol{e}_{\rm s}(s)^{\rm T}}{|s-s'|}\boldsymbol{f}(s,t)\right]ds',\end{aligned} \tag{4.A.3}$$

where $L(s) = \log\left(\frac{2(1/4-s^2)+2\sqrt{(1/4-s^2)^2+4\epsilon^2r^2(s)}}{\epsilon^2r^2(s)}\right)$. Note that we define L in this way to avoid singularities at the fiber endpoints; thus this L differs slightly from the expression given by [11] or the expression in [27].
In the closed loop setting, $\mathbf{\Lambda}$ and $\boldsymbol{K}$ are given by

$$\begin{aligned}\mathbf{\Lambda}[\boldsymbol{f}](s,t) &:= \left[(\mathbf{I}-3\boldsymbol{e}_{\rm s}\boldsymbol{e}_{\rm s}^{\rm T}) - 2(\mathbf{I}+\boldsymbol{e}_{\rm s}\boldsymbol{e}_{\rm s}^{\rm T})\log(\pi\epsilon/4)\right]\boldsymbol{f}(s,t)\\ \boldsymbol{K}[\boldsymbol{f}](s,t) &:= \int_{\mathbb{T}}\left[\left(\frac{\mathbf{I}}{|\boldsymbol{R}_0|}+\frac{\boldsymbol{R}_0\boldsymbol{R}_0^{\rm T}}{|\boldsymbol{R}_0|^3}\right)\boldsymbol{f}(s',t) - \frac{\mathbf{I}+\boldsymbol{e}_{\rm s}(s)\boldsymbol{e}_{\rm s}(s)^{\rm T}}{|\sin(\pi(s-s'))/\pi|}\boldsymbol{f}(s,t)\right]ds'.\end{aligned} \tag{4.A.4}$$

However, a spectral analysis of the Keller-Rubinow operator $-(\mathbf{\Lambda}+\boldsymbol{K})$ in the case of simple fiber geometries (see Götz [11] for the straight centerline and Shelley and Ueda [24] for the circular centerline) shows that the Keller-Rubinow expression is not suitable for inversion. In particular, the operator $-(\mathbf{\Lambda}+\boldsymbol{K})$ has a vanishing or nearly vanishing eigenvalue at some wavenumber $k \sim 1/\epsilon$. This high wavenumber instability limits the level to which the fiber can be discretized for numerics. It seems likely that more complicated centerline geometries also lead to a similar conclusion. Therefore in order to use the Keller-Rubinow expression for numerical simulations, the kernel of the operator $\boldsymbol{K}$ must be regularized. For example, in [24, 27], the denominators in the kernel of $\boldsymbol{K}$ are replaced by $\sqrt{|\boldsymbol{R}_0|^2+\delta^2}$ and $\sqrt{\sin^2(\pi(s-s'))/\pi^2+\delta^2}$, where $\delta=\delta(\epsilon)$ is chosen according to the fiber radius to maintain the same asymptotic accuracy as the Keller-Rubinow expression. This regularization, however, lacks a physical justification and clear connection to the expression (5.3.1).

Another common technique for describing the motion of the fiber centerline is to instead use the method of regularized Stokeslets (see [3, 6, 7]) to obtain an alternate version of (5.3.1). In this method, the Stokeslet is approximated by the (smooth) solution to

$$-\mu\Delta\mathbf{u}+\nabla p = \boldsymbol{f}\phi_\delta(\mathbf{x}), \quad \operatorname{div}\mathbf{u}=0$$

where ϕ_δ is a smooth, radially symmetric function with $\int_{\mathbb{R}^3}\phi_\delta = 1$. The parameter δ determines the spread of ϕ_δ and, in the case of slender body theory, is usually chosen such that $\delta \sim \epsilon$. The slender body approximation is then constructed as in (5.3.1), but now the resulting expression is not singular at $\mathbf{x} = \boldsymbol{X}(s)$, and the velocity of the slender body itself may be approximated by simply evaluating the regularized expression along the fiber centerline. The method of regularized Stokeslets can be used to construct regularized versions of the Lighthill and Keller-Rubinow expressions [7]. However, from the outset, the method of regularized Stokeslets approximates a slightly different problem from (5.3.1), and it is not entirely clear that these solutions should be close for any δ. The choice of regularization parameter δ greatly affects the resulting dynamics; however, a systematic justification for this parameter choice is lacking.

An integral model based on slender body theory, with applications to curved rigid fibers

Helge I Andersson, Elena Celledoni, Laurel Ohm, Brynjulf Owren and Benjamin K Tapley

Physics of Fluids 33 (4), 041904

An integral model based on slender body theory, with applications to curved rigid fibers

Abstract. We propose a novel integral model describing the motion of both flexible and rigid slender fibers in viscous flow, and develop a numerical method for simulating dynamics of curved rigid fibers. The model is derived from nonlocal slender body theory (SBT), which approximates flow near the fiber using singular solutions of the Stokes equations integrated along the fiber centerline. In contrast to other models based on (singular) SBT, our model yields a smooth integral kernel which incorporates the (possibly varying) fiber radius naturally. The integral operator is provably negative definite in a non-physical idealized geometry, as expected from partial differential equation (PDE) theory. This is numerically verified in physically relevant geometries. We propose a convergent numerical method for solving the integral equation and discuss its convergence and stability. The accuracy of the model and method is verified against known models for ellipsoids. Finally, a fast algorithm for computing dynamics of rigid fibers with complex geometries is developed.

5.1 Introduction

The dynamics of thin fibers immersed in fluid play an important role in many biological and engineering processes, including microorganism propulsion [8, 28, 43, 49], rheological properties of fiber suspensions used to create composite materials [16, 20, 41], and deposition of microplastics in the ocean [31]. Here the term ‘fiber’ is used to refer to a particle with a very large aspect ratio. In many of the applications mentioned, the cross sectional radius of the fiber is small compared to the length scales of the surrounding fluid, which can be well approximated locally by Stokes flow. This allows for the development of computationally tractable mathematical models describing the interaction between the fiber and the surrounding fluid.

Due to the linearity of the Stokes equations, the three dimensional flow about a body can be fully described by an expression over only the two dimensional surface of the body [42]; however, for flexible particles with complex shapes or for multiple interacting particles, this quickly becomes both analytically and computationally prohibitive. In the case of slender fibers, a more tractable option is to exploit the thinness of the fiber by approximating it as a one dimensional curve. This idea forms the basis for slender body theory (SBT). Models based on slender body theory in general are popular because they yield simple, efficient expressions for the velocity of filaments in fluid, allowing for the simulation of many interacting fibers with complex, semiflexible shapes. The most basic form of SBT (placing singular point forces known as Stokeslets along the fiber centerline) dates back to works by Hancock [21], Cox [14], and Batchelor [4]. Later developments in singular SBT, due to Keller and Rubinow [25], Lighthill [29], and Johnson [24], involved adding higher order corrections to the point force to account for the finite radius of the fiber. The most natural choice of higher order correction is often referred to as the doublet (see discussion following equation (5.3.1)). We will refer to these methods based on distributing Stokeslets and doublets along the fiber as *classical* nonlocal SBT to distinguish from some more recent developments.

Classical SBT gives rise to an expression which exactly satisfies the unforced Stokes equations away from the fiber, and, to leading order (with respect to the fiber radius) satisfies the boundary conditions for a well-posed boundary value problem for the Stokes equations [35, 36]. This expression has served as the basis for various numerical methods [45, 54, 55]. However, one issue with classical SBT is that the velocity expression is singular along the fiber centerline, and the usual methods for obtaining an expression for the velocity of the fiber itself – involving a nonstandard finite part integral – give rise to high wavenumber instabilities [18, 45, 55]. To address this, Tornberg–Shelley [55] regularize the integral kernel using an additional parameter proportional to the fiber radius.

To more generally avoid some of the difficulties of integrating a singular kernel, Cortez [10, 12, 13] developed the method of regularized Stokeslets. Here, instead of placing singular solutions of the Stokes equations along the fiber centerline, *regularized Stokeslets* are used. Regularized Stokeslets satisfy the Stokes equations with forcing given by a smooth approximation to the identity – or blob function – whose width is controlled by a parameter which can be chosen to be proportional to the fiber radius. Unlike classical SBT, this results in an expression for the fluid velocity that is nonsingular along the actual centerline of the fiber, allowing for a simpler representation of the fiber velocity. Many recent computational models for thin fibers rely on the method of regularized Stokeslets [5, 11, 48, 58, 59]. However, many choices of blob function are possible and there is not a canonical procedure for choosing one. Additionally, many commonly used blob functions introduce an additional nonzero body force into the fluid away from the fiber surface [61].

Most recently, Maxian et al. [32] developed a fiber model that is asymptotically equivalent to SBT but based on the Rotne-Prager-Yamakawa (RPY) tensor [44, 60] commonly used to model hydrodynamically interacting spheres. The model also places a curve of (singular) Stokeslets plus doublets along the fiber centerline, but replaces the region around the singular part of the Stokeslet/doublet kernel with the RPY regularization. The RPY kernel is divergence-free and known to be positive definite, making it a good choice for modeling particles in close proximity. The discontinuous kernel, however, makes the model more difficult to compare to the PDE solution of (Refs. [35, 36]), which is one of the main goals of the model presented here.

We aim to make use of the fact that classical SBT closely approximates the solution to a well-posed boundary value problem [35, 36] for the fluid velocity outside of the fiber, although the conventional way to obtain an expression for the velocity of the filament itself gives rise to instabilities which must later be corrected. Regularized Stokeslets yield a simpler expression for the fiber velocity, but can introduce errors outside of the filament and give rise to a fiber velocity which may fundamentally differ from the aforementioned PDE solution (see Remark 5.1). Thus we consider a different approach to deriving a fiber velocity expression from classical SBT. Beginning with the fundamental premise of classical SBT – placing singular Stokeslets along the fiber centerline along with doublets to cancel the angular dependence across each fiber cross section – we aim to devise a model which is analytically and computationally attractive (in that it does not exhibit high wavenumber instabilities) with a physically meaningful derivation.

Our integral model is based on classical SBT but involves a smooth kernel which incorporates the (possibly varying) fiber radius in a natural way. Since the integral kernels are smooth, the model resembles the method of regularized Stokeslets with an arclength-dependent regularization similar to (Ref. [58]); however, we derive our model from usual (singular) Stokeslets and doublets. As such, we avoid introducing a nonzero body force throughout the fluid outside of the fiber [61], and avoid introducing additional parameters into the basic first-kind formulation of the model. The model relies on the asymptotic cancellation of angular-dependent terms along the fiber surface (see Section 5.3 for details), leaving an expression that retains a dependence on the fiber radius in a natural way.

Furthermore, we include a systematic way of comparing mapping properties among different fiber models based on (Ref. [34]), which involves calculating the spectra of the integral operators from various models in the toy scenario of a straight-but-periodic fiber with constant radius. In this model geometry, our integral operator is negative definite, as is the well-posed partial differential equation (PDE) operator of (Refs. [35, 36]) which it is designed to approximate (see Ref. [34]). This is in contrast to other models based on (non-regularized) slender body theory which give expressions for the fiber velocity involving further asymptotic expansion with respect to the fiber radius [24, 25, 29]. These models exhibit an instability as the eigenvalues of the operator cross zero at a high but finite wavenumber.

The model we derive initially yields a first-kind Fredholm integral equation for the force density along the fiber centerline. Such integral equations are known for being ill-posed (see Ref. [26, Chapt. 15.1]), as they do not necessarily have a bounded inverse at the continuous level. Numerical discretization alone can provide sufficient regularization to invert first-kind integral equations at the discrete level, but to make our model more suitable for inversion, we use an integral identity to regularize the expression into a second-kind equation. The second-kind regularization preserves the asymptotic accuracy of the model while improving the conditioning and invertibility of the corresponding numerical method. The regularization also serves to ensure that the discretized operator is negative definite, even in the presence of numerical errors, by bounding the spectrum away from zero. We distinguish this type of regularization from the method of regularized Stokeslets, since our regularization is not a key component of the model derivation. In particular, we can directly compare our model *with* regularization to our model *without*, which we will do repeatedly throughout the paper. We also distinguish this regularization from the procedure used by Tornberg and Shelley [55], since we are not correcting for a high wavenumber instability. This allows us to compare the numerical behavior of our regularized and unregularized models at the discrete level even for very fine discretization. Moreover, the regularization used here affects all directions (both normal and tangent to the slender body centerline) in the same way.

The solution of the resulting second-kind Fredholm integral equation is a force density along the slender body centerline which we integrate to find the total force and torque on the rigid fiber. We implement a numerical method based on the Nyström method for solving second-kind Fredholm integral equations (see Ref. [2, Chapt. 12.4]). Numerical tests confirm its convergence. Not surprisingly, we note significant improvements in the conditioning of the second-kind versus first-kind formulation of the model. We also numerically verify the spectral properties of the model in different geometries.

The model applies to both semiflexible and rigid fibers; however, the invertibility properties of the second kind model make it particularly well suited for simulating rigid filaments. We present an algorithm for dynamic simulations of a rigid fiber using Gauss-Legendre quadrature. The rigidity of the fiber can be exploited such that only matrix-vector products need to be performed within the time loop. We compare the dynamics of our model to the well-studied dynamics of a slender prolate spheroid [6, 9, 23]. We then apply our model to compare the dynamics of curved fibers whose centerlines deviate randomly from straight lines by varying magnitudes.

The structure of the paper is as follows. Section 5.2 presents the slender body model, which is derived in greater detail and justified via spectral comparisons with other slender body theories in Section 5.3. In Section 5.4 we discuss a method for numerically solving Fredholm integral equations and integrating the result, and demonstrate the convergence of the method for our model. Section

5.5 outlines a fast algorithm for computing the dynamics of a rigid slender fiber in viscous flow. We apply the dynamical algorithm to simulate the dynamics of fibers with complex shapes. Finally, we comment on conclusions and outlook for the model in Section 5.6.

5.1.1 Fiber geometry

We begin by introducing some notation for the slender geometries considered throughout the paper. Fix ϵ, L with $0<\epsilon \ll L$ and let $\boldsymbol{X}_{\rm ext}:[-\sqrt{L^2+\epsilon^2},\sqrt{L^2+\epsilon^2}]\to\mathbb{R}^3$ denote the coordinates of a C^2 curve in $\mathbb{R}^3$, parameterized by arclength s. Defining $\boldsymbol{e}_{\rm s}(s)=\frac{d\boldsymbol{X}_{\rm ext}}{ds}/\left|\frac{d\boldsymbol{X}_{\rm ext}}{ds}\right|$, the unit tangent vector to $\boldsymbol{X}_{\rm ext}(s)$, we parameterize points near $\boldsymbol{X}_{\rm ext}(s)$ with respect to the orthonormal frame $(\boldsymbol{e}_{\rm s}(s),\boldsymbol{e}_{n_1}(s),\boldsymbol{e}_{n_2}(s))$ defined in (Ref. [36]). Letting

$$\boldsymbol{e}_r(s,\theta):=\cos\theta\,\boldsymbol{e}_{n_1}(s)+\sin\theta\,\boldsymbol{e}_{n_2}(s),$$

we define the slender body Σ_ϵ as

$$\Sigma_\epsilon:=\left\{\mathbf{x}\in\mathbb{R}^3\,:\,\mathbf{x}=\boldsymbol{X}_{\rm ext}(s)+\rho\boldsymbol{e}_r(s,\theta),\ \rho<\epsilon r(s),\ s\in[-\sqrt{L^2+\epsilon^2},\sqrt{L^2+\epsilon^2}]\right\}. \tag{5.1.1}$$

Here the radius function $r\in C^2(-\sqrt{L^2+\epsilon^2},\sqrt{L^2+\epsilon^2})$ is required to satisfy $0<r(s)\le 1$ for each $s\in(-\sqrt{L^2+\epsilon^2},\sqrt{L^2+\epsilon^2})$, and $r(s)$ must decay smoothly to zero at the fiber endpoints $\pm\sqrt{L^2+\epsilon^2}$. There are many admissible radius functions r which can be considered. For the simulations in this paper, we will use a thin prolate spheroid as our geometrical model for a slender fiber. In this case, the radius function $r(s)$ is given by

$$r(s)=\frac{1}{\sqrt{L^2+\epsilon^2}}\sqrt{L^2+\epsilon^2-s^2}. \tag{5.1.2}$$

We consider the subset

$$\boldsymbol{X}:=\{\boldsymbol{X}_{\rm ext}(s)\,:\,-L\le s\le L\} \tag{5.1.3}$$

extending from focus to focus of the prolate spheroid (5.1.2), and define $\boldsymbol{X}(s)$ to be the effective centerline of the slender body so that $r=O(\epsilon)$ at the effective endpoints $s=\pm L$.

The slender body model described in Section 5.2 may also be used in the case of a closed curve, in which case we take $\boldsymbol{X}(L)=\boldsymbol{X}(-L)$ and consider $s\in\mathbb{R}/2L$. We may take the radius function $r\equiv 1$ in this case.

5.2 Slender body model

To describe the motion of the thin fiber Σ_ϵ (5.1.1) in Stokes flow, we will use an expression derived from classical nonlocal slender body theory [18, 24, 55]. Letting $\boldsymbol{f}(s,t)$ denote the force per unit length exerted by the fiber on the surrounding fluid at time t, we approximate the velocity $\frac{\partial \boldsymbol{X}}{\partial t}$ of the fiber relative to a given background flow $\mathbf{u}_0$ by

$$8\pi\mu\left(\frac{\partial \boldsymbol{X}}{\partial t} - \mathbf{u}_0(\boldsymbol{X}(s,t),t)\right) = -2\log(\eta)\,\boldsymbol{f}(s,t) - \int_{-L}^{L}\left(\boldsymbol{S}_{\epsilon,\eta} + \frac{\epsilon^2 r^2(s')}{2}\boldsymbol{D}_\epsilon\right)\boldsymbol{f}(s',t)\,ds', \tag{5.2.1}$$

$$\boldsymbol{S}_{\epsilon,\eta}(s,s',t) = \frac{\mathbf{I}}{(|\overline{\boldsymbol{X}}|^2 + \eta^2\epsilon^2 r^2(s))^{1/2}} + \frac{\overline{\boldsymbol{X}\boldsymbol{X}}^{\mathrm{T}}}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{3/2}} \tag{5.2.2}$$

$$\boldsymbol{D}_\epsilon(s,s',t) = \frac{\mathbf{I}}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{3/2}} - \frac{3\overline{\boldsymbol{X}\boldsymbol{X}}^{\mathrm{T}}}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{5/2}} \tag{5.2.3}$$

where $\overline{\boldsymbol{X}}(s,s',t) = \boldsymbol{X}(s,t) - \boldsymbol{X}(s',t)$. Here $\eta \geq 1$ is a parameter which can be chosen to yield either a first kind ($\eta = 0$) or a second-kind ($\eta > 0$) Fredholm equation for $\boldsymbol{f}$. Notice that η must also appear in the first term of $\boldsymbol{S}_{\epsilon,\eta}$ in order to retain the asymptotic consistency of the model (5.2.1). This is due to an integral identity (5.3.6) used to convert the integral model from a first-kind equation for $\boldsymbol{f}$. The model accounts for a varying radius $r(s)$ through the denominators of each term as well as the coefficient of $\boldsymbol{D}_\epsilon$. Note that since $r(s)$ is nonzero for $-L \leq s \leq L$, the integral kernel is smooth for each $s \in [-L,L]$. We provide a more detailed derivation of (5.2.1)–(5.2.3) in Section 5.3.

The model given by equations (5.2.1)–(5.2.3) and the analysis in Section 5.3 can be used to describe both flexible and rigid fibers. In Section 5.5 we apply our model to the dynamics of a rigid fiber, since the invertibility properties of (5.2.1)–(5.2.3) make the model especially suitable for simulating rigid filaments.

In the case of a rigid fiber, at each time t we additionally impose the constraint

$$\frac{\partial \boldsymbol{X}}{\partial t} = \mathbf{v} + \boldsymbol{\omega} \times \boldsymbol{X}(s), \tag{5.2.4}$$

where $\mathbf{v}$, $\boldsymbol{\omega} \in \mathbb{R}^3$ are the linear and angular velocity of the fiber (see Refs. [19, 33, 54]). The total force $\boldsymbol{F}(t)$ and torque $\boldsymbol{T}(t)$ exerted on the slender body at time t are computed from the line force density $\boldsymbol{f}(s,t)$ via

$$\int_{-L}^{L}\boldsymbol{f}(s,t)\,ds = \boldsymbol{F}(t), \qquad \int_{-L}^{L}\boldsymbol{X}(s,t)\times\boldsymbol{f}(s,t) = \boldsymbol{T}(t). \tag{5.2.5}$$

When $\mathbf{v}$ and $\boldsymbol{\omega}$ are prescribed and one aims to solve for $\boldsymbol{F}$ and $\boldsymbol{T}$, this is known as the resistance problem. Conversely, the case when $\boldsymbol{F}$ and $\boldsymbol{T}$ are given and the rigid fiber velocity is sought is known as the mobility problem. Note that for both the resistance and mobility problems along a thin fiber, using (5.2.1) to relate fiber velocity to force involves inverting the integral equation to solve for the force density $\boldsymbol{f}$. Thus we are particularly concerned with the invertibility of (5.2.1). In Section 5.5, we use (5.2.1), (5.2.4), and (5.2.5) to solve the resistance problem, which is of interest when the density of the fiber is much larger than the density of the fluid.

5.3 Derivation and justification of the slender body model

Our model for the motion of the fiber is based on classical nonlocal slender body theory, where the fluid velocity $\mathbf{u}^{\mathrm{SB}}(\mathbf{x},t)$ at any point $\mathbf{x}$ away from the fiber centerline $\boldsymbol{X}(s,t)$ is approximated by the integral expression

$$8\pi\mu\big(\mathbf{u}^{\mathrm{SB}}(\mathbf{x},t)-\mathbf{u}_0(\mathbf{x},t)\big)=-\int_{-L}^{L}\left(\mathscr{S}\big(\mathbf{x}-\boldsymbol{X}(s',t)\big)+\frac{\epsilon^2r^2(s')}{2}\mathscr{D}\big(\mathbf{x}-\boldsymbol{X}(s',t)\big)\right)\boldsymbol{f}(s',t)\,ds'$$
$$\mathscr{S}(\mathbf{x})=\frac{\mathbf{I}}{|\mathbf{x}|}+\frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^3},\ \mathscr{D}(\mathbf{x})=\frac{\mathbf{I}}{|\mathbf{x}|^3}-\frac{3\mathbf{x}\mathbf{x}^{\mathrm{T}}}{|\mathbf{x}|^5}. \tag{5.3.1}$$

where $\mathbf{u}_0(\mathbf{x},t)$ is the fluid velocity in the absence of the fiber and μ is the fluid viscosity. The force-per-unit-length $\boldsymbol{f}(s,t)$ exerted by the fluid on the body is distributed between the generalized foci of the slender body at $s=\pm L$. The expression $\frac{1}{8\pi\mu}\mathscr{S}(\mathbf{x})$ is the free space Green's function for the Stokes equations in $\mathbb{R}^3$, commonly known as the Stokeslet, while $\frac{1}{8\pi\mu}\mathscr{D}(\mathbf{x})=\frac{1}{16\pi\mu}\Delta\mathscr{S}(\mathbf{x})$ is a higher order correction to the velocity approximation, often known as a doublet. The doublet coefficient $\frac{\epsilon^2r^2}{2}$ is chosen to cancel the leading order (in ϵ) angular dependence in the fluid velocity at the surface of the actual 3D filament. This coefficient can be obtained via matched asymptotics, or by the following heuristic. Since the purpose of the doublet is to cancel the angular dependence over each 2D cross section of the fiber, we consider Stokes flow in $\mathbb{R}^2$ due to a point force at the origin of strength $\boldsymbol{f}$. In polar coordinates $\mathbf{x}=(\rho\cos\theta,\rho\sin\theta)^{\mathrm{T}}$, the velocity due to the Stokeslet at $\rho>0$ is given by

$$\mathbf{u}^{\mathscr{S}}(\rho,\theta)=\frac{1}{4\pi}\left(-\log\rho\mathbf{I}+\frac{1}{2}\begin{pmatrix}1+\cos2\theta & \sin2\theta\\ \sin2\theta & 1-\cos2\theta\end{pmatrix}\right)\begin{pmatrix}f_1\\ f_2\end{pmatrix},$$

where $\mathbf{I}$ is the 2D identity matrix. To eliminate the θ-dependence on the circle $\rho=\epsilon$, we note that

$$\begin{aligned}\Delta\mathbf{u}^{\mathscr{S}}(\rho,\theta)&=\frac{\partial^2\mathbf{u}^{\mathscr{S}}}{\partial\rho^2}+\frac{1}{\rho}\frac{\partial\mathbf{u}^{\mathscr{S}}}{\partial\rho}+\frac{1}{\rho^2}\frac{\partial^2\mathbf{u}^{\mathscr{S}}}{\partial\theta^2}\\ &=-\frac{1}{2\pi\rho^2}\begin{pmatrix}\cos2\theta & \sin2\theta\\ \sin2\theta & -\cos2\theta\end{pmatrix}\begin{pmatrix}f_1\\ f_2\end{pmatrix}.\end{aligned}$$

Therefore the θ-dependence in the velocity due to the Stokeslet at $r = \epsilon$ can be canceled by adding a doublet term ($\frac{1}{2}\Delta\mathbf{u}^{\mathscr{S}}$) with coefficient $\frac{\epsilon^2}{2}$:

$$\mathbf{u}^{\text{SB}} = \mathbf{u}^{\mathscr{S}} + \frac{\epsilon^2}{4}\Delta\mathbf{u}^{\mathscr{S}}.$$

The expression (5.3.1) is valid for describing flows around fibers which are not highly curved (i.e. with maximum centerline curvature $\ll 1/\epsilon$) and do not come close to self-intersection ($\left|\boldsymbol{X}(s) - \boldsymbol{X}(s')\right| / \left|s - s'\right| \geq C$ for C independent of ϵ). The force density $\boldsymbol{f}$ must also be sufficiently regular. Given these constraints, in the stationary setting, the velocity field given by (5.3.1) is an asymptotically accurate approximation to the velocity field around a three-dimensional semi-flexible rod satisfying a well-posed *slender body PDE*, defined in (Refs. [35, 36]) as the following boundary value problem for the Stokes equations:

$$\begin{aligned} -\mu\Delta\mathbf{u} + \nabla p = 0, \quad \operatorname{div}\mathbf{u} = 0 \qquad & \text{in } \mathbb{R}^3\backslash\overline{\Sigma_\epsilon} \\ \int_0^{2\pi} (\boldsymbol{\sigma}\boldsymbol{n})\big|_{(\varphi(s),\theta)} \mathscr{J}_\epsilon(\varphi(s),\theta)\varphi'(s)\, d\theta = -\boldsymbol{f}(s) \qquad & \text{on } \partial\Sigma_\epsilon \\ \mathbf{u}\big|_{\partial\Sigma_\epsilon} = \mathbf{u}(s), \qquad & \text{unknown but independent of } \theta \\ |\mathbf{u}| \to 0 \text{ as } |\mathbf{x}| \to \infty. \qquad & \end{aligned} \tag{5.3.2}$$

Here $\boldsymbol{\sigma} = \mu\big(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\text{T}}\big) - p\mathbf{I}$ is the fluid stress tensor, $\boldsymbol{n}(\mathbf{x})$ denotes the unit normal vector pointing into Σ_ϵ at $\mathbf{x} \in \partial\Sigma_\epsilon$, $\mathscr{J}_\epsilon(s,\theta)$ is the Jacobian factor on $\partial\Sigma_\epsilon$, and $\varphi(s) := \frac{s\sqrt{L^2+\epsilon^2}}{L}$ is a stretch function to address the discrepancy between the extent of $\boldsymbol{f}$ and the extent of the actual slender body surface. Given a force density $\boldsymbol{f} \in C^1(-L,L)$ which decays like $r(s)$ at the fiber endpoints ($\boldsymbol{f}(s) \sim r(\varphi(s))$ as $s \to \pm L$), the difference between the slender body approximation $\mathbf{u}^{\text{SB}}$ and the solution of (5.3.2) is bounded by an expression proportional to $\epsilon\left|\log\epsilon\right|$. Note that $r(s)$ need not be spheroidal (5.1.2) for this error analysis to hold, but $r(s)$ must decay smoothly to zero at the physical endpoints of the fiber at $s = \pm\sqrt{L^2 + \epsilon^2}$.

A key component of the well-posedness theory for the slender body PDE to which (5.3.1) is an approximation is the*fiber integrity condition* on $\mathbf{u}\big|_{\partial\Sigma_\epsilon}$. The fiber integrity condition requires the velocity across each cross section s of the slender body to be constant; i.e. the velocity $\mathbf{u}(\mathbf{x})$ at any point $\mathbf{x}(s,\theta) = \boldsymbol{X}(s) + \epsilon r(s)\boldsymbol{e}_r(s,\theta) \in \partial\Sigma_\epsilon$ satisfies $\partial_\theta\mathbf{u}(\mathbf{x}(s,\theta)) = 0$. This is to ensure that the cross sectional shape of the fiber does not deform over time. An important aspect of the accuracy of slender body theory is that the expression (5.3.1) satisfies this fiber integrity condition to leading order in ϵ. Specifically, by Propositions 3.9 and 3.11 in (Refs. [35, 36]), respectively, we have that for

$\mathbf{x}(s,\theta) \in \partial\Sigma_\epsilon$,

$$\left|\partial_\theta \mathbf{u}^{\mathrm{SB}}(\mathbf{x}(s,\theta))\right| \le C\left(\epsilon|\log\epsilon|\|\boldsymbol{f}\|_{C^1(-L,L)} + \epsilon\left\|\frac{\boldsymbol{f}}{r}\right\|_{C^0(-L,L)}\right); \tag{5.3.3}$$

i.e. the angular dependence in $\mathbf{u}^{\mathrm{SB}}(\mathbf{x})$ over each cross section s of the slender body is only $\mathcal{O}(\epsilon\log\epsilon)$.

Another important general feature of the slender body PDE (5.3.2) is that the operator mapping the force data $\boldsymbol{f}(s)$ to the θ-independent fiber velocity $\mathbf{u}|_{\partial\Sigma_\epsilon}(s)$ is negative definite (see (Ref. [34]); note that the sign convention for $\boldsymbol{f}$ is opposite).

Now, the velocity expression (5.3.1) is singular at $\mathbf{x} = \boldsymbol{X}(s,t)$ and can be used only away from the fiber centerline; however, (5.3.1) presents a starting point for approximating the velocity of the slender body itself. Various methods can be used to obtain an expression for the relative velocity of the fiber centerline $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$ which depends only on the arclength parameter s and time t. The most common way to go from equation (5.3.1) to an expression independent of θ is to perform an asymptotic expansion about $\epsilon = 0$ [18, 24, 40, 55]. However, as alluded to in the introduction, this leads to issues at high frequency modes along the fiber (we will come back to this point later). Here we consider a different approach to deriving a limiting centerline expression from (5.3.1) which evidently results in a negative definite integral operator mapping $\boldsymbol{f}$ to $\mathbf{u}|_{\partial\Sigma_\epsilon}$. We then regularize this first-kind integral equation in an asymptotically consistent way to yield the second-kind integral equation (5.2.1). We detail our approach here and provide further justification in Section 5.3.1 using a model geometry.

The first step in approximating $\frac{\partial \boldsymbol{X}(s,t)}{\partial t}$ is to evaluate (5.3.1) on the surface of the slender body at $\mathbf{x} = \boldsymbol{X}(s,t) + \epsilon r(s)\boldsymbol{e}_r(s,\theta,t)$. Written out, the velocity field along the fiber surface is given by

$$\begin{aligned} 8\pi\mu\Big(\mathbf{u}^{\mathrm{SB}}(\mathbf{x}(s,\theta,t),t) - \mathbf{u}_0(\boldsymbol{X}(s,t),t)\Big) &= \\ -\int_{-L}^{L}\Bigg(\frac{\mathbf{I}}{|\boldsymbol{R}|} &+ \frac{\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathrm{T}} + \epsilon r(\overline{\boldsymbol{X}}\boldsymbol{e}_r^{\mathrm{T}} + \boldsymbol{e}_r\overline{\boldsymbol{X}}^{\mathrm{T}}) + \epsilon^2 r^2 \boldsymbol{e}_r\boldsymbol{e}_r^{\mathrm{T}}}{|\boldsymbol{R}|^3} \\ &+ \frac{\epsilon^2 r^2(s')}{2}\left(\frac{\mathbf{I}}{|\boldsymbol{R}|^3} - 3\frac{\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathrm{T}} + \epsilon r(\overline{\boldsymbol{X}}\boldsymbol{e}_r^{\mathrm{T}} + \boldsymbol{e}_r\overline{\boldsymbol{X}}^{\mathrm{T}}) + \epsilon^2 r^2 \boldsymbol{e}_r\boldsymbol{e}_r^{\mathrm{T}}}{|\boldsymbol{R}|^5}\right)\Bigg)\boldsymbol{f}(s',t)\,ds', \end{aligned} \tag{5.3.4}$$

where unless otherwise specified, we have $r = r(s)$, $\overline{\boldsymbol{X}} = \overline{\boldsymbol{X}}(s,s',t) = \boldsymbol{X}(s,t) - \boldsymbol{X}(s',t)$ and $\boldsymbol{R} = \boldsymbol{R}(s,s',\theta,t) = \overline{\boldsymbol{X}} + \epsilon r(s)\boldsymbol{e}_r(s,\theta,t)$. Now, along the fiber surface, the expression (5.3.4) satisfies the fiber integrity condition to leading order in ϵ; i.e. the terms containing $\boldsymbol{e}_r(s,\theta,t)$ in (5.3.4) vanish to $\mathcal{O}(\epsilon\log\epsilon)$, by equation (5.3.3). Because of this, to obtain an approximation to the velocity of the fiber itself which depends only on arclength, we could simply select a single curve along the length of the filament – i.e. fix $\theta=\theta^*$ or even $\theta=\theta^*(s)$ – and use the expression (5.3.4) evaluated along this curve as the approximate velocity of the fiber [40]. This yields an integral expression with a smooth, divergence-free kernel with clear physical meaning. However, this also involves a choice of θ^* and subsequent computation of a normal vector at each point along the fiber, which is unnecessarily complicated given that we know from (5.3.3) that the terms containing θ are small.

In particular, both the Stokeslet and doublet include a θ-dependent term with $\epsilon^2 r^2 \boldsymbol{e}_r \boldsymbol{e}_r^{\mathrm{T}}$ in the numerator. Due to the form of $\boldsymbol{R}$ in the denominator, both of these terms are $\mathscr{O}(1)$ at $s=s'$; however, upon integrating in s', these terms cancel each other asymptotically to order $\epsilon\log\epsilon$. In particular, by Lemmas 3.5 and 3.7 in (Refs. [35, 36]), respectively, we have

$$\left|\int_{-L}^{L} \frac{\epsilon^2 r^2 \boldsymbol{e}_r \boldsymbol{e}_r^{\mathrm{T}}}{|\boldsymbol{R}|^3} \boldsymbol{f}(s')\,ds' - 2\boldsymbol{e}_r \boldsymbol{e}_r \cdot \boldsymbol{f}(s)\right| \le C\epsilon\Big(\|\boldsymbol{f}\|_{C^1(-L,L)} + \Big\|\frac{\boldsymbol{f}}{r}\Big\|_{C^0(-L,L)}\Big),$$

$$\left|-\int_{-L}^{L} \frac{\epsilon^2 r^2(s')}{2} \frac{3\epsilon^2 r^2 \boldsymbol{e}_r \boldsymbol{e}_r^{\mathrm{T}}}{|\boldsymbol{R}|^3} \boldsymbol{f}(s')\,ds' + 2\boldsymbol{e}_r \boldsymbol{e}_r \cdot \boldsymbol{f}(s)\right| \le C\epsilon\Big(\|\boldsymbol{f}\|_{C^1(-L,L)} + \Big\|\frac{\boldsymbol{f}}{r}\Big\|_{C^0(-L,L)}\Big).$$

As we can see, the $O(1)$ contributions from both of these terms exactly cancel, leaving only higher order (in ϵ) contributions. Furthermore, the terms $\epsilon r(\overline{\boldsymbol{X}}\boldsymbol{e}_r^{\mathrm{T}} + \boldsymbol{e}_r\overline{\boldsymbol{X}}^{\mathrm{T}})$ in both the Stokeslet and doublet approximately integrate to zero in s', since, by Lemmas 3.4 and 3.6 in (Refs. [35, 36]), respectively, we have

$$\left|\int_{-L}^{L} \epsilon^m r^m(s') \frac{\epsilon r(\overline{\boldsymbol{X}}\boldsymbol{e}_r^{\mathrm{T}} + \boldsymbol{e}_r\overline{\boldsymbol{X}}^{\mathrm{T}})}{|\boldsymbol{R}|^{m+3}} \boldsymbol{f}(s')\,ds'\right| \le C\epsilon\Big(|\log\epsilon|\,\|\boldsymbol{f}\|_{C^1(-L,L)} + \Big\|\frac{\boldsymbol{f}}{r}\Big\|_{C^0(-L,L)}\Big), \quad m=0,2.$$

Finally, the $\boldsymbol{e}_r$ term in each denominator from $\left|\boldsymbol{R}(s,\theta,t)\right|^2 = \left|\overline{\boldsymbol{X}}\right|^2 + 2\epsilon r \boldsymbol{e}_r \cdot \overline{\boldsymbol{X}} + \epsilon^2 r^2$ is also only $O(\epsilon\log\epsilon)$, since, again using Lemmas 3.4 and 3.6 in (Refs. [35, 36]),

$$\left|\int_{-L}^{L} \left(\frac{\epsilon^m r^m(s')\boldsymbol{f}(s')}{|\boldsymbol{R}|^{m+1}} - \frac{\epsilon^m r^m(s')\boldsymbol{f}(s')}{(|\overline{\boldsymbol{X}}|^2+\epsilon^2 r^2)^{\frac{m+1}{2}}}\right) ds'\right| \le C\epsilon\Big(|\log\epsilon|\,\|\boldsymbol{f}\|_{C^1(-L,L)} + \Big\|\frac{\boldsymbol{f}}{r}\Big\|_{C^0(-L,L)}\Big), \quad m=0,2.$$

Due to these cancellations and the fact that dropping these terms still approximates the slender body PDE solution of (Refs. [35,36]) to at least $O(\epsilon\log\epsilon)$, we may eliminate all terms containing $\boldsymbol{e}_r(s,\theta,t)$ in (5.3.4) to obtain a θ-independent expression which approximates the velocity of the fiber itself:

$$\begin{aligned} 8\pi\mu\left(\frac{\partial \boldsymbol{X}}{\partial t} - \mathbf{u}_0(\boldsymbol{X}(s,t),t)\right) = -\int_{-L}^{L} \Bigg(&\frac{\mathbf{I}}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{1/2}} + \frac{\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^{\mathrm{T}}}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{3/2}} \\ &+ \frac{\epsilon^2 r^2(s')}{2}\left(\frac{\mathbf{I}}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{3/2}} - \frac{3\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^{\mathrm{T}}}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{5/2}}\right)\Bigg)\boldsymbol{f}(s',t)\,ds'. \end{aligned} \tag{5.3.5}$$

The expression (5.3.5) serves as the model underlying our final slender body velocity expression (5.2.1). One further limitation to note about the centerline expressions (5.2.1) and (5.3.5) is that because the model is essentially 1D, in certain special cases (i.e. when the fiber is straight and its axis is perfectly aligned with the flow), the slender body approximation, in contrast to a truly 3D fiber, does not pick up on fluid gradients (see Section 5.5.2).

In Section 5.3.1, we show that in a simplified setting, (5.3.5) results in a negative definite operator mapping the force density $\boldsymbol{f}$ to the fiber velocity $\frac{\partial \boldsymbol{X}}{\partial t}$, whereas other models which rely on further asymptotic expansion of (5.3.4) about $\epsilon = 0$ do not, and incur high wavenumber instabilities. This phenomenon is well known for the Keller–Rubinow model [18,25], but for other possible centerline expressions, including models similar to Lighthill [29], this high wavenumber instability has not been documented previously. It seems that our model (5.3.5) may be the simplest that can be obtained by expanding from (5.3.4) while still guaranteeing a negative definite operator.

Now, since the integral operator in (5.3.5) has a smooth kernel, the expression (5.3.5) yields a first-kind Fredholm integral equation for $\boldsymbol{f}$ when the fiber velocity $\frac{\partial \boldsymbol{X}}{\partial t}$ is supplied. Describing the motion of a rigid fiber involves inverting this expression to solve for $\boldsymbol{f}$, which in general is an ill-posed problem for a first-kind equation. Thus we want to regularize the integral operator (5.3.5) to create a second-kind integral equation while keeping the same order of accuracy in the map $\boldsymbol{f} \mapsto \frac{\partial \boldsymbol{X}}{\partial t}$.

We first note that, for $\eta > 1$, we have the following identity:

$$\int_{-L}^{L} \left(\frac{1}{(|\overline{\boldsymbol{X}}|^2 + \epsilon^2 r^2(s))^{1/2}} - \frac{1}{(|\overline{\boldsymbol{X}}|^2 + \eta^2 \epsilon^2 r^2(s))^{1/2}} \right) g(s')\, ds' = 2\log(\eta)\, g(s) + \mathcal{O}(\eta\epsilon \log(\eta\epsilon)). \tag{5.3.6}$$

Proof. By Lemma 3.8 in (Ref. [36]), for $a > 0$ sufficiently small, we have

$$\int_{-L}^{L} \left(\frac{g(s')}{(|\overline{\boldsymbol{X}}|^2 + a^2 r^2(s))^{1/2}} - \frac{g(s')}{|\overline{\boldsymbol{X}}|} + \frac{g(s)}{|s - s'|} \right) ds' \\ = \log\left(\frac{2(L^2 - s^2) + 2\sqrt{(L^2 - s^2)^2 + a^2 r^2(s)}}{a^2 r^2(s)} \right) + \mathcal{O}(a \log a). \tag{5.3.7}$$

Subtracting (5.3.7) with $a = \eta\epsilon$ from (5.3.7) with $a = \epsilon$ and using that

$$\left| \log\left(\frac{(L^2 - s^2) + \sqrt{L^2 + \epsilon^2 r^2}}{(L^2 - s^2) + \sqrt{L^2 + \eta^2 \epsilon^2 r^2}} \right) \right| = \left| \log\left(\frac{(L^2 - s^2) + \sqrt{L^2 + \epsilon^2 r^2}}{(L^2 - s^2) + \sqrt{L^2 + \eta^2 \epsilon^2 r^2}} \right) - \log(1) \right| \le C\epsilon^2,$$

we obtain (5.3.6). □

Using (5.3.6), we replace the first term in the integrand of (5.3.5) to obtain (5.2.1). We can compare the expression (5.2.1) to that of Tornberg and Shelley

[55], where a regularization of the Keller–Rubinow model is used to obtain a second-kind integral equation for $\boldsymbol{f}$. One thing to note is that, due to the form of the local term in our model (5.2.1), the effect of the regularization parameter η is the same in all directions (both tangent and normal to the fiber centerline). This is not necessarily the case for the Tornberg and Shelley model (see Section 5.3.1 for a spectral comparison given a simplified fiber geometry).

5.3.1 Spectral comparison of slender body integral operators

In this subsection we provide evidence that our model (5.2.1) is well suited for approximating the map $\frac{\partial \boldsymbol{X}}{\partial t} \mapsto \boldsymbol{f}$ needed to simulate the motion of a rigid fiber. Here we consider the spectrum of the integral operator taking the force density $\boldsymbol{f}$ to the fiber velocity $\frac{\partial \boldsymbol{X}}{\partial t}$ in the non-physical but nevertheless instructive case of a straight, periodic fiber with constant radius ϵ. In this scenario we can explicitly calculate the eigenvalues of both the slender body PDE operator (5.3.2) as well as the integral operator (5.3.5) and related models. This allows us to directly compare the properties of different models in the same simple setting and serves as a starting point for understanding more complicated geometries. In particular, we expect this analysis to roughly capture the high wavenumber behavior of these models in different geometries – on length scales much smaller than the variation in curvature and fiber radius. The high wavenumber behavior is of particular interest for the invertibility and stability of the slender body theory integral operator.

For comparison, we first recall the form of the eigenvalues of the slender body PDE (5.3.2), calculated in (Ref. [34]). In Section 5.3.1, we consider the model (5.3.5), before regularization, and show that the integral operator is negative definite. We compare the spectrum of (5.3.5) to three other possible models based on slender body theory which do not result in negative definite operators. Then in Section 5.3.1, we consider the regularized version of our model (5.2.1) and compare its spectrum to the regularized model of Tornberg and Shelley [55]. We note that in our model, a uniform regularization parameter appears to give the best approximation of the slender body PDE spectrum in directions both normal and tangent to the slender body centerline, whereas in the Tornberg–Shelley model, the parameter required by the tangential direction may not be optimal in the normal direction.

Spectrum of the slender body PDE

Here we consider a straight, periodic fiber with constant radius ϵ. We take the fiber centerline to be 2-periodic and lie along the z-axis, $\boldsymbol{X}(z) = z\boldsymbol{e}_z$, $z \in \mathbb{R}/2\mathbb{Z}$, and for simplicity take $\mu = 1$ and zero background flow. We consider the stationary setting and omit the time dependence in our notation; in particular, we denote the fiber velocity by $\overline{\mathbf{u}}(z)$ to distinguish from the fluid velocity away from the fiber.

We consider this scenario because we can explicitly calculate the eigenvalues of the slender body PDE (5.3.2) as well as various possible integral expressions for approximating the map $\boldsymbol{f} \mapsto \overline{\mathbf{u}}$. In particular, the eigenvectors of this map can be decomposed into tangential ($\boldsymbol{e}_z$) and normal ($\boldsymbol{e}_x, \boldsymbol{e}_y$) directions and are given by $\boldsymbol{f}_m(z) = e^{i\pi kz}\boldsymbol{e}_m$, $m = x, y, z$. We may then explicitly solve for λ_k^m satisfying

$$\overline{\mathbf{u}}(z) = \lambda_k^m \boldsymbol{f}_m(z), \qquad m = x, y, z \tag{5.3.8}$$

for both the slender body PDE operator and various approximations based on slender body theory. To avoid logarithmic growth of the corresponding bulk velocity field at spatial infinity, we will ignore translational modes ($k = 0$) in the following spectral analysis. Clearly these modes are important, especially for a rigid body; however, we are mainly interested in the high wavenumber behavior of these operators. High wavenumber instabilities are a known issue for nonlocal slender body theory [18,45,55], and the following analysis likely captures the behavior of these models at high wavenumbers (small length scales) even in curved geometries.

To begin, the eigenvalues of the slender body PDE operator (5.3.2) mapping $\boldsymbol{f}$ to $\overline{\mathbf{u}}$ were calculated in (Ref. [34], Proposition 1.4). Note that the sign convention in this paper is opposite, as we are considering $\boldsymbol{f}$ to be the hydrodynamic force exerted *by* rather than *on* the slender body. For the slender body PDE, the eigenvalues satisfying (5.3.8) in the tangential and normal directions, respectively, are given by

$$\lambda_k^m = \begin{cases} -\frac{2K_0K_1+\pi\epsilon|k|\left(K_0^2-K_1^2\right)}{4\pi^2\epsilon|k|K_1^2}, & m = z \\ -\frac{2K_0K_1K_2+\pi\epsilon|k|\left(K_1^2(K_0+K_2)-2K_0^2K_2\right)}{2\pi^2\epsilon|k|\left(4K_1^2K_2+\pi\epsilon|k|K_1(K_1^2-K_0K_2)\right)}, & m = x, y \end{cases} \tag{5.3.9}$$

where each $K_j = K_j(\pi\epsilon|k|)$, $j = 0, 1, 2$, is a j^{th} order modified Bessel function of the second kind. Note that both sets of eigenvalues λ_k^z and λ_k^x, λ_k^y are strictly negative and decay to 0 at a rate proportional to $1/|k|$ as $|k| \to \infty$. We will compare our approximation and various other slender body approximations to (5.3.9).

Pre-regularization comparison

Before we consider the regularized version (5.2.1) of our model, we consider the base model (5.3.5) and compare its spectrum to other existing models based on slender body theory, before regularization. In the straight-but-periodic scenario, our model (5.3.5) becomes the periodization of the expression

$$\overline{\mathbf{u}}(z) = -\frac{1}{8\pi}\int_{-1}^{1}\left(\frac{\mathbf{I}}{(\overline{z}^2+\epsilon^2)^{1/2}}+\frac{\overline{z}^2\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{3/2}}+\frac{\epsilon^2}{2}\left(\frac{\mathbf{I}}{(\overline{z}^2+\epsilon^2)^{3/2}}-3\frac{\overline{z}^2\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{5/2}}\right)\right)\boldsymbol{f}(z-\overline{z})\,d\overline{z}. \tag{5.3.10}$$

For this geometry, we may calculate the eigenvalues λ_k^m satisfying (5.3.8), which are given by

$$\lambda_k^m = \begin{cases} -\dfrac{1}{8\pi}\displaystyle\int_{-1}^{1}\frac{2\overline{z}^4+2\epsilon^2\overline{z}^2+\frac{3}{2}\epsilon^4}{(\overline{z}^2+\epsilon^2)^{5/2}}e^{-i\pi k\overline{z}}\,d\overline{z}, & m=z \\ -\dfrac{1}{8\pi}\displaystyle\int_{-1}^{1}\frac{\overline{z}^2+\frac{3}{2}\epsilon^2}{(\overline{z}^2+\epsilon^2)^{3/2}}e^{-i\pi k\overline{z}}\,d\overline{z}, & m=x,y. \end{cases} \tag{5.3.11}$$

These integrals may be computed explicitly to obtain

$$\lambda_k^m = \begin{cases} -\dfrac{1}{8\pi}\Big((4+\pi^2\epsilon^2k^2)K_0(\pi\epsilon|k|)-2\pi\epsilon|k|\,K_1(\pi\epsilon|k|)\Big), & m=z \\ -\dfrac{1}{8\pi}\Big(2K_0(\pi\epsilon|k|)+\pi\epsilon|k|\,K_1(\pi\epsilon|k|)\Big), & m=x,y. \end{cases} \tag{5.3.12}$$

Here K_0 and K_1 are zero and first order modified Bessel functions of the second kind, respectively. The eigenvalues λ_k^m lie along the curves plotted in Figure 5.3.1. Importantly, these eigenvalues satisfy the following lemma.

Lemma 5.1. *For all $|k| \geq 1$ and $m = x, y, z$, the eigenvalues λ_k^m given by* (5.3.12) *satisfy $\lambda_k^m < 0$.*

Proof. The case $m = x, y$ is immediate, since $K_0(t) > 0$ and $K_1(t) > 0$ for any $t > 0$.

For the tangential direction $m = z$, we first note that, by Lemma 1.16 in (Ref. [34]), we have

$$1 \leq \frac{K_1(t)}{K_0(t)} \leq 1+\frac{1}{2t}$$

for all $t > 0$. Letting $g(t) = (4+t^2)K_0(t) - 2tK_1(t)$, it suffices to show that $g(t)/K_0(t) > 0$. But

$$\frac{g(t)}{K_0(t)} = 4+t^2-2t\frac{K_1(t)}{K_0(t)} \geq 3+t^2-2t > (t-\sqrt{3})^2 \geq 0.$$

□

Now, at a continuous level, regularization is necessary to make sense of inverting the integral operator (5.3.10), since K_0 and K_1 decay exponentially as $|k| \to \infty$. However, at a discrete level, numerical approximation of (5.3.10) will be invertible, albeit with a large condition number, due to Lemma 5.1. This negativity does not hold for other popular slender body approximations which rely on further asymptotic expansion of (5.3.5) with respect to ϵ to obtain a limiting centerline velocity expression. In particular, we consider the models of Keller and Rubinow [25] and of Lighthill [29].

The Keller–Rubinow model, proposed in (Ref. [25]) and further studied by (Refs. [18, 24, 45, 55]), is equivalent to a full matched asymptotic expansion of (5.3.4) about $\epsilon = 0$. In the straight-but-periodic setting, the Keller–Rubinow expression for the slender body velocity is given by

$$8\pi\overline{\mathbf{u}}(z) = -\Big((\mathbf{I}-3\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})-2\log(\pi\epsilon/8)(\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\Big)\boldsymbol{f}(z)-(\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\frac{\pi}{2}\int_{-1}^{1}\frac{\boldsymbol{f}(z-\overline{z})-\boldsymbol{f}(z)}{\left|\sin(\pi\overline{z}/2)\right|}\,d\overline{z}. \tag{5.3.13}$$

The eigenvalues of the periodic Keller–Rubinow operator taking $\boldsymbol{f}$ to $\overline{\mathbf{u}}$ have been calculated in (Refs. [18, 45, 55]) and are given by

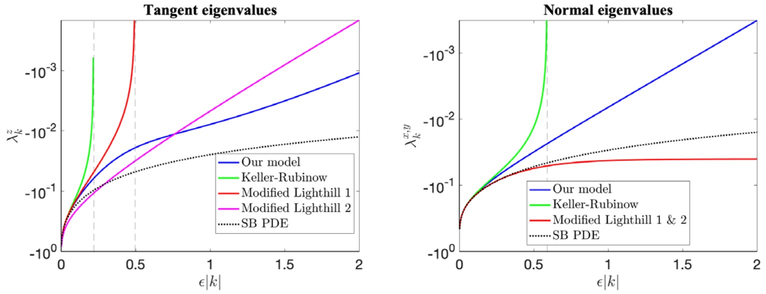
$$\lambda_k^m = \begin{cases} \frac{1}{4\pi}\big(1+2\log(\pi\epsilon\left|k\right|/2)+2\gamma\big), & m = z \\ -\frac{1}{8\pi}\big(1-2\log(\pi\epsilon\left|k\right|/2)-2\gamma\big), & m = x, y. \end{cases} \tag{5.3.14}$$

Here $\gamma \approx 0.5772$ is the Euler gamma.

In both the tangent and normal directions, however, the Keller–Rubinow approximation runs into stability issues at moderately high wavenumbers, apparent in Figure 5.3.1 at $\left|k\right| = \frac{2e^{-\gamma-1/2}}{\pi\epsilon} \approx 0.217/\epsilon$ (tangent) and $\left|k\right| = \frac{2e^{-\gamma+1/2}}{\pi\epsilon} \approx 0.589/\epsilon$ (normal). In particular, the curve containing the eigenvalues λ_k^m crosses zero and becomes negative. This is an issue both because the slender body PDE eigenvalues (5.3.9) are strictly negative, and because, for arbitrary ϵ, there is no clear way to guarantee that $\lambda_k^m \neq 0$, especially for more complicated fiber geometries. Thus some sort of regularization of (5.3.13) is necessary before approximating the inverse map $\overline{\mathbf{u}} \mapsto \boldsymbol{f}$.

In addition to the Keller–Rubinow model, we consider what we will term the *modified Lighthill* approach to deriving a fiber velocity approximation. This approach, due to Lighthill [29], also begins with the classical SBT expression (5.3.4) but uses asymptotic integration of the doublet term to arrive at an expression for the fiber velocity. We explore the Lighthill method in detail in Appendix 5.A, but plot the resulting spectrum in Figure 5.3.1.

The takeaway here is that, at least in the case of a straight, periodic fiber, our model (5.3.5), before regularization, captures the negative-definiteness of the the slender body PDE and provides a better approximation than other models based on classical SBT.



(a) **(b)**

Figure 5.3.1: Log-scale plot of the tangential (a) and normal (b) eigenvalues λ_k^m of the operator mapping $\boldsymbol{f} \mapsto \bar{\mathbf{u}}$ in various slender body models for a straight-but-periodic fiber. Our model (blue) results in strictly negative eigenvalues in both the tangential and normal directions, as does the slender body PDE (dotted). The Keller–Rubinow approximation (green) exhibits instabilities at wavenumbers $|k| \approx 0.2/\epsilon$ (tangential direction) and $|k| \approx 0.6/\epsilon$ (normal direction) as the eigenvalues of the operator mapping $\boldsymbol{f} \mapsto \bar{\mathbf{u}}$ become positive. For the modified Lighthill models, the normal direction eigenvalues λ_k^x and λ_k^y (red) remain negative at high wavenumber, but in the tangential direction, the eigenvalues of Modified Lighthill 1 (red) become positive when $|k| > 0.5/\epsilon$. Furthermore, the tangential eigenvalues of Modified Lighthill 2 (magenta) do not agree with the slender body PDE at low wavenumber.

Regularized comparison

To make our model truly suitable for inversion, we need to regularize the integral kernel as in (5.2.1). In the straight-but-periodic setting, the operator in (5.2.1) becomes the periodization of

$$
\begin{aligned}
8\pi\bar{\mathbf{u}}(z) = -2\log(\eta)\,\boldsymbol{f}(z) - \int_{-1}^{1}\Bigg(&\frac{\mathbf{I}}{(\overline{z}^2+\eta^2\epsilon^2)^{1/2}} + \frac{\overline{z}^2\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{3/2}} \\
&+\frac{\epsilon^2}{2}\left(\frac{\mathbf{I}}{(\overline{z}^2+\epsilon^2)^{3/2}} - 3\frac{\overline{z}^2\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{5/2}}\right)\Bigg)\boldsymbol{f}(z-\overline{z})\,d\overline{z}.
\end{aligned}
\tag{5.3.15}
$$

The eigenvalues of (5.3.15) are then given by

$$
\lambda_k^m = \begin{cases} -\dfrac{1}{8\pi}\Big(2\log(\eta) + 2K_0(\eta\pi\epsilon|k|) + (2+\pi^2\epsilon^2k^2)K_0(\pi\epsilon|k|) - 2\pi\epsilon|k|\,K_1(\pi\epsilon|k|)\Big), & m = z \\ -\dfrac{1}{8\pi}\Big(2\log(\eta) + 2K_0(\eta\pi\epsilon|k|) + \pi\epsilon|k|\,K_1(\pi\epsilon|k|)\Big), & m = x, y. \end{cases}
\tag{5.3.16}
$$

For $\eta > 1$, the spectrum of our operator is bounded away from 0 and (5.3.15) is a second-kind integral equation for $\boldsymbol{f}$.

We can compare the behavior of (5.3.15) with the Tornberg–Shelley regularization of the Keller–Rubinow model. In (Refs. [45, 55]), the high wavenumber instability in (5.3.13) is removed by replacing the denominator of the integral term, which vanishes at $\overline{z}=0$, with an expression proportional to ϵ at $\overline{z}=0$. Using the relation

$$\int_{-1}^{1}\left(\frac{\pi}{\left|2\sin(\pi z/2)\right|}-\frac{1}{|z|}\right)dz=-2\log(\pi/4) \tag{5.3.17}$$

to rewrite (5.3.13), a regularization $\delta\epsilon$, $\delta>0$, is added to the denominator to obtain

$$8\pi\overline{\mathbf{u}}(z)=-\left((\mathbf{I}-3\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})+2\log(\delta)(\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\right)\boldsymbol{f}(z)-(\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\int_{-1}^{1}\frac{\boldsymbol{f}(z-\overline{z})}{(\overline{z}^2+\delta^2\epsilon^2)^{1/2}}\,d\overline{z}. \tag{5.3.18}$$

Here we have also used that the second term in the original Keller–Rubinow integral expression can now be integrated up to $O(\epsilon^2)$ errors to nearly cancel the logarithmic term in (5.3.13), leaving only $\log(\delta)$. The idea is to then choose δ such that all eigenvalues of the operator taking $\boldsymbol{f}\mapsto\overline{\mathbf{u}}$ are negative. Since the integral kernel is now smooth, (5.3.18) is now a second-kind integral equation for $\boldsymbol{f}$.

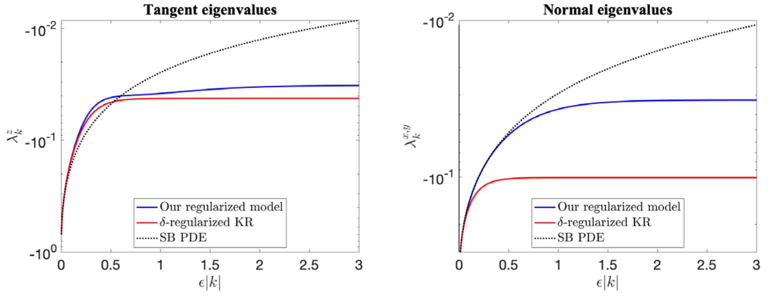
The eigenvalues of this δ-regularized Keller–Rubinow operator are given by

$$\lambda_k^m=\begin{cases}-\dfrac{1}{4\pi}\left(-1+2\log\delta+2K_0(\delta\pi\epsilon|k|)\right), & m=z\\[1ex] -\dfrac{1}{8\pi}\left(1+2\log\delta+2K_0(\delta\pi\epsilon|k|)\right), & m=x,y.\end{cases} \tag{5.3.19}$$

Since K_0 is positive, λ_k^z is guaranteed to be negative and bounded away from 0 as long as $\delta>\sqrt{e}$ (see Figure 5.3.2).

Note that in our model (5.3.15), the regularization parameter η affects the spectrum of the operator mapping $\boldsymbol{f}$ to $\overline{\mathbf{u}}$ in the same way in both the tangential and normal directions. In particular, in both directions, $\eta>1$ is required to obtain the desired second-kind integral equation. In the Tornberg–Shelley model, the bound $\delta>\sqrt{e}\approx 1.649$ is required to ensure negativity of the tangential eigenvalues, but this lower bound does not apply to the normal direction; in fact, $\delta>e^{-1}\approx 0.368$ is sufficient for ensuring strictly negative normal eigenvalues. This may mean that the η-regularization in our model is more physically reasonable; see Figure 5.3.2.

In (Ref. [34]), it is shown that using the δ-regularized model (5.3.18) to approximate the map $\overline{\mathbf{u}}\mapsto\boldsymbol{f}$ yields ϵ^2 convergence to the slender body PDE for sufficiently smooth $\overline{\mathbf{u}}$. It is also shown that the constant in the resulting error estimate has the form $C_1\delta^2(1+\log(\delta))+C_2/(-1+\log(\delta))$ for constants C_1 and C_2. We expect that a similar error estimate and analogous η dependence hold for our model (5.3.15); i.e. the constant should look like $C_1\eta^2+C_2/\log(\eta)$. If $C_1\approx C_2$, this yields an optimal η of approximately 1.5. This should give a



(a) **(b)**

Figure 5.3.2: Log-scale plot of the tangential (a) and normal (b) eigenvalues λ_k^m of our regularized model (5.3.15) (blue) with $\eta = 1.5$ and the Tornberg–Shelley δ-regularized model (5.3.18) (red) with $\delta = \sqrt{e}+0.5$. Note that the regularization parameter η in our model affects the tangential and normal eigenvalues in a similar way; in particular, $\eta > 1$ is required in both cases to ensure that (5.3.15) is a second-kind integral equation. In the δ-regularized model, the tangential direction requires $\delta > \sqrt{e}$, but the normal direction does not, resulting at least visually in a greater disparity between the λ_k^x, λ_k^y for the PDE (dotted) and the δ-regularized approximation.

rough guideline for a good choice of η for more general curved geometries, at least in the periodic setting.

Remark 5.1. We can also consider using the method of regularized Stokeslets to rederive the Keller–Rubinow model (see Ref. [13]). Here the following choices of blob functions are used in place of Dirac deltas to derive the regularized Stokeslet and doublet, respectively:

$$\phi_S(\boldsymbol{R}) = \frac{15}{8\pi}\frac{\delta^4\epsilon^4}{(|\boldsymbol{R}|^2+\delta^2\epsilon^2)^{7/2}}, \qquad \phi_D(\boldsymbol{R}) = \frac{3}{4\pi}\frac{\delta^2\epsilon^2}{(|\boldsymbol{R}|^2+\delta^2\epsilon^2)^{5/2}}.$$

Note that we have modified the notation from (Ref. [13]) to emphasize that the blob "width" will be taken to be proportional to the fiber radius ϵ, and to more easily compare with the δ-regularization of Tornberg–Shelley. For the straight-but-periodic fiber, this method yields a nearly identical expression to (5.3.18), but with a different logarithmic factor in front of the local terms: $-\log(\sqrt{\delta^2+1}/\delta)$ in place of $\log(\delta)$. Due to the low wavenumber expansion (5.A.6) of the Bessel function K_0, however, we note that the $\log(\delta)$ term in (5.3.19) exactly cancels the leading order dependence of $K_0(\delta\pi\epsilon|k|)$ on δ, yielding an expression consistent with the slender body PDE (5.3.9) when $|k|$ is small. When $\delta \ll 1$, we have $-\log(\sqrt{\delta^2+1}/\delta) \approx \log(\delta)$, but recall that $\delta > \sqrt{e}$ is required for (5.3.19) to be negative for all k. Thus this particular choice of blob function in the method of regularized Stokeslets appears to yield an expression for the fiber velocity which fundamentally differs from the slender body PDE solution, although a different choice of blob function may yield closer agreement. Note that this low wavenumber descrepancy occurs whether we start from the non-periodic or periodic regularized expressions mentioned in (Ref. [13]), due to the identity (5.3.17).

5.4 Numerical discretization of the slender body model

We turn now to numerically simulating thin rigid fibers in flows. We begin by generally discussing the numerical solution of Fredholm integral equations where the result must be integrated (i.e. to find the total force and torque on a rigid fiber). We apply these general methods to the slender body model (5.2.1) and perform convergence tests. We note improvements in conditioning and stability for the second kind ($\eta > 1$) versus first kind ($\eta = 1$) integral equation. Finally, we look at the spectrum of the discretized integral operator in different geometries to verify the negative definite nature of the operator.

5.4.1 Solving the second-kind Fredholm integral equation

Denote by $\mathbf{K}\colon L^2([-L,L],\mathbb{R}^3) \to L^2([-L,L],\mathbb{R}^3)$ the integral operator

$$\mathbf{K}[\mathbf{f}](s) := \int_{-L}^{L} K(s,s')\mathbf{f}(s')ds'. \tag{5.4.1}$$

Then a Fredholm integral equation of the first kind reads

$$\mathbf{y}(s) = \mathbf{K}[\mathbf{f}](s). \tag{5.4.2}$$

It is well known that the inversion of such an integral operator is an ill-posed problem, meaning that the solution may not be unique or not even exist [2, 22, 26]. Furthermore, small perturbations to the left hand side of (5.4.2) can lead to relatively large perturbations of the solution $\mathbf{f}(s)$. The ill-posedness of this problem can be circumvented by regularizing the integral operator into a second-kind Fredholm integral equation, which takes the form

$$\mathbf{y}(s) = (\alpha\mathbf{I} + \mathbf{K})[\mathbf{f}](s) \tag{5.4.3}$$

for some parameter α. Discretization of (5.4.3) yields a linear system with a far better condition number. The connection between equation (5.4.3) and our model is illustrated in Section 5.4.2.

Numerical methods for solving Fredholm integral equations are well documented [22, 57] and the approach we adopt is based on the Nyström method (see Ref. [2, Chapt. 12.4]). For rigid fibers, after numerically inverting a second-kind Fredholm integral equation, linear functionals (5.2.5) will also need to be applied to the resulting $\mathbf{f}(s)$ to find the total force and torque.

We consider the numerical approximation of a general linear functional of $\mathbf{f}(s)$, given by

$$\phi_M(\mathbf{f}) = \int_{-L}^{L} M(s)\mathbf{f}(s)\mathrm{d}s. \tag{5.4.4}$$

Here $M(s) \in \mathbb{R}^{3\times3}$ is a bounded, smooth operator and $\mathbf{f}(s)$ is found by numerically inverting a second-kind Fredholm integral equation of the form (5.4.3). The numerical method is obtained discretizing the equation (5.4.3) by replacing the integral with a convergent quadrature formula with nodes $-L = s_1 < s_2 < ... < s_n = L$ and weights $\mathbf{w} = (w_1, w_2, ..., w_n)^T \in \mathbb{R}^n$, and requiring the numerical approximation $\mathbf{f}_i^{[n]} \approx \mathbf{f}(s_i)$ to satisfy

$$\mathbf{y}(s_i) = \alpha \mathbf{f}_i^{[n]} + \sum_{j=1}^{n} w_j K(s_i, s_j)\mathbf{f}_j^{[n]} \quad \text{for} \quad i = 1, ..., n. \tag{5.4.5}$$

Introducing the vectors $\underline{\mathbf{f}}^{[n]} = ((\mathbf{f}_1^{[n]})^T, ..., (\mathbf{f}_n^{[n]})^T)^T$ and $\underline{\mathbf{y}} = (\mathbf{y}(s_1)^T, ..., \mathbf{y}(s_n)^T)^T$, equation (5.4.5) can be written compactly as

$$\underline{\mathbf{y}} = \left(\alpha I + \underline{K}\,\underline{W}\right)\underline{\mathbf{f}}^{[n]}. \tag{5.4.6}$$

Here I denotes the $3n \times 3n$ identity matrix, and

$$\underline{W} = \mathrm{diag}(\boldsymbol{w}) \otimes \mathbf{I}, \quad \text{and} \quad \underline{K} = \begin{pmatrix} K(s_1, s_1) & \dots & K(s_1, s_n) \\ \vdots & \ddots & \vdots \\ K(s_n, s_1) & \dots & K(s_n, s_n) \end{pmatrix} \in \mathbb{R}^{3n\times3n} \tag{5.4.7}$$

with $\otimes : \mathbb{R}^{n_1\times m_1} \times \mathbb{R}^{n_2\times m_2} \to \mathbb{R}^{(n_1n_2)\times(m_1m_2)}$ the Kronecker product of matrices and $\mathbf{I}$ the 3×3 identity matrix. We then approximate (5.4.4) by the same quadrature formula

$$\phi_M(\mathbf{f}) \approx \sum_{i=1}^{n} w_i M(s_i)\mathbf{f}_i^{[n]} = (\vec{1}^T \otimes \mathbf{I})\underline{M}\,\underline{W}\,\underline{\mathbf{f}}^{[n]} := \phi_M^{[n]}, \tag{5.4.8}$$

where

$$\underline{M} = \begin{pmatrix} M(s_1) & & 0 \\ & \ddots & \\ 0 & & M(s_n) \end{pmatrix} \in \mathbb{R}^{3n\times3n} \tag{5.4.9}$$

and $\vec{1} = (1, \dots, 1)^T \in \mathbb{R}^n$. Here we have used $\phi_M^{[n]}$ to denote the approximation of $\phi_M(\mathbf{f})$ obtained by quadrature. After inserting the solution of (5.4.6), we obtain

$$\phi_M^{[n]} = (\vec{1}^T \otimes \mathbf{I})\underline{M}\,\underline{W}\left(\alpha I + \underline{K}\,\underline{W}\right)^{-1}\underline{\mathbf{y}}. \tag{5.4.10}$$

Remark 5.2. The numerical approximation $\phi_M^{[n]}$ shares the same convergence as the underlying quadrature method. This is illustrated in appendix 5.B.

5.4.2 Application to the slender body model and convergence tests

We apply the numerical method from Section 5.4.1 to approximate the force and torque on a slender body. Note that the equations (5.2.5) are given by setting $M(s) = \mathbf{I}$ and $M(s) = \widehat{\boldsymbol{X}}(s)$ in the functional (5.4.4). That is,

$$\boldsymbol{F} = \phi_{\mathbf{I}}(\mathbf{f}) \quad \text{and} \quad \boldsymbol{T} = \phi_{\widehat{\mathbf{X}}}(\mathbf{f}). \tag{5.4.11}$$

Letting $\alpha = 2\log(\eta)$ and

$$K(s, s') = \boldsymbol{S}_{\epsilon,\eta}(s, s') + \frac{\epsilon^2 r^2(s')}{2}\boldsymbol{D}_{\epsilon}(s, s'), \tag{5.4.12}$$

$$\mathbf{y}(s) = -8\pi\mu(\mathbf{v} - \widehat{\boldsymbol{X}}(s)\boldsymbol{\omega} - \mathbf{u}_0(\boldsymbol{X}(s,t),t)), \tag{5.4.13}$$

our model (5.2.1) is of the form (5.4.3), and we may write the discretization of (5.2.1) in the form (5.4.6). Here we have introduced the hat operator $\widehat{\cdot} : \mathbb{R}^3 \to \mathfrak{so}(3)$ which maps vectors in $\mathbb{R}^3$ to 3 × 3 skew symmetric matrices by

$$\boldsymbol{\omega} = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} \mapsto \widehat{\boldsymbol{\omega}} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}. \tag{5.4.14}$$

Here, $\mathfrak{so}(3)$ is the Lie algebra of $SO(3)$, and such that $\boldsymbol{\omega} \times \mathbf{v} = \widehat{\boldsymbol{\omega}}\mathbf{v}$ for $\boldsymbol{\omega}, \mathbf{v} \in \mathbb{R}^3$.

Denote the numerical approximations to (5.4.11) by

$$\boldsymbol{F}^{[n]} = \phi_{\mathbf{I}}^{[n]} \quad \text{and} \quad \boldsymbol{T}^{[n]} = \phi_{\widehat{\mathbf{X}}}^{[n]}. \tag{5.4.15}$$

Defining the matrices Φ and $\Psi \in \mathbb{R}^{3\times 3n}$ as

$$\Phi = (\vec{1}^T \otimes \mathbf{I})\underline{W}\left(\alpha I + \underline{K}\,\underline{W}\right)^{-1}, \tag{5.4.16}$$

$$\Psi = (\vec{1}^T \otimes \mathbf{I})\underline{W}\,\underline{X}\left(\alpha I + \underline{K}\,\underline{W}\right)^{-1}, \tag{5.4.17}$$

we may then write equations (5.4.15) as

$$\boldsymbol{F}^{[n]} = \Phi\underline{\mathbf{y}} \quad \text{and} \quad \boldsymbol{T}^{[n]} = \Psi\underline{\mathbf{y}}. \tag{5.4.18}$$

In the next section we perform convergence tests for our discrete model (5.4.18) for both a thin ring and a prolate spheroid. With these geometries we are able to calculate accurate reference solutions against which we can compare the accuracy of our numerical solution. Furthermore, we will look at how the conditioning of the linear system associated with the discretized integral operator improves as the regularization parameter η is increased from $\eta = 1$ to $\eta > 1$.

Remark 5.3. For very large aspect ratios, e.g., $L/\epsilon \approx O(10^3)$ or larger, the kernel becomes very nearly singular meaning one must take n very large to accurately resolve the $O(\epsilon)$ length scales in the kernel. In this case, the quadrature can be improved by implementing special quadrature methods that take into account the near singular nature of the integral kernel [1,53]. For modest aspect ratios, e.g., $L/\epsilon \approx O(10^2)$, this is not an issue as one can accurately resolve the kernel with a few hundred points. It has been shown that slender body theories are good approximations for particles of aspect ratios larger than 20 [46].

Thin ring translating with unit velocity

As a convergence test, we use (5.4.18) to calculate the force on a thin ring of unit length in the xy-plane translating in the z direction with unit velocity in zero background flow. We will consider both the first- and second-kind formulations of the model. In this setting, the force on the ring can be calculated to arbitrarily high precision by evaluating elliptic integrals, which can be used as a reference solution. For a circular centerline parametrized by

$$\boldsymbol{X}(s) = \left(\frac{\cos(\pi s)}{2\pi}, \frac{\sin(\pi s)}{2\pi}, 0\right)^T,$$

the z-component of our unregularized ($\eta = 1$) model becomes

$$8\pi\mu = -\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{\sqrt{2}\pi\left(3\epsilon^2\pi^2 - \cos\left(2\pi\left(s-s'\right)\right) + 1\right)}{\left(2\epsilon^2\pi^2 - \cos\left(2\pi\left(s-s'\right)\right) + 1\right)^{3/2}} f^z(s')\, ds'. \tag{5.4.19}$$

As in the straight-but-periodic geometry of Section 5.3.1, the eigenfunctions of this operator are the Fourier modes $f_k^z(s) = \exp(i2\pi ks)$. The force $\boldsymbol{F} = (F, 0, 0)^T$ is therefore given by

$$F = \int_{-\frac{1}{2}}^{\frac{1}{2}} f^z(s)\mathrm{d}s = \frac{8\pi\mu}{\lambda_0^z} \tag{5.4.20}$$

where λ_0^z is the $k = 0$ eigenvalue. This can be found by evaluating the integral in equation (5.4.19) with $f^z(s) = f_0^z(s) = 1$, which gives

$$\lambda_0^z = -c_\epsilon\left(2\phi_K\left(c_\epsilon\right) + \phi_E\left(c_\epsilon\right)\right). \tag{5.4.21}$$

Here $c_\epsilon = \sqrt{\left(\epsilon^2\pi^2 + 1\right)^{-1}}$, and

$$\phi_K(x) = \int_0^1 \frac{1}{\sqrt{1-\theta^2}\sqrt{1-x^2\theta^2}}\mathrm{d}\theta \quad \text{and} \quad \phi_E(x) = \int_0^1 \frac{\sqrt{1-x^2\theta^2}}{\sqrt{1-\theta^2}}\mathrm{d}\theta \tag{5.4.22}$$

are the complete elliptic integrals of the first and second kind, respectively.

For $\epsilon = 0.05, 0.025, 0.01$ and 0.005, equation (5.4.19) is discretized using trapezoidal quadrature, and we numerically approximate F by equation (5.4.18). Figure 5.4.1 plots the numerical error as a function of n for four different values of ϵ. We observe spectral convergence of the numerical error to machine precision, which is consistent with the error estimates (5.B.21). We note that the condition number of the unregularized discrete integral operator grows exponentially as n increases, as shown in Figure 5.4.2a. However, because we are considering a rigid fiber with constant radius, computing F has a regularizing effect which lessens the impact of this ill-conditioning in the final force calculation. This may be contrasted with the prolate spheroid, where, as we will see in Section 5.4.2, the conditioning does have a noticeable effect on the error. Nevertheless, we note that by setting $\eta > 1$ we can improve the condition number of the linear system (see Figure 5.4.2b). We also note that there is a $1/\epsilon$ dependence on n for a given accuracy. This can be circumvented by using a special quadrature method that takes into account the kernel (see remark 5.3).

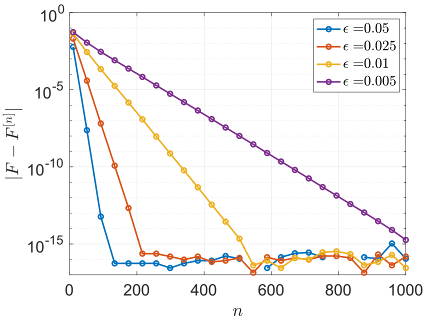
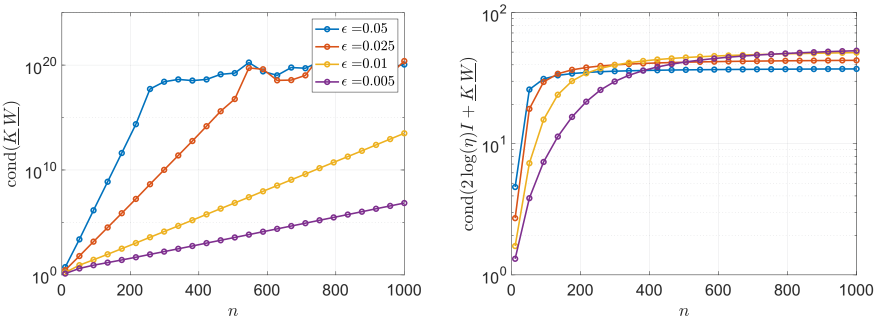


Figure 5.4.1: The approximate drag force $F^{[n]}$ on a thin ring translating broadwise with unit velocity converges with spectral accuracy to the true force F.

Prolate spheroid with artificial fluid velocity field

We next use (5.4.18) to compute the drag force for a stationary prolate spheroid immersed in an artificial fluid velocity field. The particle centerline is aligned in the z-direction, parameterized by $\boldsymbol{X}(s) = (0, 0, s)^T$, $s \in [-1, 1]$. The fluid velocity field $\mathbf{u}(s) = (u(s), 0, 0)^T$ is designed such that $\mathbf{f}(s) = (f^x(s), 0, 0)^T$ is a known analytic function. We choose this function to be a Gaussian $f^x(s) = \exp\left(-\frac{s^2}{\epsilon^2}\right)$ such that the force decays to zero at the fiber endpoints and use high order Gauss-Lobatto quadrature for the discretization of the integral operator. Denote the set of n quadrature nodes by $\{s_i\}_{i=1}^n$. Inserting the above expression for $f^x(s)$ into our model (5.3.10), the fluid velocity at s_i is found by solving the



(a) Unregularized ($\eta = 1$). **(b)** Regularized ($\eta = 1.5$).

Figure 5.4.2: The condition numbers associated with the discretized versions of the unregularized ($\eta = 1$) and regularized ($\eta = 1.5$) slender body models for calculating the force on a thin ring. Note the change in scale between the two figures.

integral

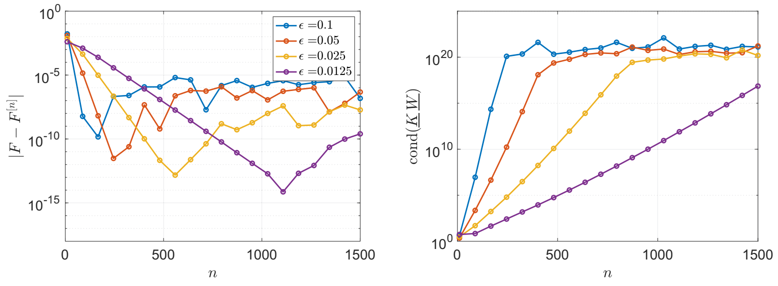
$$u(s_i) = \frac{-1}{8\pi}\left(2\log(\eta)\exp\left(-\frac{s_i^2}{\epsilon^2}\right) + \int_{-1}^{1}\frac{\epsilon^2 r\left(s_i\right)^2 + \frac{1}{2}\epsilon^2 r\left(s'\right)^2 + \left(s_i - s'\right)^2}{\left(\epsilon^2 r\left(s_i\right)^2 + \left(s_i - s'\right)^2\right)^{3/2}}\exp\left(-\frac{s'^2}{\epsilon^2}\right)ds'\right) \tag{5.4.23}$$

where the ellipsoidal radius function is given by equation (5.1.2). We also take the viscosity $\mu = 1$. To solve for $u(s_i)$ for $i = 1,...,n$, the integral in equation (5.4.23) is evaluated to machine precision using MATLAB's built-in `integral` function, which uses adaptive quadrature. For this fluid velocity field, the total force $\boldsymbol{F} = (F, 0, 0)^T$ on the ellipsoid is found by

$$F = \int_{-1}^{1}\exp\left(-\frac{s^2}{\epsilon^2}\right)ds = \sqrt{\pi}\,\epsilon\,\mathrm{erf}\left(\frac{1}{\epsilon}\right). \tag{5.4.24}$$

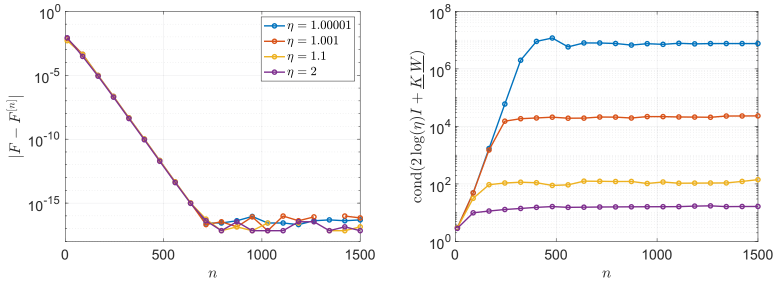
We compute numerical approximations to F using equation (5.4.18) for four choices of ϵ. We initially set $\eta = 1$ and compute these numerical approximations for the non-regularized, first-kind equation. The numerical errors are presented in Figure 5.4.3a. We see that the error converges spectrally up to a certain point where the method begins to diverge due to numerical instabilities and poor conditioning of the discrete integral operator, which is plotted in Figure 5.4.3b.

However, by choosing $\eta > 1$, we can amend the condition number and therefore improve the accuracy of the numerical solution. In Figure 5.4.4, we fix $\epsilon = 0.025$ and calculate the numerical errors for four choices of η. We see from Figure 5.4.4a that the numerical error converges spectrally to machine precision for all such choices of η. Furthermore, we observe from Figure 5.4.4b that the condition number of the discrete integral operator is bounded by a value that becomes smaller for larger η. We note that in practice, the modeling error is much larger than machine precision as we will see in section 5.5.2.



(a) **(b)**

Figure 5.4.3: The errors (a) and condition numbers (b) associated with the unregularized ($\eta = 1$) numerical method for the calculation of the force on a prolate spheroid for different values of ϵ.



(a) **(b)**

Figure 5.4.4: The errors (a) and condition numbers (b) associated with the regularized numerical method for the calculation of the force on a prolate spheroid for $\epsilon = 0.025$. Similar results are observed for other values of ϵ.

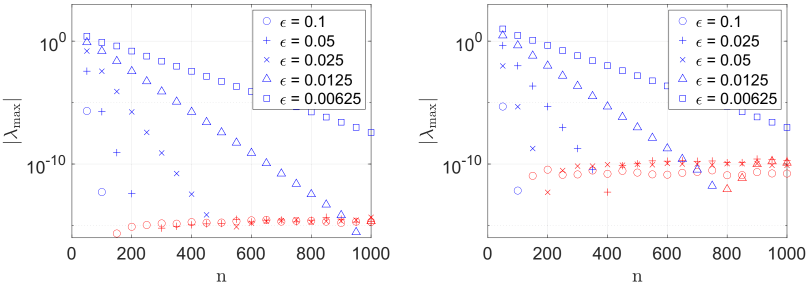
5.4.3 Spectrum of the slender body operator in different geometries

One important unresolved question about the slender body model (5.2.1) is the effect of different geometries, including curvature, endpoints, and non-uniform fiber radius, on the spectrum of the integral operator. The main difficulty is that the integral kernel (5.2.2),(5.2.3) is only well defined along the centerline of the fiber. Since the kernel is so dependent on the shape of the fiber centerline, it is difficult to prove general properties for it. Although we cannot analytically determine the spectrum of the continuous operator in general, we *can* determine the eigenvalues of the discrete operator $(2\log(\eta)I + \underline{K}\,\underline{W})$ (5.4.6).

We consider first the unregularized version $\eta = 1$, recalling that in the straight-but-periodic geometry of Section 5.3.1, the continuous operator was provably negative definite. Ideally we would like to see evidence that this negative definiteness persists in general geometries, as this would be the physically correct behavior and also would agree with the underlying slender body PDE operator (5.3.2).

We begin by calculating the eigenvalues $\{\lambda_i\}_{i=1}^{3n}$ of $\underline{K}\,\underline{W}$ for the thin ring. Letting $\lambda_{\max} = \max_i(\lambda_i)$, in Figure 5.4.5a we plot $\lambda_{\max}$ versus n for five different values of ϵ. Note that for very large n relative to ϵ^{-1} (roughly $n = O(\epsilon^{-2})$), we begin to see numerical error resulting in very small positive eigenvalues of $\underline{K}\,\underline{W}$ (denoted by red markers). However, the magnitude of these positive eigenvalues are on the order of machine precision and may be attributed to round-off errors.

We next consider the effects of endpoints and a non-uniform radius by calculating the eigenvalues of $\underline{K}\,\underline{W}$ for a slender prolate spheroid (5.1.2), keeping in mind the above level of numerical error. In Figure 5.4.5b we again plot $\lambda_{\max}$ versus n for four different values of ϵ. Again for $n = O(\epsilon^{-2})$ we begin to see small positive eigenvalues which are significantly larger than for the thin ring (around $O(10^{-10})$). However, the magnitude of the positive eigenvalues is still very small and bounded as n increases. It is not clear whether this is a numerical artifact or an actual eigenvalue crossing 0 for the continuous operator. At any rate, the non-regularized operator would never actually be used for simulations with such large n because the condition number of $\underline{K}\,\underline{W}$ is prohibitive (see Figure 5.4.3b). It appears that a very reasonable choice of regularization parameter η will ensure that none of these near-zero eigenvalues actually cross zero.

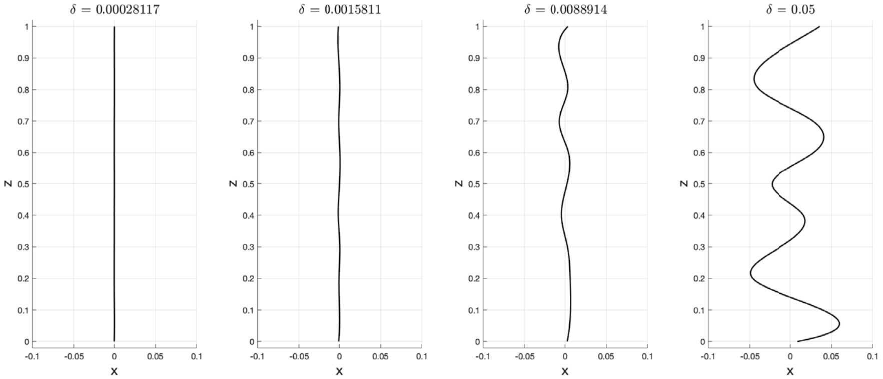


(a) Thin ring

(b) Spheroid

Figure 5.4.5: Magnitude of the maximum eigenvalue of the non-regularized discrete slender body operator $\underline{K}\,\underline{W}$. Blue markers mean $\lambda_{\max} < 0$ while red markers mean that $\lambda_{\max} > 0$.

As a final test, we calculate the spectrum of $\underline{K}\,\underline{W}$ for randomly but systematically generated curvy fibers with complicated shapes (Figure 5.4.6). Here the magnitude of the fiber's deviation from a straight line is controlled by a small parameter $\delta \geq 0$. The fiber shapes are generated by interpolating m points $(x_i, y_i, z_i) \in \mathbb{R}^3$, $i = 1, ..., m$, with cubic splines. Here $z_i = (i-1)\frac{2L}{m}$ while $x_i, y_i \in [-\delta, \delta]$ are given by a random number generator and are of size at most δ. Setting $\delta = 0$ corresponds to a straight fiber. Examples of the fiber centerline for $m = 10$ and four different values of δ are given in Figure 5.4.6.



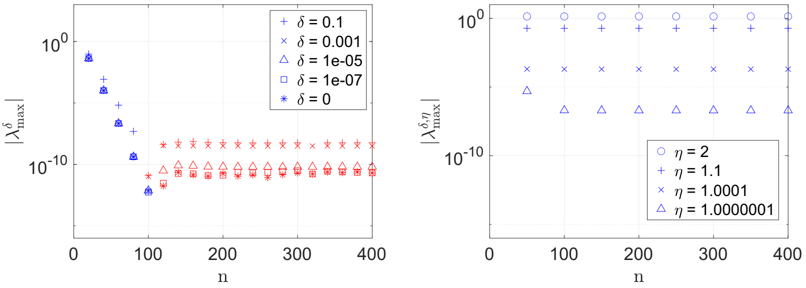
(a) **(b)** **(c)** **(d)**

Figure 5.4.6: The centerlines of four curved fiber shapes.

We fix $\epsilon = 0.1$ and use the spheroidal radius function (5.1.2). Taking $m = 10$, we generate 6 different curvy fibers for different magnitudes $\delta \in [0, \frac{1}{10}]$. For each fiber we compute the spectrum $\{\lambda_i^\delta\}_{i=1}^n$ of its corresponding (non-regularized) integral operator $\underline{K}\,\underline{W}$. We plot the most positive eigenvalue $\lambda_{\max}^\delta = \max_i(\lambda_i^\delta)$ for each fiber in Figure 5.4.7a. For each value of δ we note that there is an eigenvalue crossing zero when $n = O(\epsilon^{-2})$. As δ increases and the magnitude of the curviness of the fiber increases, we can note a slight increase in the magnitude of the largest positive eigenvalue, but $\lambda_{\max}^\delta$ is still small – roughly $O(10^{-8})$. Again, we can be assured to have a negative spectrum bounded away from 0 by a reasonable choice of regularization $\eta > 1$. This effect is displayed in Figure 5.4.7b, which shows the maximum eigenvalue $\lambda_{\max}^{\delta,\eta}$ of the now *regularized* discrete integral operator $(2\log(\eta)I + \underline{K}\,\underline{W})$ for a fixed value of ϵ and δ and varying values of η. We see here that for all choices of $\eta > 1$ in this range, the spectrum of $(2\log(\eta)I + \underline{K}\,\underline{W})$ remains negative definite.

5.5 Dynamics of curved rigid fibers

We next use the slender body model (5.2.1) and the discretization procedure of Section 5.4 to simulate the dynamics of curved rigid fibers in Stokes flow.



(a) **(b)**

Figure 5.4.7: The magnitude $\lambda^{\delta,\eta}_{\max}$ of the maximum eigenvalues for the unregularized (a) and regularized (b) discrete integral operators for the curved fibers. For (b) we fix $\epsilon = 0.1$ and $\delta = 0.001$ and consider different regularizations η. The color blue denotes a negative maximum eigenvalue and red denotes a positive maximum eigenvalue.

After outlining the dynamical equations, we validate the model against known dynamical models for a slender prolate spheroid. Finally, we compare the rotational dynamics of randomly curved fibers as in Figure 5.4.6 to straight fibers.

5.5.1 Dynamical equations

Assuming that the particle to fluid density ratio is large $\rho_p/\rho_f \gg 1$, such as in gas-solid fiber suspensions [15, 27, 30, 38], the dynamics of the slender body are governed by the following rigid body equations. The angular momentum $\boldsymbol{m}$ of a rigid particle with torque $\boldsymbol{T}(t)$ is found by solving

$$\dot{\boldsymbol{m}} = \boldsymbol{m} \times \boldsymbol{\omega} + \boldsymbol{T}, \tag{5.5.1}$$

where $\boldsymbol{\omega} = J^{-1}\boldsymbol{m}$ for moment of inertia tensor J. Each of these quantities are defined in a reference frame whose axes are co-rotating and co-translating with the fiber. The fiber orientation (with respect to a fixed inertial reference frame) is specified using Euler parameters $q \in \mathbb{R}^4$ which satisfy the constraint $||q||_2 = 1$ and are determined by solving the ODE

$$\dot{q} = \frac{1}{2} q\, w, \tag{5.5.2}$$

where $w = (0, \boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^4$. Here, $q\,w$ is the Hamilton product of two quaternions [17]. That is, by letting $q = (q_0, \boldsymbol{q})$ and $r = (r_0, \boldsymbol{r})$ denote quaternions for $q_0, r_0 \in \mathbb{R}$ and $\boldsymbol{q}, \boldsymbol{r} \in \mathbb{R}^3$, then their Hamilton product is given by

$$q\,r = (q_0\,r_0 - \boldsymbol{q}\cdot\boldsymbol{r},\; q_0\boldsymbol{r} + r_0\boldsymbol{q} + \boldsymbol{q}\times\boldsymbol{r}). \tag{5.5.3}$$

The translational dynamics are given by Newton's second law

$$\dot{\boldsymbol{p}} = \boldsymbol{F}, \tag{5.5.4}$$

where $\boldsymbol{p} = \boldsymbol{v}m$ is the inertial frame linear momentum for a fiber of mass m. The position of the fiber center of mass is found by solving

$$\dot{\mathbf{x}} = \boldsymbol{v}. \tag{5.5.5}$$

The ODEs (5.5.1) - (5.5.5) are integrated using the second order Strang splitting method of (Ref. [50]).
Recall the equations (5.4.18) for $\boldsymbol{F}^{[n]}$ and $\boldsymbol{T}^{[n]}$. Since $\boldsymbol{F}^{[n]}$ and $\boldsymbol{T}^{[n]}$ depend linearly on the linear and angular momenta $\mathbf{p}$ and $\boldsymbol{m}$, we may update them according to the linear equation

$$\begin{pmatrix} \boldsymbol{F}^{[n]} \\ \boldsymbol{T}^{[n]} \end{pmatrix} = A \begin{pmatrix} \mathbf{p} \\ \boldsymbol{m} \end{pmatrix} + \boldsymbol{b}, \tag{5.5.6}$$

where A is a negative definite dissipation matrix and $\boldsymbol{b}$ is due to the background fluid velocity and is independent of $\mathbf{p}$ and $\boldsymbol{m}$. We have that

$$A = \begin{pmatrix} \Phi\left(\vec{1} \otimes (I/m)\right), & \Phi\left(-\underline{X}(\vec{1} \otimes J^{-1})\right) \\ \Psi\left(\vec{1} \otimes (I/m)\right), & \Psi\left(-\underline{X}(\vec{1} \otimes J^{-1})\right) \end{pmatrix} \quad \text{and} \quad \boldsymbol{b} = -\begin{pmatrix} \Phi\underline{\mathbf{u}} \\ \Psi\underline{\mathbf{u}} \end{pmatrix}, \tag{5.5.7}$$

where m and J are the filament mass and moment of inertia tensor, respectively. We have also introduced the vector $\underline{\mathbf{u}} = (\mathbf{u}_0(\boldsymbol{X}(s_1))^T, ..., \mathbf{u}_0(\boldsymbol{X}(s_n))^T)^T$ containing the background fluid velocities at the location of the quadrature nodes along the centerline.

Overview and cost of algorithm

The algorithm used to compute the dynamics of a slender fiber is as follows:

1. Define particle geometry $\boldsymbol{X}(s)$, ϵ, regularization parameter η and discretization n.
2. Choose a quadrature rule and compute the matrices $\underline{W}$ and $\underline{K}$.
3. Compute the matrices Φ, Ψ and A from equations (5.4.16), (5.4.17) and (5.5.7).
4. Time loop: for $t = 0, \Delta t, ..., m\Delta t$

   a) Compute $\boldsymbol{F}^{[n]}$ and $\boldsymbol{T}^{[n]}$ using equation (5.5.6)

   b) Numerically integrate the ODEs (5.5.1) - (5.5.5) .

For step (2), we use the trapezoidal quadrature rule for closed fibers (i.e., a periodic integration interval) or Gauss-Lobatto quadrature rule for fibers with open ends. For step (4b), we use a splitting method [50]. We note that for simulations where the fluid velocity field is calculated from a direct numerical simulation of the Navier-Stokes equations, the fluid field needs to be approximated onto the centerline of the particle using an interpolation method [51].

The above algorithm exploits the rigidity of the fiber by using the fact that A, Φ and Ψ are constant in time and therefore can be computed outside of the time loop. The calculation of these matrices, which involves solving a linear system, is the most costly operation in the algorithm but only needs to be done once. If, for example, Gaussian elimination is used, this step has complexity of $O(n^3)$. Within the time loop, however, the most costly operation is the calculation of $\boldsymbol{F}^{[n]}$ and $\boldsymbol{T}^{[n]}$, which involves only $3 \times 3n$ by $3n \times 1$ matrix-vector products, which has $O(n)$ complexity. We assume that the cost of numerically integrating the ODEs is negligible compared to this. For a single fiber, the total complexity of the algorithm is therefore $O(n^3 + nm)$, where m is the total number of time steps used in the simulation. Hence, for simulations where many time steps are needed, the algorithm scales by $O(n)$. We remark that for problems where the background flow is zero, the cost of computing $\boldsymbol{F}^{[n]}$ and $\boldsymbol{T}^{[n]}$ is independent of n (after A has been computed) and therefore is $O(1)$. This is relevant, for example, when simulating fibers sedimenting in a still fluid under the influence of gravity [37].

5.5.2 Numerical validation of model dynamics

Dissipation matrix of a prolate spheroid

Here we compare our model and numerical method with accurate closed form expressions for the force and torque given by Brenner [6] and Jeffery [23]. These expressions are valid for an ellipsoid when the fluid Jacobian is approximately constant throughout the volume of the particle. When the flow is linear, these terms are essentially exact and therefore serve as a good reference model against which to validate our model.

The purpose of this numerical experiment is therefore twofold. Firstly, we aim to show that our model converges to the reference model as $\epsilon \to 0$. This is primarily to validate the accuracy of the model. However, the numerical approximation of the force and torques also introduces a numerical error that is related to the discretization parameter n. Clearly, taking n too small means that we will not exploit the accuracy of the model to its entirety. On the other hand, it is unwise to take n as large as possible as this will incur unnecessary computational costs that go to minimizing numerical error beyond the accuracy of the model. So the second question we address here is what is an ideal choice of discretization parameter to use such that the numerical error is roughly the same as the modeling error.

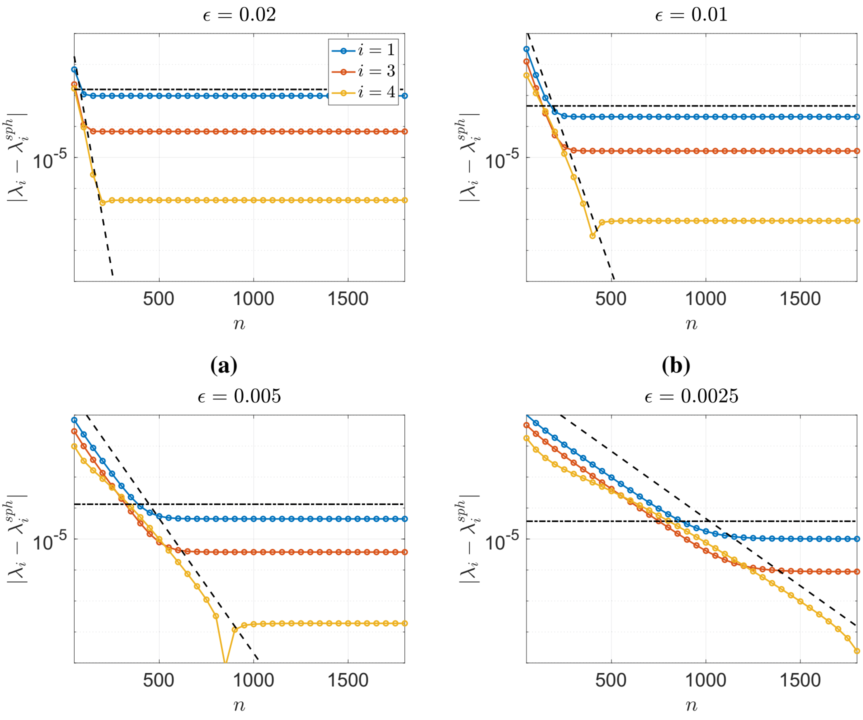
Using $\eta = 1+\epsilon^2$, the dissipation matrix for our slender body model A is numerically approximated by equation (5.5.7). The reference dissipation matrix A_{sph} is found using the closed form expressions from Jeffery and Brenner, which are given in Appendix 5.C. Denote the six eigenvalues of A and A_{sph}, by λ_i and λ_i^{sph}, respectively. Note that due to symmetry of the spheroid, $\lambda_1 = \lambda_2$ and $\lambda_4 = \lambda_5$ and similarly for the eigenvalues of A_{sph}. Furthermore, the slender body model is essentially a one dimensional filament and therefore $\lambda_6 = 0$ meaning that spinning motion about the centerline doesn't dissipate. This is in contrast to the Jeffrey term, which does dissipate spinning motion. We remark that this phenomenon only occurs in the case where the centerline is perfectly straight. Hence for curved fiber geometries where the application of the slender body is most useful, this nonphysical phenomenon is not observed. Note that for this geometry the dissipation matrices are diagonal and therefore the eigenvalues are directly proportional to the calculation of $\boldsymbol{F}^{[n]}$ and $\boldsymbol{T}^{[n]}$ in zero background flow.

The eigenvalues of A are calculated using equation (5.5.7) after discretizing equation (5.4.3) on the Gauss-Lobatto nodes. The values $|\lambda_i - \lambda_i^{sph}|$ for $i = 1,3,4$ are plotted in Figure 5.5.1 as a function of the discretization parameter n. We see that λ_i converges exponentially to a point near λ_i^{sph}, which is likely due to the slender body modelling error. As ϵ decreases, we make two observations. First, for large n the rate at which λ_i converges to λ_i^{sph} is approximately $-\epsilon^2\eta^2\log(\epsilon\eta)$, as seen by the horizontal dash-dot lines. Second, as ϵ decreases, the convergence rate slows down and one must use a larger value of n to reach the most accurate solution. This means that one must pay careful attention to the choice of n when taking ϵ to be very small. In fact, we observe empirically that the convergence rate is approximately bounded by $e^{-4\epsilon n}$. Motivated by this, we will take n in future experiments to be approximately the intersection of these two lines, that is

$$n \approx -\frac{\log(-\epsilon^2\eta^2\log(\epsilon\eta))}{4\epsilon}. \tag{5.5.8}$$

Prolate spheroids rotating in shear flow

Now we calculate the dynamics of a prolate spheroid in shear flow $\mathbf{u} = (z,0,0)^T$ using our model and compare it with that of the accurate Jeffrey model. The fiber is initially aligned at rest in the z-direction and its rotational dynamics are calculated by integrating equation (5.5.1) on the interval $t \in [0,100]$ using the splitting method of (Ref. [50]) with a small step size of $h = 0.01$. The



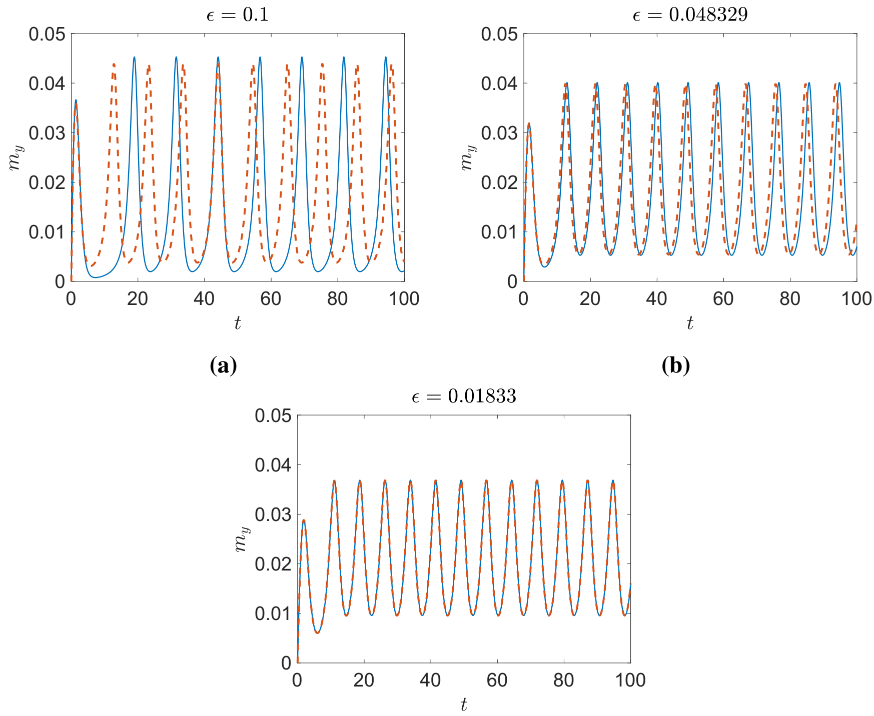
(a) **(b)**

(c) **(d)**

Figure 5.5.1: The difference in the dissipation matrix eigenvalues $|\lambda_i - \lambda_i^{sph}|$, $i = 1, 3, 4$ as a function of n for three different values of ϵ. The black dashed lines are $e^{-4\epsilon n}$ and the horizontal dash-dot lines are $-\epsilon^2\eta^2 \log(\epsilon\eta)$.

simulation was repeated with $h = 0.05$ with no significant changes to the results and it is therefore concluded that time integration errors are negligible. We repeat the experiment for 20 values of ϵ logarithmically spaced in the interval $[0.1, 0.001]$ and choose n using equation (5.5.8) and $\eta = 1 + \epsilon^2$. As the spheroids are axisymmetric, they only experience a torque about their y axis, hence all of other angular momentum components are zero (to machine precision). Three examples of the rotational dynamics are shown in Figure 5.5.2. It is seen here that as ϵ becomes smaller, the dynamics more closely resemble the Jeffery model.

The relative difference between the angular momenta of the Jeffery and slender body solutions are calculated and averaged over the simulation. This average relative error is then plotted against the corresponding value of ϵ in Figure 5.5.3. We see that the average relative error decreases with ϵ. It is observed that in the region $0.01 < \epsilon < 0.1$ the error converges at a faster rate than in the region $0.001 < \epsilon < 0.01$. This could be partially explained by the fact that wider



(a)

(b)

(c)

Figure 5.5.2: The y component of a spheroid rotating in shear flow for three different values of ϵ. The solid line is the our slender body model and the dashed line is due to Jeffery.

particles (larger ϵ) experience a greater resistive force as seen by the regions where m_y nearly reaches zero. This means that the particle spends more time in the shear plane where the fluid velocity is zero and hence the slender body model does not experience a large torque. However, the fluid gradient is non-zero in this orientation and therefore the Jeffery model, which depends only on the fluid gradient, still experiences a constant torque. This means that compared to the Jeffery model, thicker fibers will see a greater difference in the torque term when the fiber is aligned in the shear plane than thinner fibers.

5.5.3 Dynamics of randomly curvy fibers

Understanding how different shaped particles rotate in shear flow is an important step in understanding their dynamics in more complex flows [52]. Here we simulate the dynamics of the randomly curvy fibers of Figure 5.4.6 as they rotate in shear flow. In particular, we show how the rotational variables deviate from a straight fiber as δ becomes larger.

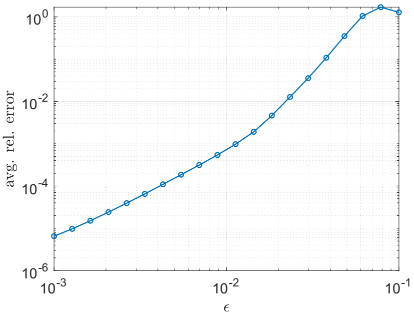


Figure 5.5.3: The relative difference in m_y between the slender body and Jeffery solutions averaged over the interval $[0,100]$.

We generate 100 different fiber shapes with $m = 10$ using 10 different values of δ logarithmically spaced in the interval $[5\times10^{-5}, 5\times10^{-2}]$. The 100 fibers are placed in shear flow $\mathbf{u} = (z,0,0)^T$ and their rotational dynamics are calculated on the interval $t \in [0,100]$. The moment of inertia tensor is approximated by placing point masses along the centerline and using the formula

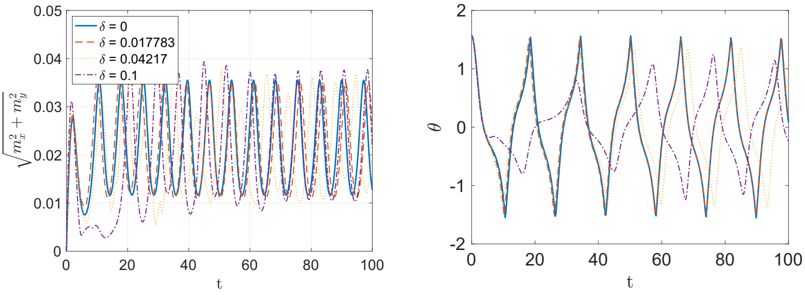
$$J_{i,i} = \sum_{j=1}^{k} m_j (X_i(s_j) - c_i)^2, \quad \text{for} \quad i = 1,...,3 \tag{5.5.9}$$

where $X_i(s_j)$ is the ith component of the centerline function at the point s_j on the centerline and c_i is the ith component of the fiber center of mass. We weight m_j by the cross sectional radius and use a very large value for k, e.g., $k = 10^4$. Here we take $\epsilon = 0.01$ and use the spheroidal radius function (5.1.2) along with $\eta = 1 + \epsilon^2$.

Figure 5.5.4a shows the angular momentum $\boldsymbol{m}$ of three fibers compared to the $\delta = 0$ case. As the $\delta = 0$ fiber is perfectly straight, it does not exhibit spinning motion and its angular momentum is purely in the m_y component. This is in contrast to the fibers with a non-zero value of δ, in which case some of the momentum is transferred to m_x. We therefore compare the value $\sqrt{m_x^2 + m_y^2}$ between the fibers to account for this. We see here that the $\delta = 0.017783$ solution is visually very similar to the $\delta = 0$ solution. We notice a significant difference between the other two solutions. Figure 5.5.4b shows the angle θ between the z-axis of the particle reference frame (that is, a frame that is rotating with the fiber) and the x-axis of a fixed inertial reference frame. As the $\delta \neq 0$ fibers are not symmetric, they slowly rotate out of the xz-plane and therefore after a long time, we see much more significant discrepancies in θ.

To quantify the effect that δ has on the angular momentum, we calculate the difference in the angular momentum Δm by subtracting off the $\delta = 0$ solution and averaging over the time interval $t \in [92,100]$, which corresponds to roughly one period of rotation. This value is averaged over all the fibers with similar values of δ and is expressed as a percentage of the $\delta = 0$ solution, which we denote by $\%\Delta m$. The results are plotted in Figure 5.5.5a. We notice that the $\%\Delta m$ is linearly proportional to δ. We observe that at the end of the simulation the $\delta = 0.0003$ fibers correspond to roughly 1% discrepancy in angular momentum and $\delta = 0.0015$ corresponds to roughly 7.5% discrepancy.

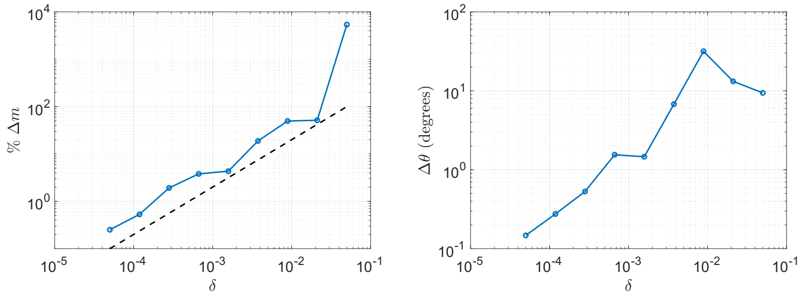
The difference in θ after one rotation as a function of δ is displayed in Figure 5.5.5b. The $\delta = 0.0003$ solution corresponds to about a 3° difference in θ and the $\delta = 0.0015$ solution corresponds to about an 8° difference.



(a)

(b)

Figure 5.5.4: The rotational variables of four fibers with different values of δ. Figure (a) shows the angular momentum and Figure (b) is the angle between the fiber's long axis and the x-axis of the inertial frame.



(a)

(b)

Figure 5.5.5: Figure (a) shows the difference in angular momentum Δm between the curved fibers and the $\delta = 0$ solution after 100 time units and averaged over all the fibers with similar δ. The black dashed line is $O(\delta)$. Figure (b) shows the discrepancy $\Delta\theta$ in the angle between the centerline and the x-axis after roughly one rotation.

5.6 Conclusions

We have developed an integral model for the motion of a thin filament in a viscous fluid based on nonlocal slender body theory. The model relies on standard singular Stokeslets and doublets but makes use of the fiber integrity condition – the near-cancellation of angular-dependent terms along the fiber surface – in a novel way to yield an integral expression for the fiber velocity with a smooth kernel which retains dependence on the (possibly varying) fiber radius in a natural way. We include a systematic way of comparing mapping properties of different models using the simplified geometry of a straight-but-periodic filament. In this simple geometry, we can show that our integral operator is negative definite and compares favorably to other models, and we expect similar high wavenumber behavior for curved filaments with constant radius. It is less clear how a non-constant radius affects the spectrum; however, numerical tests indicate that the discretized integral operator is very close to negative definite. Nevertheless, to ensure invertibility, we develop an asymptotically consistent regularization to convert the first-kind Fredholm integral equation for the force density along the fiber into a second-kind equation and show that this second-kind regularization improves the stability and conditioning of the discretized equation. We numerically solve the integral equation using the Nyström method [2] and show how constraining the fiber motion to be rigid can be exploited for fast computation of fiber dynamics. We validate the method and model against the prolate spheroid model of Jeffery [23], and apply the method to study the rotational deviation of randomly curved rigid fibers from straight fibers.

While the fibers considered here are rigid, the model can also be used to simulate the dynamics of semiflexible filaments. The invertibility properties of the integral equation make it particularly well suited for handling simulations involving inextensible fibers, where an additional line tension equation must be solved at each time step [32, 55]. We may also consider the effects of different choices of radius functions on the model properties, similar to what is done in (Ref. [58]), although we note the necessity of smooth decay in our radius function near the fiber endpoints.

To build on the dynamic simulations for rigid fibers, we aim to consider the effects of fiber shape on particle deposition and aggregation. We are especially interested in more complicated background flows, including suspensions of rigid fibers in turbulence. The novel modelling approach advocated herein will enable earlier explorations based on the point-particle approach [7] to be extended to curved fibers particles.

Acknowledgments

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (No. 691070) as well as the SPIRIT project (No. 231632) under the Research Council of Norway FRIPRO funding scheme. E. Celledoni, B. Owren and B. K. Tapley would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme *Geometry, compatibility and structure preservation in computational differential equations* (2019) where part of the work on this paper was undertaken. E. Celledoni and B. Owren also acknowledge funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 860124. L. Ohm was supported by a University of Minnesota Doctoral Dissertation Fellowship and NSF Postdoctoral Research Fellowship Grant No. 2001959. The authors are listed in alphabetical order.

Bibliography

[1] L. af Klinteberg and A. H. Barnett, *Accurate quadrature of nearly singular line integrals in two and three dimensions by singularity swapping*, BIT Numerical Mathematics, (2020), pp. 1–36.

[2] K. Atkinson and W. Han, *Theoretical numerical analysis*, vol. 39, Springer, 2005.

[3] K. E. Atkinson, *An introduction to numerical analysis*, (1978).

[4] G. Batchelor, *Slender-body theory for particles of arbitrary cross-section in Stokes flow*, J. Fluid Mech., 44 (1970), pp. 419–440.

[5] E. L. Bouzarth and M. L. Minion, *Modeling slender bodies with the method of regularized stokeslets*, Journal of Computational Physics, 230 (2011), pp. 3929–3947.

[6] H. Brenner, *The stokes resistance of an arbitrary particle—iv arbitrary fields of flow*, Chemical Engineering Science, 19 (1964), pp. 703–727.

[7] N. R. Challabotla, L. Zhao, and H. I. Andersson, *On fiber behavior in turbulent vertical channel flow*, Chemical Engineering Science, 153 (2016), pp. 75–86.

[8] S. Chattopadhyay and X.-L. Wu, *The effect of long-range hydrodynamic interaction on the swimming of a single bacterium*, Biophys. J., 96 (2009), pp. 2023–2028.

[9] A. T. Chwang and T. Y.-T. Wu, *Hydromechanics of low-Reynolds-number flow. Part 2: Singularity method for Stokes flows*, J. Fluid Mech., 67 (1975), pp. 787–815.

[10] R. Cortez, *The method of regularized stokeslets*, SIAM Journal on Scientific Computing, 23 (2001), pp. 1204–1225.

[11] R. Cortez, *Regularized Stokeslet segments*, J. Comp. Phys., 375 (2018), pp. 783 – 796.

[12] R. Cortez, L. Fauci, and A. Medovikov, *The method of regularized Stokeslets in three dimensions: analysis, validation, and application to helical swimming*, Phys. Fluids, 17 (2005), p. 031504.

[13] R. Cortez and M. Nicholas, *Slender body theory for Stokes flows with regularized forces*, Commun. Appl. Math. Comput. Sci., 7 (2012), pp. 33–62.

[14] R. Cox, *The motion of long slender bodies in a viscous fluid part 1. general theory*, J. Fluid Mech., 44 (1970), pp. 791–810.

[15] A. R. Esmaeili, B. Sajadi, and M. Akbarzadeh, *Numerical simulation of ellipsoidal particles deposition in the human nasal cavity under cyclic inspiratory flow*, Journal of the Brazilian Society of Mechanical Sciences and Engineering, 42 (2020), pp. 1–13.

[16] X. Fan, N. Phan-Thien, and R. Zheng, *A direct simulation of fibre suspensions*, J. Non-Newton. Fluid Mech., 74 (1998), pp. 113–135.

[17] H. Goldstein, C. Poole, and J. Safko, *Classical mechanics*, 2002.

[18] T. Götz, *Interactions of fibers and flow: asymptotics, theory and numerics*, Doctoral dissertation, University of Kaiserslautern, 2000.

[19] K. Gustavsson and A.-K. Tornberg, *Gravity induced sedimentation of slender fibers*, Phys. Fluids, 21 (2009), p. 123301.

[20] J. Hämäläinen, S. B. Lindström, T. Hämäläinen, and H. Niskanen, *Papermaking fibre-suspension flow simulations at multiple scales*, J. Engrg. Math., 71 (2011), pp. 55–79.

[21] G. Hancock, *The self-propulsion of microscopic organisms through liquids*, Proc. R. Soc. Lond. A, 217 (1953), pp. 96–121.

[22] P. C. Hansen, *Numerical tools for analysis and solution of fredholm integral equations of the first kind*, Inverse problems, 8 (1992), p. 849.

[23] G. B. Jeffery, *The motion of ellipsoidal particles immersed in a viscous fluid*, Proc. R. Soc. Lond. A, 102 (1922), pp. 161–179.

[24] R. E. Johnson, *An improved slender-body theory for Stokes flow*, J. Fluid Mech., 99 (1980), pp. 411–431.

[25] J. B. Keller and S. I. Rubinow, *Slender-body theory for slow viscous flow*, J. Fluid Mech., 75 (1976), pp. 705–714.

[26] R. Kress, V. Maz'ya, and V. Kozlov, *Linear integral equations*, vol. 82, Springer, 1989.

[27] S. Kuperman, L. Sabban, and R. van Hout, *Inertial effects on the dynamics of rigid heavy fibers in isotropic turbulence*, Physical Review Fluids, 4 (2019), p. 064301.

[28] E. Lauga and T. R. Powers, *The hydrodynamics of swimming microorganisms*, Rep. Progr. Phys., 72 (2009), p. 096601.

[29] J. Lighthill, *Flagellar hydrodynamics*, SIAM review, 18 (1976), pp. 161–230.

[30] C. Marchioli, M. Fantoni, and A. Soldati, *Orientation, distribution, and deposition of elongated, inertial fibers in turbulent channel flow*, Physics of fluids, 22 (2010), p. 033301.

[31] J. Martin, A. Lusher, R. C. Thompson, and A. Morley, *The deposition and accumulation of microplastics in marine sediments and bottom water from the irish continental shelf*, Sci. Rep, 7 (2017), p. 10772.

[32] O. Maxian, A. Mogilner, and A. Donev, *Integral-based spectral method for inextensible slender fibers in stokes flow*, Physical Review Fluids, 6 (2021), p. 014102.

[33] Y. Mori and L. Ohm, *An error bound for the slender body approximation of a thin, rigid fiber sedimenting in Stokes flow*, Res. Math. Sci., 7 (2020).

[34] Y. Mori and L. Ohm, *Accuracy of slender body theory in approximating force exerted by thin fiber on viscous fluid*, Stud. Appl. Math., published online, (2021).

[35] Y. Mori, L. Ohm, and D. Spirn, *Theoretical justification and error analysis for slender body theory*, Comm. Pure Appl. Math., 73 (2020), pp. 1245–1314.

[36] Y. Mori, L. Ohm, and D. Spirn, *Theoretical justification and error analysis for slender body theory with free ends*, Arch. Ration. Mech. Anal., 235 (2020), pp. 1905–1978.

[37] R. Newsom and C. Bruce, *The dynamics of fibrous aerosols in a quiescent atmosphere*, Physics of Fluids, 6 (1994), pp. 521–530.

[38] D. O. Njobuenwu and M. Fairweather, *Simulation of inertial fibre orientation in turbulent flow*, Physics of Fluids, 28 (2016), p. 063307.

[39] A. OBERBECK, *Uber stationare flussigkeitsbewegungen mit berucksichtigung der inner reibung*, J. reine angew. Math., 81 (1876), pp. 62–80.

[40] L. OHM, B. K. TAPLEY, H. I. ANDERSSON, E. CELLEDONI, AND B. OWREN, *A slender body model for thin rigid fibers: validation and comparisons*, Proc. of MEKiT'19, 10th Nat. Conf. on Comp. Mech., (2019).

[41] C. J. PETRIE, *The rheology of fibre suspensions*, J. Non-Newton. Fluid Mech., 87 (1999), pp. 369–402.

[42] C. POZRIKIDIS, *Boundary integral and singularity methods for linearized viscous flow*, Cambridge University Press, 1992.

[43] B. RODENBORN, C.-H. CHEN, H. L. SWINNEY, B. LIU, AND H. ZHANG, *Propulsion of microorganisms by a helical flagellum*, Proc. Natl. Acad. Sci., 110 (2013), pp. E338–E347.

[44] J. ROTNE AND S. PRAGER, *Variational treatment of hydrodynamic interaction in polymers*, The Journal of Chemical Physics, 50 (1969), pp. 4831–4837.

[45] M. J. SHELLEY AND T. UEDA, *The Stokesian hydrodynamics of flexing, stretching filaments*, Phys. D, 146 (2000), pp. 221–245.

[46] M. SHIN AND D. L. KOCH, *Rotational and translational dispersion of fibres in isotropic turbulent flows*, Journal of Fluid Mechanics, 540 (2005), p. 143.

[47] C. SIEWERT, R. KUNNEN, M. MEINKE, AND W. SCHRÖDER, *Orientation statistics and settling velocity of ellipsoids in decaying turbulence*, Atmospheric research, 142 (2014), pp. 45–56.

[48] D. J. SMITH, *A boundary element regularized stokeslet method applied to cilia-and flagella-driven flow*, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 465 (2009), pp. 3605–3626.

[49] S. E. SPAGNOLIE AND E. LAUGA, *Comparative hydrodynamics of bacterial polymorphism*, Phys. Rev. Lett., 106 (2011), p. 058103.

[50] B. TAPLEY, E. CELLEDONI, B. OWREN, AND H. I. ANDERSSON, *A novel approach to rigid spheroid models in viscous flows using operator splitting methods*, Numer. Algorithms, (2019), pp. 1–19.

[51] B. K. Tapley, H. I. Andersson, E. Celledoni, and B. Owren, *Computational methods for tracking inertial particles in discrete incompressible flows*, arXiv preprint arXiv:1907.11936, (2019).

[52] I. Thorp and J. Lister, *Motion of a non-axisymmetric particle in viscous shear flow*, (2019).

[53] A.-K. Tornberg, *Accurate evaluation of integrals in slender-body formulations for fibers in viscous flow*, arXiv preprint arXiv:2012.12585, (2020).

[54] A.-K. Tornberg and K. Gustavsson, *A numerical method for simulations of rigid fiber suspensions*, J. Comput. Phys., 215 (2006), pp. 172–196.

[55] A.-K. Tornberg and M. J. Shelley, *Simulating the dynamics and interactions of flexible fibers in Stokes flows*, J. Comput. Phys., 196 (2004), pp. 8–40.

[56] L. N. Trefethen and J. Weideman, *The exponentially convergent trapezoidal rule*, siam REVIEW, 56 (2014), pp. 385–458.

[57] S. Twomey, *On the numerical solution of fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature*, Journal of the ACM (JACM), 10 (1963), pp. 97–101.

[58] B. J. Walker, M. P. Curtis, K. Ishimoto, and E. A. Gaffney, *A regularised slender-body theory of non-uniform filaments*, Journal of Fluid Mechanics, 899 (2020), p. A3.

[59] B. J. Walker, K. Ishimoto, H. Gadêlha, and E. A. Gaffney, *Filament mechanics in a half-space via regularised stokeslet segments*, Journal of Fluid Mechanics, 879 (2019), pp. 808–833.

[60] H. Yamakawa, *Transport properties of polymer chains in dilute solution: hydrodynamic interaction*, The Journal of Chemical Physics, 53 (1970), pp. 436–443.

[61] B. Zhao, E. Lauga, and L. Koens, *Method of regularized stokeslets: Flow analysis and improvement of convergence*, Physical Review Fluids, 4 (2019), p. 084104.

Appendix

5.A Modified Lighthill model

Here we consider the *modified Lighthill* approach to deriving a fiber velocity approximation from classical SBT (5.3.4). This approach takes advantage of the fact that the doublet term of (5.3.5) only has an $O(1)$ contribution to the fiber velocity very close to $s' = s$, and thus can be integrated asymptotically to leave only a local term. This results in a model similar to that of Lighthill [29], which was derived via different reasoning but also includes a local doublet term and a nonlocal Stokeslet contribution (see Remark 5.4).

There are two ways to consider the nonlocal Stokeslet contribution. The first expression, which we will term Modified Lighthill 1, is given by the periodization of

$$\overline{\mathbf{u}}(z) = -\frac{1}{8\pi}\left((\mathbf{I}-\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\boldsymbol{f}(s) + \int_{-1}^{1}\left(\frac{\mathbf{I}}{(\overline{z}^2+\epsilon^2)^{1/2}} + \frac{\overline{z}^2\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{3/2}}\right)\boldsymbol{f}(z-\overline{z})\,d\overline{z}\right). \tag{5.A.1}$$

Here the local term $(\mathbf{I}-\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})$ comes from asymptotically integrating the doublet term of (5.3.4) (see estimate 3.65 of (Ref. [35]) for more detail). Note that in (5.A.1), the Stokeslet term inside the integral is equal to $\boldsymbol{f}/\epsilon$ when $\overline{z} = 0$.

For the second expression, which we will call Modified Lighthill 2, the $\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}$ component of the Stokeslet term is normalized to give the same order contribution at $\overline{z} = 0$ as in (5.3.4); namely, $(\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\boldsymbol{f}/\epsilon$. This yields the periodization of the expression

$$\overline{\mathbf{u}}(z) = -\frac{1}{8\pi}\left((\mathbf{I}-\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\boldsymbol{f}(s) + \int_{-1}^{1}\frac{\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{1/2}}\boldsymbol{f}(z-\overline{z})\,d\overline{z}\right). \tag{5.A.2}$$

Remark 5.4. The actual model proposed by Lighthill in (Ref. [29]), written in the periodic, straight setting, has the form

$$\overline{\mathbf{u}}(z) = -\frac{1}{8\pi}\left(2(\mathbf{I}-\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\boldsymbol{f}(z) + \int_{|\overline{z}|>q}\frac{\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{|\overline{z}|}\boldsymbol{f}(z-\overline{z})\,d\overline{z}\right); \quad q = \epsilon\sqrt{e}/2. \tag{5.A.3}$$

At first glance, this looks like a slightly different model from (5.A.1) and (5.A.2), due to the 2 in front of the $(\mathbf{I}-\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\boldsymbol{f}(z)$ term. However, the extra factor here is precisely due to the removal of the section $|\overline{z}| \le q$ from the integral term. Indeed, if we consider the integrand of (5.A.1), we note that

$$\begin{aligned}\int_{-q}^{q}\left(\frac{\mathbf{I}}{(\overline{z}^2+\epsilon^2)^{1/2}}+\frac{\overline{z}^2\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{3/2}}\right)\boldsymbol{f}(z-\overline{z})\,d\overline{z} &= \left(2\log(2q/\epsilon)(\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})-2\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}\right)\boldsymbol{f}(z)+O(\epsilon^2/q^2)\\ &= (\mathbf{I}-\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}})\boldsymbol{f}(z)+O(\epsilon^2/q^2)\end{aligned}$$

for q as in (5.A.3). Now, this particular choice of q is not large relative to ϵ, so the $O(\epsilon^2/q^2)$ error term is not small asymptotically. However, this is merely a heuristic and we will not be considering the expression (5.A.3) in greater depth here. Furthermore, the expressions (5.A.1) and (5.A.2) are more amenable to calculating eigenvalues.

The eigenvalues of (5.A.1) are given by

$$\lambda_k^m = \begin{cases} -\dfrac{1}{4\pi}\Big(2K_0(\pi\epsilon|k|)-\pi\epsilon|k|\,K_1(\pi\epsilon|k|)\Big), & m=z\\ -\dfrac{1}{8\pi}\Big(1+2K_0(\pi\epsilon|k|)\Big), & m=x,y.\end{cases} \tag{5.A.4}$$

Now the normal eigenvalues λ_k^x and λ_k^y are always negative. However, there is still a high wavenumber instability in the tangent direction. In particular, $\lambda_k^z=0$ when $\pi\epsilon|k|\approx 1.55265$, and becomes positive at higher wavenumbers (see Figure 5.3.1). Thus the instability issue is not fully resolved by expanding only the doublet term of (5.3.4).

For Modified Lighthill 2, the eigenvalues of (5.A.2) are given by

$$\lambda_k^m = \begin{cases} -\dfrac{1}{2\pi}K_0(\pi\epsilon|k|), & m=z\\ -\dfrac{1}{8\pi}\Big(1+2K_0(\pi\epsilon|k|)\Big), & m=x,y.\end{cases} \tag{5.A.5}$$

Here the eigenvalues λ_k^x and λ_k^y in the normal directions are identical to (5.A.4), but the tangential eigenvalues λ_k^z are very different. In fact, they are too different: Recall that near $t=0$, the modified Bessel functions $K_0(t)$ and $K_1(t)$ satisfy

$$K_0(t)=-\log(t/2)-\gamma+O(t^2);\quad tK_1(t)=1+O(t^2). \tag{5.A.6}$$

Therefore, at low wavenumber ($k=O(1)$), the tangential eigenvalues of Modified Lighthill 2 (5.A.2) look like

$$\lambda_k^z=\frac{1}{2\pi}(\log(\pi\epsilon|k|/2)+\gamma)+O(\epsilon^2k^2).$$

This does not agree with the low wavenumber behavior of the slender body PDE (5.3.9) (see Figure 5.3.1). It appears that the normalization in Modified Lighthill 2 (5.A.2) results in the wrong model.

For the sake of completeness, we also consider a modification of our model (5.3.5) in which the $\overline{\boldsymbol{X}\boldsymbol{X}}^{\mathrm{T}}$ terms are normalized as in Modified Lighthill 2 (5.A.2) to yield a nonzero contribution to the fiber velocity when $s = s'$. In the case of the periodic straight centerline, the modified version of our model becomes the periodization of

$$\overline{\mathbf{u}}(z) = -\frac{1}{8\pi}\int_{-1}^{1}\left(\frac{\mathbf{I}+\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{1/2}} + \frac{\epsilon^2}{2}\frac{\mathbf{I}-3\boldsymbol{e}_z\boldsymbol{e}_z^{\mathrm{T}}}{(\overline{z}^2+\epsilon^2)^{3/2}}\right)\boldsymbol{f}(z-\overline{z})\,d\overline{z}. \tag{5.A.7}$$

The eigenvalues of (5.A.7) are given by

$$\lambda_k^m = \begin{cases} -\dfrac{1}{4\pi}\Big(2K_0(\pi\epsilon|k|) - \pi\epsilon|k|\,K_1(\pi\epsilon|k|)\Big), & m = z \\ -\dfrac{1}{8\pi}\Big(2K_0(\pi\epsilon|k|) + \pi\epsilon|k|\,K_1(\pi\epsilon|k|)\Big), & m = x, y. \end{cases} \tag{5.A.8}$$

Now, the eigenvalues λ_k^x and λ_k^y in the directions normal to the fiber are unchanged from our original expression (5.3.12). However, the tangent eigenvalues λ_k^z are now given by the same expression as Modified Lighthill 1 (5.A.4), which we recall exhibits a high wavenumber instability (Figure 5.3.1).

5.B Convergence and error bounds of numerical method

We are interested in obtaining an estimate for the error when approximating (5.4.4) by its discrete approximation (5.4.10), which we denote by

$$\mathbf{d}^{[n]} = \phi_M(\mathbf{f}) - \phi_M^{[n]} = \int_{-L}^{L} M(s)\mathbf{f}(s)ds - \sum_{j=1}^{n} w_j M(s_j)\mathbf{f}_j^{[n]}. \tag{5.B.1}$$

This error will depend on the error committed in the numerical approximation of (5.4.3) by the solution $\underline{\mathbf{f}}^{[n]}$ of (5.4.6). For this reason, we first analyze the convergence of Nyström's method (see Ref. [2, Chapt. 12.4]) in using (5.4.6) to approximate the solution of (5.4.3). At each quadrature node, we define the error of this approximation as

$$\mathbf{e}_i^{[n]} := \mathbf{f}(s_i) - \mathbf{f}_i^{[n]}, \quad \text{for} \quad i = 1, \ldots, n, \tag{5.B.2}$$

and let $\underline{\mathbf{e}}^{[n]} := ((\mathbf{e}_1^{[n]})^T, \ldots, (\mathbf{e}_n^{[n]})^T)^T$ denote the error vector. We want to show that $\|\underline{\mathbf{e}}^{[n]}\|_\infty \to 0$ as $n \to \infty$. Let $\underline{\mathbf{f}} := (\mathbf{f}(s_1)^T, \ldots, \mathbf{f}(s_n)^T)^T$ and define $\underline{\tau}^{[n]} := (\tau_1^T, \ldots, \tau_n^T)^T$ with components

$$\tau_i := \mathbf{y}(s_i) - \alpha\mathbf{f}(s_i) - \sum_{j}^{n} K_{i,j} w_j \mathbf{f}(s_j), \tag{5.B.3}$$

the truncation error for the discrete second kind equation (5.4.6) – i.e. the residual obtained replacing $\underline{\mathbf{f}}^{[n]}$ by $\underline{\mathbf{f}}$ in (5.4.6). We obtain

$$\left(\alpha I+\underline{K}\,\underline{W}\right)\underline{\mathbf{f}}=\underline{\mathbf{y}}-\underline{\tau}^{[n]}. \tag{5.B.4}$$

It is easily seen using (5.4.3) that

$$\tau_i=\int_{-L}^{L}K(s_i,s')\mathbf{f}(s')ds'-\sum_{j}^{n}K_{i,j}w_j\mathbf{f}(s_j), \tag{5.B.5}$$

which is simply quadrature error, and for any convergent quadrature formula we have

$$\lim_{n\to\infty}\|\tau^{[n]}\|_\infty=0. \tag{5.B.6}$$

We next bound the norm of the error $\underline{\mathbf{e}}^{[n]}$ by the norm of $\underline{\tau}^{[n]}$ to prove the convergence of the method. Subtracting (5.4.6) from (5.B.4) we obtain a linear system satisfied by $\underline{\mathbf{e}}^{[n]}$:

$$\left(\alpha I+\underline{K}\,\underline{W}\right)\underline{\mathbf{e}}^{[n]}=-\underline{\tau}^{[n]}. \tag{5.B.7}$$

From (Ref. [2, Chapt. 12.4], Theorem 12.4.4 and equation (12.4.51)), we have that for sufficiently large n, say $n\geq n^*$, the matrix $\left(\alpha I+\underline{K}\,\underline{W}\right)$ is invertible and

$$\|\left(\alpha I+\underline{K}\,\underline{W}\right)^{-1}\|_\infty\leq C_1 \qquad \forall n\geq n^*. \tag{5.B.8}$$

Thus we can conclude that

$$\|\underline{\mathbf{e}}^{[n]}\|_\infty\leq\|\left(\alpha I+\underline{K}\,\underline{W}\right)^{-1}\|_\infty\|\underline{\tau}^{[n]}\|_\infty\leq C_1\|\underline{\tau}^{[n]}\|_\infty. \tag{5.B.9}$$

Since C_1 is independent of n for $n\geq n^*$ and $\|\underline{\tau}^{[n]}\|_\infty\to 0$ as $n\to\infty$, this implies that

$$\lim_{n\to\infty}\|\underline{\mathbf{e}}^{[n]}\|_\infty=0.$$

Consider now the quadrature error

$$\delta^{[n]}:=\int_{-L}^{L}M(s)\mathbf{f}(s)ds-\sum_{j=1}^{n}w_jM(s_j)\mathbf{f}(s_j). \tag{5.B.10}$$

From (5.B.1) we obtain

$$\mathbf{d}^{[n]}=\delta^{[n]}-\sum_{j=1}^{n}w_jM(s_j)\mathbf{e}_j, \tag{5.B.11}$$

and using (5.B.7) the total discretization error for our methods is given by

$$\mathbf{d}^{[n]}=(\vec{1}^T\otimes\mathbf{I})\underline{W}\,\underline{M}(\alpha I+\underline{K}\,\underline{W})^{-1}\underline{\tau}^{[n]}+\delta^{[n]}. \tag{5.B.12}$$

Since both $\delta^{[n]}$ and $\tau^{[n]}$ are quadrature errors, $\|(\alpha I+\underline{K}\,\underline{W})^{-1}\|\leq C_1$ for all $n\geq n^*$, and M is bounded, the method converges at the same rate as the underlying quadrature.

5.B.1 Convergence of numerical method for closed loop geometry

By applying the formula (5.B.12), we now show how one can achieve spectral convergence in the case of a closed fiber geometry with constant radius ϵ and periodic integration domain. In this setting, we will use trapezoidal quadrature. We begin by bounding the norms of the integration kernels to which we apply the trapezoidal quadrature rules to, namely the integrals (5.4.1) and (5.4.4). Using this, and some smoothness assumptions, we are able bound the quadrature errors $\tau_i^{[n]}$ and $\delta^{[n]}$ using classical error estimates. This leads to a bound on the total error $\mathbf{d}^{[n]}$ for both the force and torque calculation.

Let C_2 be a constant such that

$$\|\mathbf{f}(s')\|_\infty \le C_2 \quad \text{for} \quad s \in [-L, L]. \tag{5.B.13}$$

From the definition of $K(s,s')$ (equations (5.2.2), (5.2.3), and (5.4.12)) in the constant radius case, we observe that

$$\|K(s,s')\|_\infty \le \frac{3}{2\epsilon} \tag{5.B.14}$$

with equality when $s = s'$. From equation (5.4.11) we have $\|M(s)\|_\infty = 1$ for the force calculation, while for the torque calculation, $M(s) = \widehat{\boldsymbol{X}}(s)$ and therefore

$$\|M(s)\|_\infty \le \max_{s\in[-L,L]} \|\boldsymbol{X}(s)\|_1. \tag{5.B.15}$$

Therefore we can bound the integration kernels of (5.4.1) and (5.4.4) by

$$\|K(s,s')\mathbf{f}(s')\|_\infty \le \frac{3}{2\epsilon} C_2 \tag{5.B.16}$$

and

$$\|M(s)\mathbf{f}(s)\|_\infty \le \|M(s)\|_\infty C_2. \tag{5.B.17}$$

Note that in the constant radius case, $K(s,s')$ has the same regularity as $\boldsymbol{X}(s)$. If we assume that $\boldsymbol{X}(s)$, $\mathbf{f}(s)$ and $M(s)$ are analytic, then using [56, Theorem 3.2] we can bound the trapezoidal rule quadrature error from equation (5.B.3) by

$$\|\tau_i^{[n]}\|_\infty \le \frac{6LC_2}{\epsilon(e^{an}-1)} \quad \text{for} \quad i = 1, \dots, n. \tag{5.B.18}$$

Similarly, we can bound equation (5.B.10) by

$$\|\delta^{[n]}\|_\infty \le \frac{4L\|M(s)\|_\infty C_2}{e^{an}-1}. \tag{5.B.19}$$

Here a is some constant. Using equation (5.B.12), the total discretization error is therefore bounded as

$$\|\mathbf{d}^{[n]}\|_\infty \le \left(\|(\vec{1}^T \otimes \mathbf{I})\underline{W}\,\underline{M}(\alpha I + \underline{K}\,\underline{W})^{-1}\|_\infty \frac{3}{2\epsilon} + \|M(s)\|_\infty \right) \frac{4LC_2}{e^{an}-1}. \tag{5.B.20}$$

Using that $\|\underline{M}\|_\infty \leq \|M(s)\|_\infty$, $\|\underline{W}\|_\infty = \frac{2L}{n}$ and C_1 is given by equation (5.B.8), this simplifies to

$$\|\mathbf{d}^{[n]}\|_\infty \leq \left(\frac{6C_1L}{2\epsilon}+1\right)\frac{4L\|M(s)\|_\infty C_2}{e^{an}-1}. \tag{5.B.21}$$

Hence, the method shares the same exponential convergence as the underlying trapezoidal rule. We remark that one could perform an analogous analysis for open ended fiber geometries with, e.g., Gauss-Lobatto quadrature, and derive similar results. Furthermore, we also remark that one could require less stringent regularity assumptions on the integration on the kernels or the fiber centreline $\boldsymbol{X}(s)$, e.g., $M(s)\mathbf{f}(s) \in C^{2m+2}[-L,L]$. Then (Ref. [3, Thm. 5.5]) can be used to derive asymptotic error estimates for $\tau_i^{[n]}$ and $\delta^{[n]}$ of order $O(h^{2m+2})$. Nonetheless, we do observe spectral convergence in numerical experiments in the following sections, as predicted by the bound (5.B.21).

5.C Dissipation matrix of a prolate spheroid

The non-dimensionalized body frame resistance tensor R_1 for a spheroid with aspect ratio λ was derived by Oberbeck [39] and is given by

$$R_1 = 16\pi\lambda \ \mathrm{diag}\left(\frac{1}{\chi_0+\alpha_0}, \frac{1}{\chi_0+\beta_0}, \frac{1}{\chi_0+\lambda^2\gamma_0}\right). \tag{5.C.1}$$

The constants χ_0, α_0, β_0 and γ_0 were calculated by Siewert [47] and are presented for a prolate ($\lambda > 1$) spheroid

$$\chi_0 = \frac{-\kappa_0\lambda}{\sqrt{\lambda^2-1}}, \tag{5.C.2}$$

$$\alpha_0 = \beta_0 = \frac{\lambda^2}{\lambda^2-1} + \frac{\lambda\kappa_0}{2(\lambda^2-1)^{3/2}}, \tag{5.C.3}$$

$$\gamma_0 = \frac{-2}{\lambda^2-1} - \frac{\lambda\kappa_0}{(\lambda^2-1)^{3/2}}, \tag{5.C.4}$$

$$\kappa_0 = \ln\left(\frac{\lambda-\sqrt{\lambda^2-1}}{\lambda+\sqrt{\lambda^2-1}}\right). \tag{5.C.5}$$

The torques $\mathbf{N} = (N_x, N_y, N_z)^\mathrm{T}$ that describe the rotational forces acting on an ellipsoid in creeping Stokes flow in the body frame were calculated by Jeffery [23] and are presented in their non-dimensional form with zero background flow

$$N_x = -\frac{16\pi\lambda}{3(\beta_0+\lambda^2\gamma_0)}\left[(1+\lambda^2)\omega_x\right], \tag{5.C.6}$$

$$N_y = -\frac{16\pi\lambda}{3(\alpha_0+\lambda^2\gamma_0)}\left[(1+\lambda^2)\omega_y\right], \tag{5.C.7}$$

$$N_z = -\frac{32\pi\lambda}{3(\alpha_0+\beta_0)}\omega_z. \tag{5.C.8}$$

Here $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)^{\mathrm{T}}$ is the body frame angular velocity, which is related to body frame angular momentum by $\mathbf{m} = J\boldsymbol{\omega}$. Taking derivatives of $\mathbf{N}$ with respect to $\mathbf{m}$ gives for the rotational dissipation matrix

$$R_2 = -\frac{16\lambda}{3}\operatorname{diag}\left(\frac{(1+\lambda^2)}{(\beta_0+\lambda^2\gamma_0)}, \frac{(1+\lambda^2)}{(\alpha_0+\lambda^2\gamma_0)}, \frac{2}{(\alpha_0+\beta_0)}\right)J^{-1}. \tag{5.C.9}$$

The full dissipation matrix used for the calculation in Figure 5.5.1 is given by

$$A_{sph} = \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix}. \tag{5.C.10}$$

Using discrete Darboux polynomials to detect and determine preserved measures and integrals of rational maps

Elena Celledoni, Charalambos A Evripidou, David I McLaren, Brynjulf Owren, G R W Quispel, Benjamin K Tapley and P H van der Kamp

Journal of Physics A: Mathematical and Theoretical 52 31LT01

Using discrete Darboux polynomials to detect and determine preserved measures and integrals of rational maps

Abstract. In this Letter we propose a systematic approach for detecting and calculating preserved measures and integrals of a rational map. The approach is based on the use of cofactors and Discrete Darboux Polynomials and relies on the use of symbolic algebra tools. Given sufficient computing power, all rational preserved integrals can be found. We show, in two examples, how to use this method to detect and determine preserved measures and integrals of the considered rational maps.

6.1 Introduction

The search for preserved measures and integrals of ordinary differential equations (ODEs) has been at the forefront of mathematical physics since the time of Galileo and Newton.

In this Letter our aim will be to develop an analogous theory for the (arguably more general) discrete-time case. This will lead to essentially linear algorithms for detecting and determining preserved measures and first and second integrals of (discrete) rational maps (both integrable and non-integrable).

But before we consider the discrete case, let us look at the continuous case, i.e. ODEs.

Consider two polynomials P_1 and P_2:

$$\begin{aligned} P_1(\mathbf{x}) &= \sum a_{i_1,\dots,i_n} x_1^{i_1} \dots x_n^{i_n} \\ P_2(\mathbf{x}) &= \sum b_{i_1,\dots,i_n} x_1^{i_1} \dots x_n^{i_n}. \end{aligned}$$

Then $I := P_1/P_2$ is a rational integral of the ODE $\frac{d\mathbf{x}}{dt} = f(\mathbf{x})$ if

$$\dot{P}_1 P_2 - P_1 \dot{P}_2 = 0$$

along solutions of the ODE. Here $\dot{}$ denotes $\frac{d}{dt}$.

For a polynomial ODE, the problem of finding P_1 and P_2, as posed, is bilinear in the parameters $a_{i_1,\dots,i_n}$ and $b_{i_1,\dots,i_n}$.

6.1.1 Darboux polynomials (ODE case)

Let $P(\mathbf{x})$ and $C(\mathbf{x})$ be polynomials.

Then $P(\mathbf{x})$ is called a Darboux polynomial of the polynomial ODE $\frac{d\mathbf{x}}{dt} = f(\mathbf{x})$, if

$$\dot{P}(\mathbf{x}) = C(\mathbf{x})P(\mathbf{x}).$$

Here $C(\mathbf{x})$ is called the co-factor of P.

Note that $P(\mathbf{x}(0)) = 0$ implies $P(\mathbf{x}(t)) = 0$ for all t. Hence the set $P(\mathbf{x}) = 0$ is an invariant set in phase space.

Consider two Darboux polynomials with the same co-factor C:

$$\begin{array}{l}\dot{P}_1 = CP_1\\ \dot{P}_2 = CP_2\end{array} \Rightarrow \frac{d}{dt}\left(\frac{P_1}{P_2}\right) = \frac{\dot{P}_1P_2 - P_1\dot{P}_2}{P_2^2} = \frac{CP_1P_2 - P_1CP_2}{P_2^2} = 0, \tag{6.1.1}$$

i.e. the ratio of two Darboux polynomials with the same cofactor is a rational integral. (The converse is also true).

However, finding C, P_1 and P_2 involves one bilinear problem, plus one linear problem. (Nevertheless, this approach can still be useful).

More generally,

$$\begin{array}{l}\dot{P}_1 = C_1P_1\\ \dot{P}_2 = C_2P_2\end{array} \Rightarrow \frac{d}{dt}(P_1P_2) = \dot{P}_1P_2 + P_1\dot{P}_2 = (C_1 + C_2)P_1P_2. \tag{6.1.2}$$

A very nice introduction to Darboux polynomials for ODEs was given by Goriely [6]. Note that Darboux polynomials were studied by Darboux, Poincaré, Painlevé and others [6], and are also known by several other names, including “second integrals" and “weak integrals".

6.1.2 Discrete Darboux Polynomials (mapping case)

Instead of polynomial ODEs $\frac{d\mathbf{x}}{dt} = f(\mathbf{x})$, we now consider rational maps $\mathbf{x}_{n+1} = \phi(\mathbf{x}_n)$ (cf [4, 5]).

Then we define $P(\mathbf{x})$ to be a Discrete Darboux Polynomial of the rational map $\mathbf{x}_{n+1} = \phi(\mathbf{x}_n)$ if

$$P(\mathbf{x}_{n+1}) = C(\mathbf{x}_n)P(\mathbf{x}_n),$$

where the co-factor C is now a rational function whose form will be presented in §1.3.

We use the shorthand notation

$$P' = CP$$

Note that, similarly to the continuous case, $P(\mathbf{x}) = 0$ is an invariant set in phase space.

Now consider again two Discrete Darboux Polynomials P_1 and P_2 with the same co-factor C:

$$\begin{array}{l} P_1' = CP_1 \\ P_2' = CP_2 \end{array} \Rightarrow \frac{P_1'}{P_2'} = \frac{P_1}{P_2},$$

i.e. the ratio of the two Discrete Darboux Polynomials with the same co-factors is again an integral (and the converse is also true).

More generally

$$\begin{array}{l} P_1' = C_1 P_1 \\ P_2' = C_2 P_2 \end{array} \Rightarrow (P_1 P_2)' = C_1 C_2 (P_1 P_2)$$

How is all this going to help us find integrals of a given map?

The answer comes in two parts:

1. In the discrete case we use a non-trivial ansatz for the co-factors $C(\mathbf{x})$. This ansatz works in all examples we have tried so far.

2. In the discrete case the co-factor of the product is the *product* of the co-factors.
   In the continuous case the co-factor of the product is the *sum* of the co-factors.

The latter point is crucial: It means that in the discrete case we can use the fact that the factorization of the co-factor C is unique. By contrast, in the ODE case we have addition, where splitting into summands is not unique.

6.1.3 Ansatz

Ansatz: The co-factors we use are of the form

$$C(\mathbf{x}) = \frac{1}{D^l(\mathbf{x})} \prod_i K_i^{a_i}(\mathbf{x})$$

where $D(\mathbf{x})$ is the common denominator of the map, and the $K_i(\mathbf{x})$ are factors of the numerator of the Jacobian determinant $J(\mathbf{x})$ of the map:

$$J(\mathbf{x}) = \frac{1}{D^m(\mathbf{x})} \prod_i K_i^{b_i}(\mathbf{x})$$

Comments:

1. There is a finite number of these co-factors up to a certain degree.
2. For each of this finite number of co-factors, we only need to solve a linear problem (up to a chosen degree).
3. If $C(\mathbf{x}) = J(\mathbf{x})$, the corresponding Darboux polynomials are (inverse) densities of preserved measures.

6.2 Determining preserved measures and first and second integrals of rational maps

In this section we study the following two-dimensional ODE as an example:

$$\begin{aligned} \frac{dx}{dt} &= x(x+6y-3) \\ \frac{dy}{dt} &= y(-3y-2x+3) \end{aligned} \tag{6.2.1}$$

The Kahan-Hirota-Kimura (KHK) discretization of (6.2.1) reads (cf [2,3,8–11, 14])

$$\begin{aligned} x' &= \frac{x(1+h(x+6y-3)+\frac{h^2}{4}(9-6x))}{D(x)} \\ y' &= \frac{y(1+h(3-2x-3y)+\frac{9h^2}{4}(1-2y))}{D(x)} \end{aligned} \tag{6.2.2}$$

where the common denominator $D(\mathbf{x})$ of the map is given by

$$D(\mathbf{x}) := 1 - \frac{h^2}{4}(9-12x-36y+4x^2+12xy+36y^2) \tag{6.2.3}$$

The Jacobian determinant $J(\mathbf{x})$ of the mapping (6.2.2) is

$$J(\mathbf{x}) = \frac{K_1(\mathbf{x})K_2(\mathbf{x})K_3(\mathbf{x})}{D^3(\mathbf{x})} \tag{6.2.4}$$

where

$$\begin{aligned} K_1 &= 1+h(x-3y)-\frac{3}{4}h^2(3-2x-6y) \\ K_2 &= 1+h(x+6y-3)-\frac{3}{4}h^2(3-2x) \\ K_3 &= 1+h(3-2x-3y)+\frac{9}{4}h^2(1-2y) \end{aligned} \tag{6.2.5}$$

We have used cofactors $C_1 = \frac{K_1}{D}$, $C_2 = \frac{K_2}{D}$, $C_3 = \frac{K_3}{D}$, $C_4 = J$ to find the corresponding Discrete Darboux Polynomials for the map (6.2.2):

$$\begin{aligned}
p_{1,1} &= x+3y-3\\
p_{2,1} &= x\\
p_{3,1} &= y\\
p_{4,1} &= xy(x+3y-3)\\
p_{4,2} &= 1-\frac{h^2}{4}(9-12x-36y+4x^2+12xy+36y^2)
\end{aligned}$$

Here $p_{i,j}$ denotes the j^{th} Darboux polynomial corresponding to the cofactor C_i.

A phase plot for the map (6.2.2), clearly exhibiting the linear Darboux polynomials $p_{1,1}$, $p_{2,1}$, and $p_{3,1}$, is given in Figure 1.

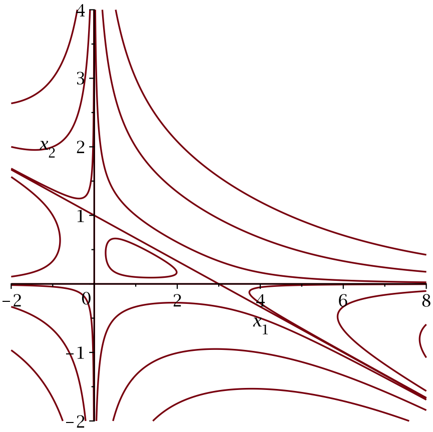


Figure 6.2.1: Phase plot for map (6.2.2) and for $h = \frac{1}{17}$

It follows that the map (6.2.2) preserves the integral

$$\tilde{I}(\mathbf{x}) = \frac{xy(x+3y-3)}{1-\frac{h^2}{4}(9-12x-36y+4x^2+12xy+36y^2)} \tag{6.2.6}$$

and the measure

$$\frac{dxdy}{1-\frac{h^2}{4}(9-12x-36y+4x^2+12xy+36y^2)} \tag{6.2.7}$$

Taking the continuum limit $h \to 0$, we obtain the cofactors $\tilde{C}_1 = x - 3y$, $\tilde{C}_2 = x + 6y - 3$, $\tilde{C}_3 = 3 - 2x - 3y$, $\tilde{C}_4 = 0$, and the corresponding Darboux polynomials

$$\begin{aligned} p_{1,1} =& x + 3y - 3 \\ p_{2,1} =& x \\ p_{3,1} =& y \\ p_{4,1} =& xy(x + 3y - 3) \\ p_{4,2} =& 1 \end{aligned}$$

It follows that the ODE (6.2.1) preserves the integral

$$I(\mathbf{x}) = xy(x + 3y - 3) \tag{6.2.8}$$

and the measure

$$dxdy. \tag{6.2.9}$$

It thus turns out that our original ODE (6.2.1) is Hamiltonian, with $H(x) = xy(x + 3y - 3)$.

Interpreted conversely, one can say that the KHK discretization (6.2.2) preserves the three affine Darboux polynomials of the ODE (6.2.1), as well as the modified integral (6.2.6) and the modified density (6.2.7). These results are no coincidences.

Indeed, the preservation of the three affine Darboux polynomials is the consequence of the following theorem (whose proof we will present elsewhere).

Theorem 6.1. *The KHK discretization preserves all affine Darboux polynomials of a given quadratic ODE.*

Theorem 6.1 is a very significant step towards the full resolution of the open problem posed in 2002 in [12]: ‘How does one preserve more than $n-1$ integrals and weak integrals (of an n-dimensional vector field)?’

The preservation of the modified integral and measure is an example of a general result in [2] giving a modified integral for all systems with a cubic Hamiltonian in any dimension.

6.3 Detecting preserved measures and first and second integrals of rational maps

In this section we consider the following three-dimensional ODE as an example:

$$\begin{aligned} \frac{dx}{dt} &= x(y-\mu z) \\ \frac{dy}{dt} &= y(\lambda z - x) \\ \frac{dz}{dt} &= z(\mu x - \lambda y), \end{aligned} \tag{6.3.1}$$

where λ and μ are arbitrary parameters.

Applying the Kahan-Hirota-Kimura discretization to (6.3.1), we obtain

$$\begin{aligned} \frac{x'-x}{h} &= \frac{x'(y-\mu z)+x(y'-\mu z')}{2} \\ \frac{y'-y}{h} &= \frac{y'(\lambda z-x)+y(\lambda z'-x')}{2} \\ \frac{z'-z}{h} &= \frac{z'(\mu x-\lambda y)+z(\mu x'-\lambda y')}{2}. \end{aligned} \tag{6.3.2}$$

Solving equation (6.3.2) for x', y', and z' we obtain the (rational) Kahan map discretizing (6.3.1). Using the Jacobian determinant $J(\mathbf{x})$ of the Kahan map as cofactor, our algorithm finds that for all (μ,λ), the map preserves the measure $\frac{dxdydz}{xyz}$ and the first integral $x+y+z$.

Moreover, the algorithm also detects the following special values of the parameters (μ,λ) where the map preserves an additional integral, and outputs the formula for the integral (cf. Table 1).

6.4 Concluding remarks

In this Letter we have presented a method for detecting and determining first and second integrals of rational maps. There are in the literature several other methods for *determining* first and second integrals of discrete systems, cf. [4, 5, 13, 16] and references therein. There are also in the literature several other methods for *detecting* first and second integrals of discrete systems, cf. [1,8,15] and references therein.

However, to our knowledge none of the above combine all the following properties of the method presented in this Letter:

Table 6.3.1: Integrable parameter values and corresponding functionally independent additional first integrals detected by our algorithm.

(μ, λ)	additional first integral
$(-1,0)$	y/z
$(1,0)$	$yz/(1-\frac{h^2}{4}x^2)$
$(0,1)$	$xz/(1-\frac{h^2}{4}y^2)$
$(0,-1)$	z/x
$(1,1)$	$xyz/(1-\frac{h^2}{4}(x^2+y^2+z^2-2xy-2xz-2yz))$
$(1,-1)$	x/yz
$(-1,-1)$	z/xy
$(-1,1)$	y/xz

1. It is algorithmic, and requires no other input than the rational map in question. At heart the algorithm is linear and, to some extent apart from birationality, requires no knowledge about the map (such as symplecticity, measure preservation, time-reversal symmetry, integrability, Lax pairs, etc) on the part of the user.

2. Up to a certain prescribed degree, it determines and outputs all:

   (a) rational first integrals

   (b) polynomial second integrals

   (c) preserved measures of the form $P(x)dx$ or $\frac{dx}{P(x)}$, where P is a polynomial.

3. It can detect special parameter values where additional preserved first and/or second integrals and/or measures exist, and output those integrals and measures.

4. It works for both integrable and non-integrable cases.

5. It allows one to take the continuum limit, if appropriate.

Acknowledgements

This work was supported by the Australian Research Council, by the Research Council of Norway, and by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 691070. GRWQ is grateful to K.Maruno for his hospitality at Waseda University, and to G.Gubbiotti for useful comments and correspondence.

Bibliography

[1] Abarenkova N, Anglès d'Auriac J-Ch, Boukraa S and Maillard J-M 2000, Real topological entropy versus metric entropy for birational measure-preserving transformations, *Physica* **D144** 387–433

[2] Celledoni E, McLachlan RI, Owren B and Quispel GRW 2013, Geometric properties of Kahan's method *J. Phys. A* **46** 12 pp. 025201

[3] Celledoni E, McLachlan RI, McLaren DI, Owren B and Quispel GRW 2014, Integrability properties of Kahan's method. *J. Phys. A* **47** 20 pp. 365202

[4] Falqui G and Viallet C-M 1993, Singularity, complexity, and quasi-integrability of rational mappings, *Comm. Math. Phys. A* **154**, 111–125

[5] Gasull A and Manosa V 2010, A Darboux-type theory of integrability for discrete dynamical systems, *Journal of Difference Equations and Applications* **8** 1171-1191

[6] Goriely A 2001, Integrability and Nonintegrability of Dynamical Systems, *World Scientific, Singapore*, section 2.5

[7] Halburd RG and Korhonen RJ 2017, Three approaches to detecting discrete integrability, *ArXiv: 1704.07927*

[8] Hirota R and Kimura K 2000, Discretization of the Euler top, *J. Phys. Soc. Jap.* **69** 627–630.

[9] Hone A N W and Petrera M 2009, Three dimensional discrete systems of Hirota-Kimura type and deformed Lie-Poisson algebras, *Journal of Geometric mechanics* **1** No.1 55–85.

[10] Kimura K and Hirota R, 2000, Discretization of the Lagrange top. *J. Phys. Soc. Japan* **69** 3193–3199.

[11] Kahan W 1993, Unconventional numerical methods for trajectory calculations, *Unpublished lecture notes*.

[12] McLachlan RI and Quispel GRW 2002, Splitting Methods, *Acta Numerica* **11** 341–434, section 6.1.

[13] Papageorgiou VG, Nijhoff FW and Capel HW 1990, Integrable mappings and nonlinear integrable lattice equations, *Phys Lett* **147A** 106–114

[14] Petrera M, Pfadler A and Suris YB 2011, On integrability of Hirota–Kimura type discretizations, *Regular and Chaotic Dynamics* **16** 245–289.

[15] JAG Roberts and F Vivaldi, 2003, Arithmetical method to detect integrability in maps, *Phys Rev Lett* **90**(3) 034102.

[16] Tran DT, van der Kamp PH and Quispel GRW 2009, Closed-form expressions for integrals of travelling wave reductions of integrable lattice equations, *J Phys A* **42** 225201

Appendix

Detecting and determining preserved measures and integrals of rational maps

Elena Celledoni, Charalambos A Evripidou, David I McLaren, Brynjulf Owren, G R W Quispel and Benjamin K Tapley

Submitted

Detecting and determining preserved measures and integrals of rational maps

Abstract. In this paper we use the method of discrete Darboux polynomials to calculate preserved measures and integrals of rational maps. The approach is based on the use of cofactors and Darboux polynomials and relies on the use of symbolic algebra tools. Given sufficient computing power, most, if not all, rational preserved integrals can be found (and even some non-rational ones). We show, in a number of examples, how it is possible to use this method to both determine and detect preserved measures and integrals of the considered rational maps. Many of the examples arise from the Kahan-Hirota-Kimura discretization of completely integrable systems of ordinary differential equations.

7.1 Introduction

Suppose $(x_1', x_2', \dots, x_n') = \mathbf{x}' = \phi(\mathbf{x}) = \phi(x_1, x_2, \dots, x_n)$ defines a rational map of $\mathbb{R}^n$, and let $\mathbb{R}_p[\mathbf{x}]$ be the class of polynomials up to degree p in n variables. We say that I is a preserved first integral of ϕ if and only if $I(\mathbf{x}') = I(\mathbf{x})$. The rational map ϕ preserves a measure of the form

$$\int \frac{1}{m(\mathbf{x})} \mathrm{d}x_1 \wedge \cdots \wedge \mathrm{d}x_n, \quad m \in \mathbb{R}_p[\mathbf{x}], \tag{7.1.1}$$

if the condition

$$m(\phi(\mathbf{x})) = J\, m(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^n, \tag{7.1.2}$$

is satisfied, where J is the determinant of the Jacobian matrix of ϕ. The reciprocal of the density m is here assumed to be a polynomial (or the reciprocal of a polynomial) and the preserved integrals are assumed rational. In this paper, we devise a systematic approach for searching for such preserved measures and integrals of a rational map.

Our interest in such properties of rational maps originates from the study of Kahan's numerical discretization of first order quadratic ordinary differential equations and its analogue for higher order and higher degree on the one hand, [3, 7–9], and from the study of discrete integrable systems [14] on the other. Kahan [5] proposed a numerical method designed for quadratic systems of differential equations in $\mathbb{R}^n$ written in component form as

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{j,k} a_{ijk} x_j x_k + \sum_j b_{ij} x_j + c_i, \quad i = 1, \dots, n, \tag{7.1.3}$$

where a_{ijk}, b_{ij}, c_i are arbitrary constants and all summation indices are ranging from 1 to n. The method of Kahan, also known as the Hirota–Kimura discretization [2,4], is a one-step method $(x_1,\ldots,x_n) \mapsto (x'_1,\ldots,x'_n)$ where

$$\frac{x'_i - x_i}{h} = \sum_{j,k} a_{ijk} \frac{x'_j x_k + x_j x'_k}{2} + \sum_j b_{ij} \frac{x_j + x'_j}{2} + c_i, \quad i = 1,\ldots,n, \tag{7.1.4}$$

where h denotes the discrete time step. The method (7.1.4) is linearly implicit and so is its inverse, hence it defines a birational map.

Much of the recent interest in Kahan's method stems from its ability to preserve modified first integrals and measures of the underlying quadratic differential equation [7, 8]. But even in cases where there are strong indications that Kahan's method preserves such a nearby invariant, it is not necessarily an easy task to determine its closed form.

Consider a rational map $\mathbf{x}_{n+1} = \phi(\mathbf{x}_n)$. We define the polynomial $P(\mathbf{x})$ to be a (discrete) Darboux polynomial of the map ϕ if there exists a rational function $C(\mathbf{x})$ s.t.

$$P(\mathbf{x}_{n+1}) = C(\mathbf{x}_n)P(\mathbf{x}_n),$$

where the form of $C(\mathbf{x}_n)$ will be prescribed below.

Note that if

$$P_i(\mathbf{x}_{n+1}) = C_i(\mathbf{x}_n)P_i(\mathbf{x}_n), \quad i = 1,\ldots,k,$$

then

$$\left(\prod_i P_i^{a_i}(\mathbf{x}_{n+1})\right) = \left(\prod_i C_i^{a_i}(\mathbf{x}_n)\right)\left(\prod_i P_i^{a_i}(\mathbf{x}_n)\right),$$

so that if

$$\prod_i C_i^{a_i}(\mathbf{x}_n) \equiv 1$$

then

$$\prod_i P_i^{a_i}(\mathbf{x}_n)$$

is an integral of the map ϕ.

The remaining question is how to prescribe the form of the discrete cofactor $C_i(\mathbf{x})$.

In [10] we introduced the following Ansatz:

Given a rational map ϕ with Jacobian determinant

$$J(\mathbf{x}) = \frac{\prod_{i=1}^{l} K_i^{b_i}(\mathbf{x})}{\prod_{j=1}^{k} D_j^{c_j}(\mathbf{x})},$$

where the K_i and D_j are distinct factors, we try all cofactors (up to a certain polynomial degree d) of the form

$$C(\mathbf{x}) = \pm \frac{\prod_{i=1}^{l} K_i^{f_i}(\mathbf{x})}{\prod_{j=1}^{k} D_j^{g_j}(\mathbf{x})},$$

where $f_i, g_j \in \mathbb{N}_0$.

For each such cofactor we try all Darboux polynomials, (again up to a certain degree). The algorithm has been implemented in a symbolic algebra system (our codes are made with Maple version 2019) and the codes are adapted to run on clusters of up to 32 cores with up to 768 GBs of memory if necessary.

We presented two examples in [10] of the use of the above approach in determining Darboux polynomials of a given map (the first example arose from the Kahan discretisation of a $2D$ Lotka-Volterra system, and the second from discrete integrable systems), plus a third example illustrating the detection of special parameter values with extra Darboux polynomials.

As additional supporting evidence for the usefulness of the above Ansatz we also proved in [10] that given a quadratic ODE in dimensions 1,2,3, or 4, possessing an affine Darboux polynomial, the Kahan discretisation of the ODE preserves the DP, and the numerator of the corresponding cofactor divided the numerator of the Jacobian determinant of the map.

In the present paper we present 11 examples exhibiting various aspects of discrete Darboux polynomials. Eight examples involve determining Darboux polynomials, three additional examples involve parametric detection. In most cases the map ϕ is obtained as the application of Kahan's method (7.1.4) to a quadratic differential equation (7.1.3). We also consider three examples which are unrelated to Kahan's method. Finally, we prove that for any quadratic Hamiltonian ODE, the modified integral as well as the modified preserved measure of the Kahan map are both found using the above Ansatz with $C(\mathbf{x}) = \pm J(\mathbf{x})$.

7.2 Determining preserved measures and integrals

7.2.1 Example 1: finding measures and integrals of a specific 2D vector field

In this subsection we study the following two-dimensional vector field:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2x_1x_2 - 4x_2 \\ -3x_1^2 - x_2^2 + 4x_1 + 1 \end{pmatrix}.$$

The Kahan discretization of this vector field is given by

$$\begin{aligned} x_1' &= \frac{x_1 + h(2x_1x_2 - 4x_2) + h^2(2x_1^2 - 2x_2^2 - 3x_1 - 2)}{D(\mathbf{x})}, \\ x_2' &= \frac{x_2 + h(-3x_1^2 - x_2^2 + 4x_1 + 1) + h^2(4x_1x_2 - 5x_2)}{D(\mathbf{x})}, \end{aligned} \tag{7.2.1}$$

where the common denominator $D(\mathbf{x})$ is given by

$$D(\mathbf{x}) = 1 + h^2(3x_1^2 - x_2^2 - 8x_1 + 4).$$

The Jacobian determinant J of the Kahan map ϕ_h (7.2.1) is given by

$$J = C_1(\mathbf{x})C_2(\mathbf{x}),$$

where

$$\begin{aligned} C_1(\mathbf{x}) &= \frac{1 + 2hx_2 + h^2(5 - 4x_1)}{D(\mathbf{x})}, \\ C_2(\mathbf{x}) &= \frac{1 - 2hx_2 + h^2(7 - 20x_1 + 9x_1^2 + x_2^2) + h^3(26x_2 - 16x_1x_2) + h^4(28 - 28x_1 + 7x_1^2 + 3x_2^2)}{D^2(\mathbf{x})}. \end{aligned}$$

Defining $C_3 := J$, we have used cofactors C_1, C_2 and C_3 to find the corresponding Darboux polynomials for the Kahan map (7.2.1):

$$\begin{aligned} p_{1,1} &= x_1 - 2, \\ p_{2,1} &= 1 - x_1^2 - x_2^2 + h^2(\frac{13}{3} - \frac{16}{3}x_1 + x_2^2 + \frac{7}{3}x_1^2), \\ p_{3,1} &= 1 + h^2(3x_1^2 - x_2^2 - 8x_1 + 4), \\ p_{3,2} &= (x_1 - 2)\,(1 - x_1^2 - x_2^2 + h^2(\frac{13}{3} - \frac{16}{3}x_1 + x_2^2 + \frac{7}{3}x_1^2)). \end{aligned}$$

Here and below, $p_{i,j}$ denotes the jth Darboux polynomial corresponding to the cofactor C_i, i.e., $p_{i,j}$ satisfies

$$p_{i,j}(\mathbf{x}') = C_i(\mathbf{x})p_{i,j}(\mathbf{x}).$$

Note also that here $p_{3,1}(\mathbf{x}) \equiv D(\mathbf{x})$. So it turns out that the Kahan map (7.2.1) possesses the second integrals $p_{1,1}(\mathbf{x})$, $p_{2,1}(\mathbf{x})$ and $p_{3,1}(\mathbf{x})$, and also preserves the measure

$$\frac{\mathrm{d}x\mathrm{d}y}{1 + h^2(3x_1^2 - x_2^2 - 8x_1 + 4)}.$$

Finally, the Kahan map preserves the first integral

$$\frac{(x_1 - 2)\,(1 - x_1^2 - x_2^2 + h^2(\frac{13}{3} - \frac{16}{3}x_1 + x_2^2 + \frac{7}{3}x_1^2))}{1 + h^2(3x_1^2 - x_2^2 - 8x_1 + 4)},$$

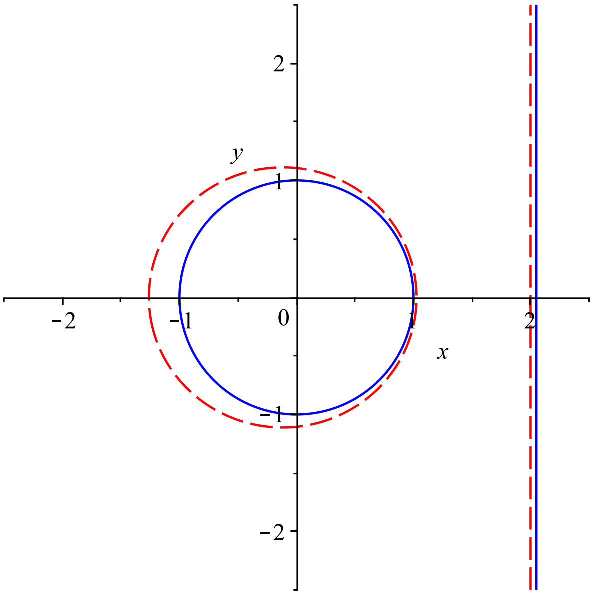


Figure 7.2.1: Plots of Darboux polynomials $p_{1,1}$ and $p_{2,1}$ of the Kahan map in Example 2 (for $h = \frac{1}{5}$), dotted red. Also shown are the corresponding second integrals of the ODE, $x_1 - 2$ and $1 - x_1^2 - x_2^2$, solid blue.

Taking the continuum limit $h \to 0$, we now see in hindsight that the vector field (7.2.1) possesses two second integrals, i.e. $x_1 - 2$ resp. $1 - x_1^2 - x_2^2$. It also preserves the measure $\mathrm{d}x_1\mathrm{d}x_2$ and the first integral $H = (x_1 - 2)(1 - x_1^2 - x_2^2)$. The fact that the affine Darboux polynomial $x_1 - 2$ is preserved by the Kahan map ϕ_h is an example of theorem (1) of [10], which states that the Kahan discretization preserves all affine Darboux polynomials in any dimension. On the other hand, the fact that the vector field (7.2.1) preserves the integral H and the measure $\mathrm{d}x_1\mathrm{d}x_2$ implies that (7.2.1) is a Hamiltonian vector field with cubic Hamiltonian H. Therefore, the fact that the Kahan method preserves a modified Hamiltonian $\tilde{H}$, and the modified densities $p_{3,1}$ and $p_{3,2}$ is a special case of the following theorem:

Theorem 7.1. *Let H be cubic in $\mathbb{R}^n$, let K be a constant rank $2l$ antisymmetric $n \times n$ matrix and let the vector field be given by $f = K\nabla H(\mathbf{x})$. Then:*

(i) $\phi_h(\mathbf{x})$ possesses the following two Darboux polynomials, both with cofactor $C_1(\mathbf{x}) = J$:

$$\begin{aligned} p_{1,1} &= H(\mathbf{x})\det(A(\mathbf{x})) + \frac{1}{3}h\nabla H(\mathbf{x})^t \mathrm{adj}(A(\mathbf{x}))f(\mathbf{x}), \\ p_{1,2} &= \det(A(\mathbf{x})). \end{aligned}$$

Here $A(\mathbf{x}) = \mathbb{I} - \frac{1}{2}hf'(\mathbf{x})$*, and* $\mathrm{adj}(A)$ *denotes the adjugate of* A*.*

(ii) moreover, the degree of $p_{1,2}$ *is at most* $2l$ *and the degree of* $p_{1,1}$ *is at most* $2l+3$*. If* $n = 2l$ *the degree of* $p_{1,1}$ *is at most* $2l+1$*.*

Proof. In the proof of Proposition 4 in [7], it is shown that ϕ_h possesses the modified integral $\tilde{H} = \frac{p_{1,1}}{p_{1,2}}$. Proposition 5 in [7] is equivalent to the statement that $p_{1,2}$ is a Darboux polynomial with cofactor J. Combining these two results, it follows that $p_{1,1}$ is also a Darboux polynomial with cofactor J. Part *(ii)* follows from proposition 4*(i)* in [7].

□

7.2.2 Example 2: An inhomogeneous Nambu system

We consider the following inhomogeneous Nambu system belonging to the class of systems considered in [1]

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} (x_1+3x_2)(5x_2+12x_3+8) \\ -(x_1+x_2)(5x_2+12x_3+8) \\ (x_1+x_2)(8x_2+5x_1+7) \end{pmatrix},$$

which has the two integrals

$$\begin{aligned} H &:= x_1^2 + 2x_1x_2 + 3x_2^2, \\ K &:= 4x_2^2 + 5x_2x_3 + 6x_3^2 + 7x_2 + 8x_3, \end{aligned}$$

and the preserved measure

$$\int \mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3.$$

We consider the Kahan discretization of these equations. The corresponding Jacobian determinant can be factorized in two irreducible factors:

$$J = C_1(\mathbf{x})C_2(\mathbf{x}).$$

Letting C_1 and C_2 play the role of cofactors, we find the Darboux polynomials $p_{1,1}, p_{1,2}, p_{2,1}$ and $p_{2,2}$:

$$\begin{aligned} p_{1,1} &= 50h^2x_1^2 + 100h^2x_1x_2 - 720h^2x_2x_3 - 864h^2x_3^2 - 480h^2x_2 \\ &\quad -1152h^2x_3 - 384h^2 - 3, \\ p_{1,2} &= 50h^2x_2^2 + 240h^2x_2x_3 + 288h^2x_3^2 + 160h^2x_2 + 384h^2x_3 \\ &\quad +128h^2 + 1, \\ p_{2,1} &= (270h^2x_1^2 + 540h^2x_1x_2 + 270h^2x_2^2 - 4x_2^2 - 5x_2x_3 \\ &\quad -6x_3^2 - 7x_2 - 8x_3)/270h^2, \\ p_{2,2} &= \frac{142x_2^2}{135} + \frac{71x_2x_3}{54} + \frac{71x_3^2}{45} + \frac{497x_2}{270} + \frac{284x_3}{135} + 1. \end{aligned}$$

From these we obtain that the preserved integrals of the Kahan map are $\frac{p_{1,1}(\mathbf{x})}{p_{1,2}(\mathbf{x})}, \frac{p_{2,1}(\mathbf{x})}{p_{2,2}(\mathbf{x})}$, and any combination

$$\frac{1}{p_{1,i}(\mathbf{x})p_{2,j}(\mathbf{x})}\mathrm{d}\mathbf{x}, \qquad i,j \in \{1,2\}$$

is a preserved measure.

7.2.3 Example 3: Quartic Nahm system in 2D

We consider the following example whose Kahan discretization was studied in [6]

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2x_1^2 - 12x_2^2 \\ -6x_1x_2 - 4x_2^2 \end{pmatrix}. \tag{7.2.2}$$

This ODE has a preserved integral

$$H := x_2(2x_1 + 3x_2)(x_1 - x_2)^2,$$

and a preserved measure

$$\int \frac{\mathrm{d}x_1 \mathrm{d}x_2}{x_2(2x_1 + 3x_2)(x_1 - x_2)}.$$

The Jacobian determinant of the Kahan discretization has the following factors

$$J = \frac{L_1\, L_2\, L_3}{D^3},$$

where the three affine L_i are given by

$$\begin{aligned} L_1(\mathbf{x}) &:= 1 + 3hx_1 - 8hx_2 \\ L_2(\mathbf{x}) &:= 1 - 5hx_1. \\ L_3(\mathbf{x}) &:= 1 + 3hx_1 + 12hx_2 \end{aligned}$$

and the quadratic D is

$$D(\mathbf{x}) := 1 + hx_1 + 4hx_2 - 6h^2x_1^2 - 8h^2x_1x_2 - 36h^2x_2^2$$

Among the cofactors $L_1^i L_2^j L_3^k / D^l$ for $i,j,k = 0,1$ and $l = 1,\ldots,3$ we consider $C_1 = \frac{L_1}{D}$, $C_2 = \frac{L_2}{D}$, and $C_3 = \frac{L_3}{D}$, satisfying $J = C_1(\mathbf{x})C_2(\mathbf{x})C_3(\mathbf{x})$. The corresponding Darboux polynomials are

$$p_{1,1}(\mathbf{x}) = 2x_1 + 3x_2, \quad p_{2,1}(\mathbf{x}) = x_2, \quad p_{3,1}(\mathbf{x}) = x_2 - x_1$$

leading to the preserved measure

$$\frac{d\mathbf{x}}{p_{1,1}(\mathbf{x})p_{2,1}(\mathbf{x})p_{3,1}(\mathbf{x})}.$$

To find the modified integral, we search for Darboux polynomials whose cofactors are of the form $C_1^i(\mathbf{x})C_2^j(\mathbf{x})$ for $i, j = 1, 2, \ldots$ (i.e., "super-factors" of J). Using the cofactor

$$C_4(\mathbf{x}) := C_1(\mathbf{x})C_2(\mathbf{x})C_3(\mathbf{x})^2,$$

we find

$$\begin{aligned}
p_{4,1}(\mathbf{x}) &= x_2(2x_1+3x_2)(x_1-x_2)^2,\\
p_{4,2}(\mathbf{x}) &= 9h^4x_1^4+272h^4x_1^3x_2-352h^4x_1x_2^3+696h^4x_2^4-10h^2x_1^2-40h^2x_2^2+1,
\end{aligned}$$

and $\frac{p_{4,1}(\mathbf{x})}{p_{4,2}(\mathbf{x})}$ is an integral of the Kahan discretization, see the corresponding example in [6].

7.2.4 Example 4: Lagrange top

Discretizations of the Lagrange top have been studied in [4] and [6], where it was shown that the Kahan map preserves a number of modified integrals. The Lagrange top reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} m_1\\ m_2\\ m_3\\ p_1\\ p_2\\ p_3 \end{pmatrix} = \begin{pmatrix} (\alpha-1)\,m_2m_3+\gamma\,p_2\\ (1-\alpha)\,m_1m_3-\gamma\,p_1\\ 0\\ \alpha\,p_2m_3-p_3m_2\\ p_3m_1-\alpha\,p_1m_3\\ p_1m_2-p_2m_1 \end{pmatrix},$$

where m_i and p_i are the angular and linear momentum components and α and γ are constant parameters. The Lagrange top admits four independent integrals

$$\begin{aligned}
H_1 &= p_1^2+p_2^2+p_3^2,\\
H_2 &= p_1m_1+p_2m_2+p_3m_3,\\
H_3 &= m_1^2+m_2^2+\alpha m_3^2+2\gamma p_3,\\
H_4 &= m_3.
\end{aligned}$$

For the Lagrange top, it suffices to treat m_3 as a free parameter by working in the variables $\bar{\mathbf{x}} = (m_1, m_2, p_1, p_2, p_3)^{\mathrm{T}}$ and look for degree-six Darboux polynomial densities in $\bar{\mathbf{x}}$. Using the cofactor $C_1(\mathbf{x}) = J$, we find the following five Darboux polynomial densities

$$\begin{aligned}
p_{1,1} = {} & -256\gamma + 64\,h^2\gamma\left(-2\,{m_3}^2\alpha^2 + 2\,{m_3}^2\alpha + \gamma\,p_3 - {m_1}^2 - {m_2}^2 - {m_3}^2\right) \\
& + h^4 Q^{(4)}_{1,2} + h^6 Q^{(6)}_{1,3} + h^8 Q^{(8)}_{1,4}, \\
p_{1,2} = {} & -2048\gamma^3 + 256\,h^2\gamma^3\left(-2\,{m_3}^2\alpha^2 + 2\,{m_3}^2\alpha + 4\gamma\,p_3 - {m_1}^2 - {m_2}^2 - 2\,{m_3}^2\right) \\
& + h^4 Q^{(4)}_{2,2} + h^6 Q^{(6)}_{2,3} + h^8 Q^{(8)}_{2,4} + h^{10} Q^{(8)}_{2,5}, \\
p_{1,3} = {} & -2048\,m_3\,(6\,\alpha - 5) + h^2(-8192\,\alpha^3{m_3}^3 + 10240\,\alpha^2{m_3}^3 + 6144\,\alpha\gamma\,m_3p_3 - 1536\,\alpha\,{m_1}^2m_3 \\
& -1536\,\alpha\,{m_2}^2m_3 - 4096\,\alpha\,{m_3}^3 - 512\gamma\,p_1m_1 - 512\gamma\,p_2m_2 - 5120\gamma\,m_3p_3 + 1536\,m_3{m_1}^2 \\
& +1536\,m_3{m_2}^2 + 1024\,{m_3}^3) + h^4 Q^{(5)}_{3,2} + h^6 Q^{(7)}_{3,3} + h^8 Q^{(9)}_{3,4} + h^{10} Q^{(11)}_{3,5} + h^{12} Q^{(11)}_{3,6}, \\
p_{1,4} = {} & -256\gamma\,m_3\,(2\,\alpha - 1) + 64\,h^2\gamma\,(-2\,\alpha^3{m_3}^3 + 3\,\alpha^2{m_3}^3 + 4\,\alpha\gamma\,m_3p_3 - \alpha\,{m_1}^2m_3 - \alpha\,{m_2}^2m_3 \\
& -3\,\alpha\,{m_3}^3 + \gamma\,p_1m_1 + \gamma\,p_2m_2 - 2\gamma\,m_3p_3 + {m_3}^3) + h^4 Q^{(5)}_{4,2} + h^6 Q^{(7)}_{4,3} + h^8 Q^{(9)}_{4,4}, \\
p_{1,5} = {} & -65536 + h^2\big(-32768\,{m_3}^2\alpha^2 + 40960\,{m_3}^2\alpha + 16384\gamma\,p_3 - 16384\,{m_1}^2 - 16384\,{m_2}^2 \\
& -32768\,{m_3}^2\big) + h^4 Q^{(4)}_{5,2} + h^6 Q^{(6)}_{5,3} + h^8 Q^{(8)}_{5,4} + h^{10} Q^{(10)}_{5,5} + h^{12} Q^{(12)}_{5,6} + h^{14} Q^{(12)}_{5,7},
\end{aligned}$$

where each $Q^{(i)}_{j,k}$ is a polynomial of degree i in the variables $\mathbf{x} = (m_1, m_2, m_3, p_1, p_2, p_3)^{\mathrm{T}}$. Taking the quotients $\frac{p_{1,1}}{p_{1,5}}$, $\frac{p_{1,2}}{p_{1,5}}$, $\frac{p_{1,3}}{p_{1,5}}$ and $\frac{p_{1,4}}{p_{1,5}}$ yields four functionally independent integrals. Taking functionally dependent combinations of these, we are able to form the following integrals that are preserved by the Kahan discretization

$$\begin{aligned}
\tilde{H}_1 = {} & \tfrac{p_1^2+p_2^2+p_3^2+\mathcal{O}(h^2)}{1+\mathcal{O}(h^2)} \\
\tilde{H}_2 = {} & \tfrac{p_1m_1+p_2m_2+p_3m_3+\mathcal{O}(h^2)}{1+\mathcal{O}(h^2)} \\
\tilde{H}_3 = {} & \tfrac{m_1^2+m_2^2+\alpha m_3^2+2\gamma p_3+\mathcal{O}(h^2)}{1+\mathcal{O}(h^2)} \\
\tilde{H}_4 = {} & m_3,
\end{aligned}$$

where the first three integrals are modified versions of the continuous integrals.

7.2.5 Example 5: An ODE with many linear Darboux polynomials

Consider the following quadratic ODE in 4 dimensions

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 9\,{x_1}^2 + 48\,x_3x_1 - 40\,x_1x_4 + 24\,{x_2}^2 - 48\,x_2x_3 + 48\,x_2x_4 + 48\,{x_3}^2 + 24\,x_4x_3 - 132\,{x_4}^2 + x_1 \\ -2\,{x_1}^2 - 12\,x_3x_1 + 12\,x_1x_4 - 5\,{x_2}^2 + 12\,x_2x_3 - 14\,x_2x_4 - 12\,{x_3}^2 - 6\,x_4x_3 + 38\,{x_4}^2 + x_2 \\ -4\,{x_1}^2 - 24\,x_3x_1 + 24\,x_1x_4 - 14\,{x_2}^2 + 28\,x_2x_3 - 28\,x_2x_4 - 25\,{x_3}^2 - 12\,x_4x_3 + 76\,{x_4}^2 + x_3 \\ -2\,{x_1}^2 - 12\,x_3x_1 + 12\,x_1x_4 - 6\,{x_2}^2 + 12\,x_2x_3 - 12\,x_2x_4 - 12\,{x_3}^2 - 6\,x_4x_3 + 37\,{x_4}^2 + x_4 \end{pmatrix}.$$

The Jacobian determinant of the Kahan discretization of this ODE is

$$J = \frac{{L_1}^4{L_2}^4}{{D_1}^2{D_2}^2{D_3}^2{D_4}^2}$$

where

$$\begin{aligned} L_1 &= h-2 \\ L_2 &= h+2 \\ D_1 &= -2+\left(2x_1+8x_4+1\right)h \\ D_2 &= -2+\left(4x_2-2x_3+1\right)h \\ D_3 &= -2+\left(2x_2-2x_4+1\right)h \\ D_4 &= -2+\left(4x_1+6x_3+2x_4+1\right)h \end{aligned}$$

The following 14 linear discrete Darboux polynomials $p_{i,1}$ for $i=1,...,14$ are found corresponding to the cofactors C_i

i	$p_{i,1}$	C_i
1	x_1+4x_4+1	L_1/D_1
2	x_1+4x_4	L_2/D_1
3	$2x_2-x_3+1$	L_1/D_2
4	$-2x_2+x_3$	L_2/D_2
5	x_2-x_4+1	L_1/D_3
6	$-x_2+x_4$	L_2/D_3
7	$2x_1+3x_3+x_4+1$	L_1/D_4
8	$2x_1+3x_3+x_4$	L_2/D_4
9	$x_1-2x_2+x_3+4x_4$	$L_1L_2/(D_1D_2)$
10	$x_1-x_2+5x_4$	$L_1L_2/(D_1D_3)$
11	$3x_3+x_1-3x_4$	$L_1L_2/(D_1D_4)$
12	$-x_2+x_3-x_4$	$L_1L_2/(D_2D_3)$
13	$-2x_2+4x_3+2x_1+x_4$	$L_1L_2/(D_2D_4)$
14	$-2x_1+x_2-3x_3-2x_4$	$L_1L_2/(D_3D_4)$

(7.2.3)

We know that if $\prod_i C_i^{\delta_i}=1$ is satisfied, then $\prod_i {p_i}^{\delta_i}$ is an integral. Solving for the δ_i, we find the following nine integrals of the Kahan map

$$\tilde{H}_1=\frac{p_{1,1}p_{4,1}}{p_{2,1}p_{3,1}},\quad \tilde{H}_2=\frac{p_{1,1}p_{6,1}}{p_{2,1}p_{5,1}},\quad \tilde{H}_3=\frac{p_{1,1}p_{8,1}}{p_{2,1}p_{7,1}},\quad \tilde{H}_4=\frac{p_{9,1}}{p_{2,1}p_{3,1}},$$

$$\tilde{H}_5=\frac{p_{10,1}}{p_{2,1}p_{5,1}},\quad \tilde{H}_6=\frac{p_{11,1}}{p_{2,1}p_{7,1}},\quad \tilde{H}_7=\frac{p_{1,1}p_{12,1}}{p_{2,1}p_{3,1}p_{5,1}},$$

$$\tilde{H}_8=\frac{p_{1,1}p_{13,1}}{p_{2,1}p_{3,1}p_{7,1}},\quad \tilde{H}_9=\frac{p_{1,1}p_{14,1}}{p_{2,1}p_{5,1}p_{7,1}}$$

of which 3 are independent e.g.,

$$\tilde{H}_4 = \frac{x_1 - 2\,x_2 + x_3 + 4\,x_4}{\left(x_1 + 4\,x_4\right)\left(2\,x_2 - x_3 + 1\right)}, \tag{7.2.4}$$

$$\tilde{H}_5 = \frac{x_1 - x_2 + 5\,x_4}{\left(x_1 + 4\,x_4\right)\left(x_2 - x_4 + 1\right)}, \tag{7.2.5}$$

$$\tilde{H}_6 = \frac{3\,x_3 + x_1 - 3\,x_4}{\left(x_1 + 4\,x_4\right)\left(2\,x_1 + 3\,x_3 + x_4 + 1\right)} \tag{7.2.6}$$

Moreover, we see that e.g.,

$$\frac{\mathrm{d}\mathbf{x}}{p_{9,1}\,p_{11,1}\,p_{12,1}\,p_{14,1}} \tag{7.2.7}$$

is a preserved measure of the Kahan map. Hence the Kahan map is super integrable. Since the integrals (7.2.4) and the measure (7.2.7) do not depend on the time step, it follows that they are all preserved by the ODE (7.2.3). We hope to present further examples of ODEs with many linear Darboux polynomials in a forthcoming paper.

7.2.6 Example 6 A Kahan map having a non-rational integral

Consider the ODE having vector field

$$\frac{d}{dt}\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\right) = \left(\begin{array}{l} -2\,{x_4}^2 + \left(-2\,x_1 - x_3 + 5\right)x_4 - {x_1}^2 + x_1 + x_2 + 3\,x_3 \\ -2\,{x_4}^2 + \left(-2\,x_1 - x_3 + 3\right)x_4 - {x_1}^2 - x_1 + x_2 + 3\,x_3 \\ -2\,{x_4}^2 + \left(-2\,x_1 - x_3 - 5\right)x_4 - {x_1}^2 - 3\,x_1 - 3\,x_2 + x_3 \\ 2\,{x_4}^2 + \left(2\,x_1 + x_3\right)x_4 + {x_1}^2 \end{array}\right) \tag{7.2.8}$$

The Kahan map for this ODE has 2 Darboux polynomials p_1 resp. p_2 with constant cofactors C_1 resp. C_2, where

$$\begin{aligned} C_1 &= \frac{1 + h/2}{1 - h/2} \\ C_2 &= \frac{1 + h + 5h^2}{1 - h + 5h^2} \\ p_1 &= 3\,x_1 - 3\,x_2 + x_3 + x_4 \\ p_2 &= 5\,{x_1}^2 + 9\,x_1x_2 - 3\,x_3x_1 + 16\,x_4x_1 + 5\,{x_2}^2 + 3\,x_2x_3 \\ &\quad +22\,x_4x_2 + 9\,{x_3}^2 + 18\,x_4x_3 + 28\,{x_4}^2 \end{aligned}$$

By definition,

$$p_i(\mathbf{x}_{n+1}) = C_i\,p_i(\mathbf{x}_n)$$

and since C_i constant, we can solve for p_i:

$$p_i(\mathbf{x}_n) = \alpha_i C_i^n$$

where the α_i are integration constants.

It also follows that

$$I := \frac{p_1(\mathbf{x}_n)^{\ln C_2/\ln C_1}}{p_2(\mathbf{x}_n)}$$

is an (h-dependent) generally non-rational integral of the Kahan map. Note that in the continuum limit $h \to 0$ this implies that

$$\frac{p_1^2(\mathbf{x})}{p_2(\mathbf{x})}$$

is a (rational) integral of the ODE (7.2.8).

7.2.7 Example 7: A 4D polarization map

Here we consider the 4-dimensional map presented in [9] (choosing $a = 2, b = 1, c = -3,\ d = -1, e = 1$ in their notation).

$$\begin{aligned}
x_1' &= x_2\\
x_2' &= \frac{2\,hx_1{}^2x_2 - 6\,hx_1{}^2x_4 - 12\,hx_1x_2x_3 - 4\,hx_1x_3x_4 - 2\,hx_2x_3{}^2 + 2\,hx_3{}^2x_4 + x_1}{D(\mathbf{x})}\\
x_3' &= x_4\\
x_4' &= \frac{-4\,hx_1{}^2x_2 - 2\,hx_1{}^2x_4 - 4\,hx_1x_2x_3 + 12\,hx_1x_3x_4 + 6\,hx_2x_3{}^2 + 2\,hx_3{}^2x_4 + x_3}{D(\mathbf{x})}
\end{aligned}$$

where the quartic D is[*]

$$\begin{aligned}
D(\mathbf{x}) := \quad & -28\,h^2x_1{}^2x_2{}^2 + 4\,h^2x_1{}^2x_2x_4 - 40\,h^2x_1{}^2x_4{}^2 + 4\,h^2x_1x_2{}^2x_3 - 28\,h^2x_1x_2x_3x_4\\
& -8\,h^2x_1x_3x_4{}^2 - 40\,h^2x_2{}^2x_3{}^2 - 8\,h^2x_2x_3{}^2x_4 - 16\,h^2x_3{}^2x_4{}^2 + 1
\end{aligned}$$

The determinant of the Jacobian of the map has the factorized form

$$J = \frac{K_1}{D^3}$$

where

$$\begin{aligned}
K_1(\mathbf{x}) := \quad & -128\,h^3x_1{}^3x_2{}^3 + 456\,h^3x_1{}^3x_2{}^2x_4 - 120\,h^3x_1{}^3x_2x_4{}^2 + 496\,h^3x_1{}^3x_4{}^3\\
& +984\,h^3x_1{}^2x_2{}^3x_3 + 432\,h^3x_1{}^2x_2{}^2x_3x_4 + 936\,h^3x_1{}^2x_2x_3x_4{}^2 + 336\,h^3x_1{}^2x_3x_4{}^3\\
& +408\,h^3x_1x_2{}^3x_3{}^2 - 864\,h^3x_1x_2{}^2x_3{}^2x_4 + 216\,h^3x_1x_2x_3{}^2x_4{}^2 - 528\,h^3x_1x_3{}^2x_4{}^3\\
& -488\,h^3x_2{}^3x_3{}^3 - 192\,h^3x_2{}^2x_3{}^3x_4 - 264\,h^3x_2x_3{}^3x_4{}^2 - 80\,h^3x_3{}^3x_4{}^3\\
& -84\,h^2x_1{}^2x_2{}^2 + 12\,h^2x_1{}^2x_2x_4 - 120\,h^2x_1{}^2x_4{}^2 + 12\,h^2x_1x_2{}^2x_3 - 84\,h^2x_1x_2x_3x_4\\
& -24\,h^2x_1x_3x_4{}^2 - 120\,h^2x_2{}^2x_3{}^2 - 24\,h^2x_2x_3{}^2x_4 - 48\,h^2x_3{}^2x_4{}^2 + 1
\end{aligned}$$

[*] Erratum: In eqs (4.1) of [9], $1 - 4h^2\Delta$ should read $1 + 4h^2\Delta$.

Using J as cofactor, the resulting functionally independent Darboux polynomials are

$$\begin{aligned}
p_1 &= D(\mathbf{x}) \\
p_2 &= \left(x_1x_4 - x_2x_3\right)(4\,hx_1{}^2x_2{}^2 + 4\,hx_1{}^2x_2x_4 - 6\,hx_1{}^2x_4{}^2 + 4\,hx_1x_2{}^2x_3 - 24\,hx_1x_2x_3x_4 \\
&\quad -4\,hx_1x_3x_4{}^2 - 6\,hx_2{}^2x_3{}^2 - 4\,hx_2x_3{}^2x_4 + 2\,hx_3{}^2x_4{}^2 + x_1x_4 - x_2x_3) \\
p_3 &= 26\,hx_1{}^3x_2{}^2x_4 + 10\,hx_1{}^3x_2x_4{}^2 + 2\,hx_1{}^3x_4{}^3 - 26\,hx_1{}^2x_2{}^3x_3 - 60\,hx_1{}^2x_2x_3x_4{}^2 \\
&\quad -8\,hx_1{}^2x_3x_4{}^3 - 10\,hx_1x_2{}^3x_3{}^2 + 60\,hx_1x_2{}^2x_3{}^2x_4 + 14\,hx_1x_3{}^2x_4{}^3 - 2\,hx_2{}^3x_3{}^3 \\
&\quad +8\,hx_2{}^2x_3{}^3x_4 - 14\,hx_2x_3{}^3x_4{}^2 + 2\,x_1{}^2x_2{}^2 + 2\,x_1{}^2x_2x_4 + 2\,x_1x_2{}^2x_3 - 18\,x_1x_2x_3x_4 \\
&\quad -2\,x_1x_3x_4{}^2 - 2\,x_2x_3{}^2x_4 + x_3{}^2x_4{}^2
\end{aligned}$$

The map thus possesses the preserved measure $\int \frac{\mathbf{dx}}{p_1}$ and the (independent) first integrals $\frac{p_2}{p_1}$ and $\frac{p_3}{p_1}$, in agreement with [9].

7.2.8 Example 8: sine-Gordon maps

In this subsection we consider $(k+1)$-dimensional maps that arise as so-called $(1,k)$ reductions of the discrete sine-Gordon equation [17].

We start with the case $k=3$, then treat the case $k=2$, before giving a general theorem for arbitrary k.

The $(1,3)$ sine-Gordon map.

The $(1,3)$ sine-Gordon map ϕ is given by

$$\begin{aligned}
x_i' &= x_{i+1}, \qquad i = 0,1,2, \\
x_3' &= \frac{1-\alpha x_1x_3}{x_0(x_1x_3-\alpha)},
\end{aligned}$$

where α is a parameter. Using $C_1(\mathbf{x}) = J$, we find the corresponding Darboux polynomials:

$$\begin{aligned}
p_{1,1} &= x_3x_2x_1x_0, \\
p_{1,2} &= x_0{}^2x_1x_2x_3{}^2 - \alpha\,x_0{}^2x_2x_3 - \alpha\,x_0x_1{}^2x_3 - \alpha\,x_0x_1x_2{}^2 \\
&\quad -\alpha\,x_0x_1x_3{}^2 - \alpha\,x_0x_2{}^2x_3 - \alpha\,x_1{}^2x_2x_3 + x_1x_2, \\
p_{1,3} &= x_0{}^2x_1{}^2x_2x_3 + x_0x_1{}^2x_2{}^2x_3 + x_0x_1x_2{}^2x_3{}^2 - \alpha\,x_0{}^2x_1x_2 \\
&\quad -\alpha\,x_1x_2x_3{}^2 + x_0x_1 + x_0x_3 + x_3x_2.
\end{aligned}$$

It follows that ϕ possesses the (independent) first integrals $\frac{p_{1,2}}{p_{1,1}}$ and $\frac{p_{1,3}}{p_{1,1}}$, and the preserved measure $\int \frac{\mathbf{dx}}{p_{1,1}}$. These results were found using different methods in ref [17].

The $(1,2)$ sine-Gordon map.

The $(1,2)$ sine-Gordon map ϕ is given by

$$\begin{aligned} x_i' &= x_{i+1}, \qquad i=0,1, \\ x_2' &= \frac{1-\alpha x_1 x_2}{x_0(x_1x_2-\alpha)}. \end{aligned}$$

Using $C_1(\mathbf{x}) = -J$, we find the corresponding Darboux polynomials:

$$\begin{aligned} p_{1,1} &= x_0x_1x_2, \\ p_{1,2} &= {x_0}^2x_1{x_2}^2 - \alpha\, {x_0}^2x_2 - \alpha\, x_0{x_1}^2 - \alpha\, x_0{x_2}^2 - \alpha\, {x_1}^2x_2 + x_1, \\ p_{1,3} &= {x_0}^2{x_1}^2x_2 + x_0{x_1}^2{x_2}^2 - \alpha\, {x_0}^2x_1 - \alpha\, x_1{x_2}^2 + x_0 + x_2. \end{aligned}$$

It follows that ϕ possesses the (independent) first integrals $\frac{p_{1,2}}{p_{1,1}}$ and $\frac{p_{1,3}}{p_{1,1}}$, and the preserved measure $\int \frac{\mathrm{dx}}{p_{1,1}}$. Note that this is one extra first integral, that was not found using the Lax representation approach of ref [17]. Moreover, we find that there is an additional cofactor $C_2(\mathbf{x}) = J$, for which we find the corresponding Darboux polynomial

$$p_{2,1} = -{x_0}^2x_1{x_2}^2 + \alpha\, {x_0}^2x_2 - \alpha\, x_0{x_1}^2 + \alpha\, x_0{x_2}^2 - \alpha\, {x_1}^2x_2 + x_1.$$

Normally, a sole Darboux polynomial does not yield an integral, but, because $C_2 = -C_1$, we find that

$$I := \frac{p_{2,1}}{p_{1,1}}$$

is a so-called 2-integral [13], i.e. an integral of $\phi\circ\phi$. In this case $I(\mathbf{x}') = -I(\mathbf{x})$. This result was not found using the Lax matrix approach in [17].

The $(1,k)$ sine-Gordon map.

The $(1,k)$ sine-Gordon map ϕ_k is given by

$$\begin{aligned} x_i' &= x_{i+1}, \qquad i=0,\ldots,k-1, \\ x_k' &= \frac{1-\alpha x_1 x_k}{x_0(x_1x_k-\alpha)}, \end{aligned} \tag{7.2.9}$$

where $\lfloor\frac{k+1}{2}\rfloor$ functionally independent rational integrals for this map were found using a Lax matrix approach in [17].

Denote these integrals by $I_k^n(\mathbf{x}) = \frac{N_k^n(\mathbf{x})}{D_k^n(\mathbf{x})}$, $n = 1,\ldots,\lfloor\frac{k+1}{2}\rfloor$, and define

$$\epsilon := (-1)^{k+1} \tag{7.2.10}$$

Theorem 7.2. *For all n and k, the Darboux polynomials N_k^n and D_k^n are given by*

$$N_k^n(\mathbf{x}') = C(\mathbf{x})N_k^n(\mathbf{x}),$$
$$D_k^n(\mathbf{x}') = C(\mathbf{x})D_k^n(\mathbf{x}),$$

where the cofactor $C(\mathbf{x})$ depends only on k, and is given by $C(\mathbf{x}) = \epsilon|D\phi_k(\mathbf{x})|$.

The proof is given in appendix A.

7.3 Detecting Darboux polynomials and integrals

Given a rational map $\mathbf{x}' = \phi(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$ containing k free parameters denoted by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$, one could ask if there exist particular choices of $\boldsymbol{\alpha}$ such that ϕ preserves additional second integrals. This amounts to solving the non-linear cofactor equation

$$p(\mathbf{x}') = C(\mathbf{x}; \boldsymbol{\alpha})p(\mathbf{x}) \tag{7.3.1}$$

for the Darboux polynomial indeterminants as well as the parameters, where $C(\mathbf{x}; \boldsymbol{\alpha})$ can be non-linear in $\boldsymbol{\alpha}$.

7.3.1 Example 9: Extended McMillan map

Consider the following rational map $\phi(\mathbf{x})$ defined by

$$\phi\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -x_2 - f(x_1) \\ x_1 \end{pmatrix},$$

where

$$f(x_1) = \frac{\alpha_1 x_1{}^3 + \alpha_2 x_1{}^2 + \alpha_3 x_1 + \alpha_4}{\alpha_5 x_1{}^2 + \alpha_2 x_1 + \alpha_6},$$

and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_6)$ are free parameters. The integrability of a special case of this map was studied in [18]. The Jacobian of the map ϕ is $J = 1$. If all parameters $\boldsymbol{\alpha}$ are arbitrary, the equation

$$m(\phi(\mathbf{x})) = m(\mathbf{x}) \tag{7.3.2}$$

has only one solution $m_1(\mathbf{x}) = 1$. Solving the cofactor equation $\boldsymbol{\alpha}$ yields the condition $\alpha_1 = 0$. This is an integrable map known as the McMillan map [19]. Enforcing this condition, one now finds two solutions to equation (7.3.2)

$$\begin{aligned} m_1(\mathbf{x}) = &\quad 1, \\ m_2(\mathbf{x}) = &\quad \alpha_5 x_1{}^2 x_2{}^2 + \alpha_2\left(x_1{}^2 x_2 + x_1 x_2{}^2\right) + \alpha_3 x_1 x_2 + \alpha_6\left(x_1{}^2 + x_2{}^2\right) \\ &\quad +\alpha_4\left(x_1 + x_2\right), \end{aligned}$$

where $m_2(\mathbf{x})$ is a preserved integral of ϕ, in agreement with [19].

7.3.2 Example 10: Two coupled Euler tops

We now consider two coupled Euler tops whose vector field is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} a_1^2 x_2 x_3 \\ a_2^2 x_3 x_1 \\ a_3^2 x_1 x_2 + a_4^2 x_4 x_5 \\ a_5^2 x_5 x_3 \\ a_6^2 x_3 x_4 \end{pmatrix}.$$

This system was first presented in [16], and its integrals after discretisation were first explored in [6], where the authors present the following three independent integrals of motion

$$H_1 = a_2^2 x_1^2 - a_1^2 x_2^2, \quad H_3 = a_6^2 x_4^2 - a_5^2 x_5^2,$$
$$H_2 = a_3^2 a_5^2 x_2^2 - a_2^2 a_5^2 x_3^2 + a_2^2 a_4^2 x_4^2,$$

however we note the existence of a fourth independent integral given by

$$H_4 = \frac{\left(a_1 x_2 + a_2 x_1\right)^{a_5 a_6}}{\left(a_5 x_5 + a_6 x_4\right)^{a_1 a_2}},$$

hence the system is super-integrable. To our knowledge, the integral H_4 may be new. The Jacobian determinant of the Kahan map has the following factors

$$J = \frac{L_1 L_2 L_3 L_4 L_5}{D^6}.$$

The cofactors $C_i = \frac{L_i}{D}$, for $i = 1, ..., 5$ admit the following linear Darboux polynomials

$$\begin{aligned} p_{1,1}(\mathbf{x}) &= \quad a_5 x_5 + a_6 x_4, \\ p_{2,1}(\mathbf{x}) &= \quad a_5 x_5 - a_6 x_4, \\ p_{4,1}(\mathbf{x}) &= \quad a_1 x_2 + a_2 x_1, \\ p_{5,1}(\mathbf{x}) &= \quad a_1 x_2 - a_2 x_1, \end{aligned}$$

however, the cofactor C_3 admits no polynomial solutions, even up to degree 6. Now we look for quadratic Darboux polynomials with the cofactors $C_6 := C_1 C_2$ and $C_7 := C_4 C_5$ and get the following

$$p_{6,1}(\mathbf{x}) = p_{1,1} p_{2,1}, \quad p_{6,2}(\mathbf{x}) = (2 - h a_5 a_6 x_3)(2 + h a_5 a_6 x_3), \tag{7.3.3}$$
$$p_{7,1}(\mathbf{x}) = p_{4,1} p_{5,1}, \quad p_{7,2}(\mathbf{x}) = (2 - h a_1 a_2 x_3)(2 + h a_1 a_2 x_3). \tag{7.3.4}$$

We note that $p_{6,2}(\mathbf{x})$ and $p_{7,2}(\mathbf{x})$ also factorise. In the ODE case, if a Darboux polynomial factorises, each factor is also a Darboux polynomial. In the discrete case that need not be the case, and indeed it often is not true as

$$p_{6,2}(\mathbf{x}) = q_{6,1}(\mathbf{x})q_{6,2}(\mathbf{x}), \quad p_{7,2}(\mathbf{x}) = q_{7,1}(\mathbf{x})q_{7,2}(\mathbf{x}),$$

where the $q_{i,j}(\mathbf{x})$ satisfy

$$\begin{aligned} q_{6,1}(\mathbf{x}') &= C_1 q_{6,2}(\mathbf{x}), \quad q_{6,2}(\mathbf{x}') = C_2 q_{6,1}(\mathbf{x}) \\ q_{7,1}(\mathbf{x}') &= C_4 q_{7,2}(\mathbf{x}), \quad q_{7,2}(\mathbf{x}') = C_5 q_{7,1}(\mathbf{x}) \end{aligned}$$

which implies that each $q_{i,j}(\mathbf{x})$ is in fact a discrete Darboux polynomial of the second iterate of the Kahan map. The Darboux polynomials from equations (7.3.3) and (7.3.4) yield two independent integrals $\frac{p_{6,1}(\mathbf{x})}{p_{6,2}(\mathbf{x})}$ and $\frac{p_{7,1}(\mathbf{x})}{p_{7,2}(\mathbf{x})}$, in agreement with [6]. We also note that no Darboux polynomial measures are found up to degree 6 using J as the cofactor.

We now now attempt to solve the cofactor equation

$$p(\mathbf{x}') = C_3(\mathbf{x};\boldsymbol{\alpha})p(\mathbf{x}). \tag{7.3.5}$$

For a polynomial basis of degree 2, equation (7.3.5) admits three conditions that yield non-trivial Darboux polynomials: $a_3 = 0$, $a_4 = 0$ and $a_1^2 a_2^2 = a_5^2 a_6^2$. The first two correspond to the decoupling of two of the equations and these two less interesting cases have three independent discrete integrals each. The third condition is presented in [6]. In this case the Jacobian determinant of the Kahan discretization now factors as

$$J = \frac{{L_6}^3 {L_7}^3}{D^6}.$$

Using $C_8 = \frac{L_6 L_7}{D^2}$ as the cofactor, we get the following six Darboux polynomials

$$\begin{aligned} p_{8,1}(\mathbf{x}) &= a_2^4 x_1^2 - a_5^2 a_6^2 x_2^2, \\ p_{8,2}(\mathbf{x}) &= a_2^2 x_1 x_5 - a_6^2 x_2 x_4, \\ p_{8,3}(\mathbf{x}) &= a_2^2 x_1 x_4 - a_5^2 x_2 x_5, \\ p_{8,4}(\mathbf{x}) &= a_2^2 a_4^2 x_4^2 - a_2^2 a_5^2 x_3^2 + a_3^2 a_5^2 x_2^2, \\ p_{8,5}(\mathbf{x}) &= a_2^2 a_4^2 x_5^2 - a_2^2 a_6^2 x_3^2 + a_3^2 a_6^2 x_2^2, \\ p_{8,6}(\mathbf{x}) &= 4 - a_5^2 a_6^2 x_3^2 h^2. \end{aligned}$$

Hence, the following measures are preserved

$$\int \frac{\mathrm{d}\mathbf{x}}{p_{8,i}(\mathbf{x})p_{8,j}(\mathbf{x})p_{8,k}(\mathbf{x})}, \quad \text{for any} \quad i,j,k = 1,\dots,6,$$

and the following integrals are preserved

$$\frac{p_{8,i}}{p_{8,k}}, \quad \text{for} \quad i \neq k,$$

of which four are independent. The choice $k = 6$ yields the integrals presented in [6].

7.3.3 Example 11: A family of Nambu systems with rational integrals

Here we will consider Nambu systems, of the form

$$\dot{\mathbf{x}} = c\big(\nabla H \times \nabla K\big), \tag{7.3.6}$$

where $c = y^{2-\alpha}$, $H = \frac{x}{y}$, $K = y^{\alpha} Q$, Q is homogeneous and quadratic and α is a free parameter. We get the following Jacobian determinant for the Kahan map

$$J = \frac{L_1^2 L_2}{D^4}.$$

The cofactor $C_1 := L_1/D$ has the following two Darboux polynomials at degree one

$$p_{1,1} = x \ \text{ and } \ p_{1,2} = y,$$

hence the integral H is preserved exactly by the Kahan method. The cofactor $C_2 := L_2/D^2$ has the following Darboux polynomial

$$p_{2,1} = Q.$$

Using $C_3 = |D\phi(\mathbf{x})|$ we find that the Kahan discretisation has three preserved measures corresponding to the densities

$$p_{3,1} = x^2 Q, \ \ p_{3,2} = xyQ \ \text{ and } \ p_{3,3} = y^2 Q,$$

which yields only one independent integral H.
Now choosing integer coefficients for Q, we can use our detection algorithm to search for any values of the parameter α such that the Kahan discretisation yields extra Darboux polynomial solutions. To do this, we solve the cofactor equation

$$p(\mathbf{x}') = |D\phi(\mathbf{x})| p(\mathbf{x})$$

for α and the Darboux polynomial p of degree 4. This gives us the following solutions for α and the corresponding additional second integral of the Kahan discretisation:

α	Integrals
-2	H and K
-1	H and K
0	H and $\tilde{K}_0$
1	H and $\tilde{K}_1$
2	H and $\tilde{K}_2$

where

$$\begin{aligned}
\tilde{K}_0 &= \frac{Q}{12+h^2(1853\,xy+3485\,xz+938\,y^2+2665\,yz+1435\,z^2)}\\
\tilde{K}_1 &= \frac{yQ}{1-h^2(226\,x^2+211\,xy+119\,xz+64\,y^2+91\,yz+49\,z^2)}\\
\tilde{K}_2 &= \frac{y^2Q}{48-h^2(26616\,x^2+23472\,xy+11424\,xz+6840\,y^2+8736\,yz+4704\,z^2)-h^4\,A}
\end{aligned}$$

and

$$\begin{aligned}
A = 6309873\,x^3 \quad z - 10784832\,x^2yz + 1918455\,x^2z^2 + 27341015\,xy^3 + 37337147\,xy^2z\\
-7467243\,xyz^2 - 559776\,xz^3 + 14528513\,y^4 + 37680292\,y^3z\\
+19891900\,y^2z^2 - 428064\,yz^3 - 115248\,z^4
\end{aligned}$$

7.4 Concluding remarks

We have proposed a systematic approach to search for the preserved measures and integrals of a rational map and applied it to a number of examples. The method is based on the use of Darboux polynomials. We have shown that the method can be used to both determine and detect measures and integrals. Some of the examples have required the use of relatively large computer memory space and computational time.

Acknowledgements

This work was partially supported by the Australian Research Council, by the Research Council of Norway, by the project “International mobilities for research activities of the University of Hradec Kralove" CZ.02.2.69/0.0/0.0/16 027/0008487, by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 691070, and by the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme GCS when work on this paper was undertaken. This work was also supported by: EPSRC grant number EP/R014604/1. Celledoni and Quispel are also grateful to the Simons Foundation for Fellowships supporting this work.

Bibliography

Appendix

Proof of Theorem 2.

In [17] it was shown that $\lfloor\frac{k+1}{2}\rfloor$ functionally independent integrals of the $(1,k)$ sine-Gordon map (7.2.9) are given by the trace of the Lax matrix $L^{1,k}$:

$$TrL^{1,k}(\mathbf{x},\lambda) = Tr\left[\left(\begin{array}{cc} qx_0/x_k & \lambda^{-2}/x_k \\ x_0 & q \end{array}\right)\prod_{l=0}^{k-1}\left(\begin{array}{cc} p & -x_{l+1} \\ -\lambda^2/x_l & px_{l+1}/x_l \end{array}\right)\right] \quad (7..1)$$

where $pq = \alpha$. The individual integrals are given by the coefficients of the various powers of the spectral parameter λ in the expansion of the rhs.

It was also shown in [17] that the sine-Gordon map ϕ_k is either measure preserving or anti measure preserving, i.e. satisfies

$$P(\mathbf{x}') = \epsilon|D\phi_k(\mathbf{x})|P(\mathbf{x}), \quad (7..2)$$

where $P(\mathbf{x}) := \prod_{l=0}^{k} x_l$, and ϵ is given by (7.2.10).

It is easy to see that the rhs of (7..1) is equal to

$$\begin{aligned} & Tr\left[\frac{1}{x_k}\left(\begin{array}{cc} qx_0 & \lambda^{-2} \\ x_0x_k & qx_k \end{array}\right)\prod_{l=0}^{k-1}\frac{1}{x_l}\left(\begin{array}{cc} px_l & -x_lx_{l+1} \\ -\lambda^2 & x_{l+1} \end{array}\right)\right] \\ = \quad & Tr\left[\left(\begin{array}{cc} qx_0 & \lambda^{-2} \\ x_0x_k & qx_k \end{array}\right)\prod_{l=0}^{k-1}\left(\begin{array}{cc} px_l & -x_lx_{l+1} \\ -\lambda^2 & x_{l+1} \end{array}\right)\right] / \left[\prod_{l=0}^{k} x_l\right] \quad (7..3) \end{aligned}$$

We now recognize that the denominator of the integrals (7..3) equals the Darboux polynomial P in (7..2). Bearing in mind that the matrices in the trace in (7..3) are all polynomial, it follows using Theorem 1 that this trace is also a Darboux polynomial with the same cofactor $C = \epsilon|D\phi_k(\mathbf{x})|$, for all values of λ.
□

Bibliography

[1] Celledoni E, McLaren DI, Owren B, Quispel GRW, Geometric and integrability properties of Kahan's method: the preservation of certain quadratic integrals, *J. Phys. A* **52** 065201 9 pp.

[2] Hirota R and Kimura K 2000, Discretization of the Euler top, *J. Phys. Soc. Jap.* **69** 627–630.

[3] Hone ANW, and Quispel GRW, Analogues of Kahan's method for higher order equations of higher degree, accepted for publication in Proceedings in Mathematics & Statistics (arXiv: 1911.03161)

[4] Kimura K, and Hirota R, 2000, Discretization of the Lagrange top. *J. Phys. Soc. Japan* **69** 3193–3199.

[5] Kahan W 1993, Unconventional numerical methods for trajectory calculations, Unpublished lecture notes.

[6] Petrera M, Pfadler A, and Suris YB 2011, On integrability of Hirota–Kimura type discretizations, *Regular and Chaotic Dynamics* **16** 245–289.

[7] Celledoni E, McLachlan RI, Owren B, Quispel, GRW 2013, Geometric properties of Kahan's method *J. Phys. A* **46** 12 025201

[8] Celledoni E, McLachlan RI, McLaren DI, Owren B, and Quispel GRW 2014, Integrability properties of Kahan's method. *J. Phys. A* **47** 20 365202

[9] Celledoni E, McLachlan RI, McLaren DI, Owren B, and Quispel GRW 2014, Discretization of polynomial vector fields by polarizatopn *Proc.R.Soc. A* **471** 20150390

[10] Celledoni E, Evripidou C, McLaren DI, Owren B, Quispel GRW, Tapley BK, and van der Kamp P, Using discrete Darboux polynomials to detect and determine preserved measures and integrals of rational maps, *J. Phys. A* **52** 31 31LT01

[11] G Falqui and C-M Viallet, *Singularity, complexity, and quasi-integrability of rational mappings*, Commun. Math. Phys. **154** (1993) 111-125

[12] A Gasull and V Manosa, A Darboux-type theory of integrability for discrete dynamical systems, *Journal of Difference Equations and Applications* **8** (2010) 1171–1191

[13] Haggar FA, Byrnes GB, Quispel GRW, and Capel HW, k-integrals and k-Lie symmetries in discrete dynamical systems, *Physica A* **233** (1996), 379–394

[14] J Hietarinta, N Joshi, and FW Nijhoff, Discrete Systems and Integrability CUP, 2016

[15] Hone A N W, and Petrera M 2009, Three dimensional discrete systems of Hirota-Kimura type and deformed Lie-Poisson algebras, Journal of Geometric mechanics **1** No.1 55–85

[16] Golse F, Mahalov A, and Nicolaenko B 2008, Bursting dynamics of the 3D Euler equations in cylindrical domains, in *Instability in models connected with fluid flows: 1* Int. Math. Ser. (N.Y.), vol **6**, New York:Springer 300–338

[17] Quispel GRW, HW Capel, VG Papageorgiou, and FW Nijhoff, *Integrable mappings derived from soliton equations* Physica A, 173 (1991), 243–266.

[18] JAG Roberts and F Vivaldi, 2003, Arithmetical method to detect integrability in maps, *Phys Rev Lett* **90**(3) 034102

[19] E. M. McMillan, in *Topics in Modern Physics. A Tribute to E.U. Condon*, edited by E. Britton and H. Odabasi (Colorado University Press, Boulder, 1971), 219–244.

On the preservation of affine second integrals by Runge-Kutta methods

Benjamin K Tapley

Preprint

On the preservation of affine second integrals by Runge-Kutta methods

Abstract. One can elucidate integrability properties of ordinary differential equations (ODEs) by knowing the existence of second integrals. However, little is known about how they are preserved, if at all, under numerical methods. Here, we present a number of novel results about the preservation of affine second integrals of ODEs when discretised by Runge-Kutta methods. In particular, we show that all Runge-Kutta methods preserve all affine second integrals with a modified discrete cofactor. We also discuss the preservation of higher affine integrals and show that Runge-Kutta methods can preserve some rational integrals for certain ODEs.

8.1 Introduction

Consider an autonomous ODE in $\mathbb{R}^n$

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \tag{8.1.1}$$

then a *second integral* [6] of (8.1.1) is a function $p(\mathbf{x})$ that satisfies

$$\dot{p}(\mathbf{x}) = c(\mathbf{x})p(\mathbf{x}) \tag{8.1.2}$$

where the dot denotes $\frac{\mathrm{d}}{\mathrm{d}t}$ and $c(\mathbf{x})$ is called the *cofactor* of $p(\mathbf{x})$. Polynomial second integrals of polynomial ODEs are also referred to as Darboux polynomials. In contrast to (8.1.2), a *discrete second integral* (or *discrete Darboux polynomial* when $\varphi_h(\mathbf{x})$ is rational polynomial) of a map $\varphi_h : \mathbb{R}^n \to \mathbb{R}^n$ is a function $p(\mathbf{x})$ that satisfies

$$p(\varphi_h(\mathbf{x})) = \tilde{c}(\mathbf{x})p(\mathbf{x}) \tag{8.1.3}$$

where $\tilde{c}(\mathbf{x})$ is called the *discrete cofactor* of $p(\mathbf{x})$. This is a discrete analogue of equation (8.1.2) and was recently introduced in [1,2].

In this paper we will consider ODEs with one or more affine second integrals of the form $p(\mathbf{x}) = \mathbf{p}^T\mathbf{x} + r$ where $\mathbf{p} \in \mathbb{R}^n$ and $r \in \mathbb{R}$. Note that we can take the constant $r = 0$ without loss of generality. In particular, we focus on the case where the map φ_h comes about as a Runge-Kutta method applied to an ODE that possesses one or more second integrals. Letting $\varphi_h(\mathbf{x})$ denote one step of an s-stage Runge-Kutta method applied to (8.1.1) with initial condition $\mathbf{x}$ and Butcher tableau given by

$$\begin{array}{c|c} \mathscr{C} & \mathscr{A} \\ \hline & \boldsymbol{b}^T \end{array} \tag{8.1.4}$$

then this defines the method

$$\mathbf{g}_i = \mathbf{x} + h\sum_{j=1}^{s} a_{ij}\mathbf{f}(\mathbf{g}_j), \quad \text{for } i = 1, ..., s \tag{8.1.5}$$

$$\varphi_h(\mathbf{x}) = \mathbf{x} + h\sum_{j=1}^{s} b_j\mathbf{f}(\mathbf{g}_j) \tag{8.1.6}$$

where $\varphi_h(\mathbf{x})$ is the Runge-Kutta map with time-step h applied to the initial point $\mathbf{x}$ and $\mathscr{A} = [a_{ij}]$. For explicit Runge-Kutta maps the sum in equation (8.1.5) runs from 1 to $i-1$ as $\mathscr{A}$ is strictly lower triangular.

The paper begins with some general results about the preservation of affine second integrals with general cofactors. We then focus on the special case where the cofactor is constant. The final section is on the preservation of affine higher integrals.

8.2 Preservation of affine second integrals

We begin with an example of a planar ODE and its discretisation by a second order Runge-Kutta method.

Example 4. Consider the following ODE in two dimensions

$$\dot{x} = x^2 + 2xy + 3y^2, \quad \dot{y} = 2y\big(2x + y\big), \tag{8.2.1}$$

This ODE was studied in [5] and has the following three linear Darboux polynomials

$$p_1(\mathbf{x}) = x + y, \quad p_2(\mathbf{x}) = x - y, \quad p_3(\mathbf{x}) = y, \tag{8.2.2}$$

that correspond to the cofactors

$$c_1(\mathbf{x}) = x + 5y, \quad c_2(\mathbf{x}) = x - y, \quad c_3(\mathbf{x}) = 4x + 2y. \tag{8.2.3}$$

The system is discretised using Ralston's method with a time step of $h = 0.001$ and the phase portrait on the square $[-10, 10]^2$ is presented in figure 8.2.1a. Here, the level sets $p_i(\mathbf{x}) = 0$ for $i = 1, 2, 3$ are represented by blue dashed lines. We see that numerical solutions starting on one of these zero level sets remain on the level set. This is exemplified in figure 8.2.1b which shows the errors $p_i(\mathbf{x}_n) - p_i(\mathbf{x}_0)$ for the numerical solutions starting from $p_i(\mathbf{x}_0) = 0$ and for $i = 1, 2, 3$. Here, we see that the errors are all within machine precision, implying that these three second integrals are preserved by the Ralston method. ▶

Second integrals are important as they divide phase space into sections with qualitatively different behavior. We see that Ralston's method has preserves the Darboux polynomials $p_i(\mathbf{x})$, meaning that it produces a qualitatively similar phase portrait. This fact is due to the following theorem.

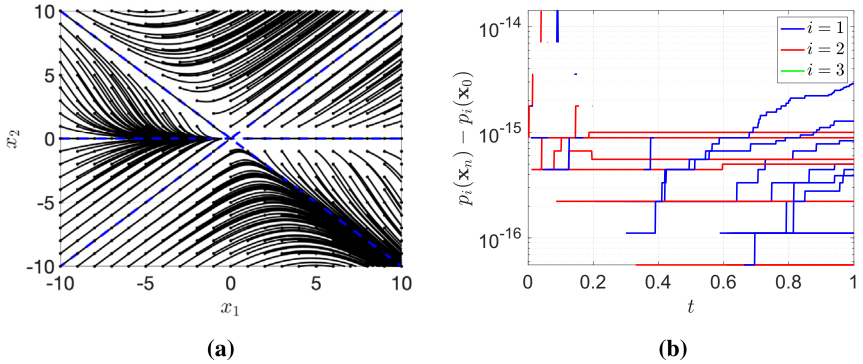


Figure 8.2.1: The phase portrait of the ODE (8.2.1) and the errors of the second integrals $p_i(\mathbf{x})$ for initial conditions satisfying $p_i(\mathbf{x}_0) = 0$. Note that $p_3(\mathbf{x}_n) - p_3(\mathbf{x}_0) = 0$ and therefore does not show on the semi-log axis. The initial conditions are shown by black dots and are located on the grid $(-10+i, -10+j)$ for $i, j = 0, ..., 20$.

Theorem 8.1. *If an autonomous ODE* $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ *possesses an affine second integral* $p(\mathbf{x}) = \mathbf{p}^T\mathbf{x}$ *with cofactor* $c(\mathbf{x})$ *satisfying* $\mathbf{p}^T\mathbf{f} = c(\mathbf{x})\mathbf{p}^T\mathbf{x}$ *then a Runge-Kutta map* φ_h *of the ODE possesses the discrete second integral* $\mathbf{p}^T\mathbf{x}$ *that satisfies* $\mathbf{p}^T\varphi_h(\mathbf{x}) = \tilde{c}(\mathbf{x})\,\mathbf{p}^T\mathbf{x}$ *where the discrete cofactor is given by*

$$\tilde{c}(\mathbf{x}) = 1 + h\boldsymbol{b}^T D_c (I - h\mathscr{A}D_c)^{-1}\mathbb{1}_s \tag{8.2.4}$$

and $D_c := \mathrm{diag}([c(\boldsymbol{g}_1), ..., c(\boldsymbol{g}_s)]) \in \mathbb{R}^{s\times s}$.

Proof. Let φ_h denote the Runge-Kutta map defined by equations (8.1.4), (8.1.5) and (8.1.6). Now let $G := (\boldsymbol{g}_1, ..., \boldsymbol{g}_s)^T$ and $F := (\mathbf{f}(\boldsymbol{g}_i), ..., \mathbf{f}(\boldsymbol{g}_s))^T$ denote the $s \times n$ matrices whose i'th rows are $\boldsymbol{g}_i^T$ and $\mathbf{f}_i^T$, respectively. Then

$$p(\varphi_h(\mathbf{x})) = \mathbf{p}^T\mathbf{x} + h\sum_{j=1}^{s} b_j\mathbf{p}^T\mathbf{f}(\boldsymbol{g}_j) = \mathbf{p}^T\mathbf{x} + h\boldsymbol{b}^T F\mathbf{p} = \mathbf{p}^T\mathbf{x} + h\boldsymbol{b}^T D_c G\mathbf{p} \tag{8.2.5}$$

We have for $G\mathbf{p}$ the following

$$G\mathbf{p} = \mathbb{1}_s\mathbf{p}^T\mathbf{x} + h\mathscr{A}F\mathbf{p} = \left(I - h\mathscr{A}D_c\right)^{-1}\mathbb{1}_s\mathbf{p}^T\mathbf{x} \tag{8.2.6}$$

due to the fact that $F\mathbf{p} = D_c G\mathbf{p}$. Inserting (8.2.6) into (8.2.5) and dividing by $\mathbf{p}^T\mathbf{x}$ we arrive at the desired result

$$\frac{\mathbf{p}^T\varphi_h(\mathbf{x})}{\mathbf{p}^T\mathbf{x}} = \tilde{c} = 1 + h\boldsymbol{b}^T D_c (I - h\mathscr{A}D_c)^{-1}\mathbb{1}_s. \tag{8.2.7}$$

□

This is a generalisation of theorem 1 in [1]. The discrete cofactor $\tilde{c}(\mathbf{x})$ of theorem 8.1 depends only on the Butcher table coefficients, the vector field $\mathbf{f}(\mathbf{x})$ and the (continuous) cofactor $c(\mathbf{x})$. Furthermore, $\tilde{c}(\mathbf{x})$ is in general rational and implicitly defined due the dependence of D_c on $\boldsymbol{g}_i$. However, explicit Runge-Kutta maps applied to polynomial ODEs yield polynomial maps and one would therefore expect $\tilde{c}$ to be known explicitly and be polynomial.

Remark 8.2. For all explicit Runge-Kutta methods, $\tilde{c}(\mathbf{x})$ is polynomial and can be written explicitly. This can be shown by observing that the matrix $I - h\mathscr{A}D_c = I + L$ where $L := -h\mathscr{A}D_c$ is strictly lower triangular and therefore $(I+L)^{-1} = I + \tilde{L}$ where $\tilde{L}$ is also strictly lower triangular. Moreover, as $\det(I+L) = 1$, its inverse is equal to its adjugate, that is $(I+L)^{-1} = \operatorname{adj}(I+L) = I + \tilde{L}$ and therefore $\tilde{L}$ is polynomial in the components of L.

An important implication of theorem 8.1 is that if an ODE possesses two linear second integrals with the same cofactor, then so does φ_h. This leads to the following corollary about the preservation of rational integrals.

Corollary 8.3. All Runge-Kutta methods preserve rational first integrals of the form $H(\mathbf{x}) = Q(\mathbf{x})/R(\mathbf{x})$ for $Q(\mathbf{x})$ and $R(\mathbf{x})$ affine.

Proof. Any ODE with first integral $H(\mathbf{x}) = Q(\mathbf{x})/R(\mathbf{x})$ can be written as the following system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) = S(\mathbf{x})\nabla\left(\frac{Q(\mathbf{x})}{R(\mathbf{x})}\right) \tag{8.2.8}$$

for some skew-symmetric matrix $S(\mathbf{x}) = -S(\mathbf{x})^T$. Without loss of generality we can let $Q(\mathbf{x}) = \mathbf{q}^T\mathbf{x}$ and $R(\mathbf{x}) = \boldsymbol{r}^T\mathbf{x}$ for constant vectors $\mathbf{q}, \boldsymbol{r} \in \mathbb{R}^n$. Then computing their time derivatives gives

$$\frac{\mathrm{d}}{\mathrm{d}t}Q(\mathbf{x}) = \mathbf{q}^T\mathbf{f} = \frac{\mathbf{q}^T S(\mathbf{x})\left(\mathbf{q}^T\mathbf{x}\boldsymbol{r} - \boldsymbol{r}^T\mathbf{x}\mathbf{q}\right)}{R(\mathbf{x})^2} = \frac{\mathbf{q}^T S(\mathbf{x})\boldsymbol{r}}{R(\mathbf{x})^2}Q(\mathbf{x}) \tag{8.2.9}$$

due to skew-symmetry of $S(\mathbf{x})$. Similarly, we can show that

$$\frac{\mathrm{d}}{\mathrm{d}t}R(\mathbf{x}) = \frac{\mathbf{q}^T S(\mathbf{x})\boldsymbol{r}}{R(\mathbf{x})^2}R(\mathbf{x}) \tag{8.2.10}$$

that is, $Q(\mathbf{x})$ and $R(\mathbf{x})$ are second integrals with cofactor $\frac{\mathbf{q}^T S(\mathbf{x})\boldsymbol{r}}{R(\mathbf{x})^2}$. Due to theorem 8.1, all Runge-Kutta methods preserve these second integrals and they both correspond to the same discrete cofactor. Therefore their quotient is an integral of the Runge-Kutta map. □

Note that corollary 8.3 applies to general ODEs. The same statement can me made for polynomial ODEs by scaling (8.2.8) by $R(x)^a$, $a \geq 2$ and $a \in \mathbb{N}$. We now give two examples of explicit Runge-Kutta methods preserving a rational integral. The first is of a simple Lotka-Volterra system, the second is of a non-polynomial vector field.

Example 5 (A Lotka-Volterra system with a rational integral)**.** Consider the following 2D Lotka-Volterra system

$$\dot{x} = x(x-y), \tag{8.2.11}$$

$$\dot{y} = y(x-y). \tag{8.2.12}$$

x and y are clearly Darboux polynomials with cofactor $c = x - y$ implying that $H(\mathbf{x}) = \frac{x}{y}$ is a first integral of the ODE. Now consider the generic second order explicit Runge-Kutta map φ_h defined by the Butcher Tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \theta & \theta & 0 \\ \hline & 1-\frac{1}{2\theta} & \frac{1}{2\theta} \end{array} \tag{8.2.13}$$

Note that setting $\theta = 1/2, 2/3, 1$ yields the explicit midpoint, Ralston's and Heun's method respectively. Then according to theorem 8.1 the discrete polynomial cofactor is

$$\tilde{c}(\mathbf{x}) = 1 + \left(x_1 - x_2\right) h + \left(x_1 - x_2\right)^2 h^2 + \frac{\theta}{2} \left(x_1 - x_2\right)^3 h^3. \tag{8.2.14}$$

Indeed we can show that the Runge-Kutta map satisfies

$$\varphi_h(\mathbf{x}) = \tilde{c}(\mathbf{x})\mathbf{x} \tag{8.2.15}$$

meaning that x and y are discrete Darboux polynomials of φ_h with cofactor $\tilde{c}$. This implies that $H(\mathbf{x}) = \frac{x}{y}$ is a first integral of φ_h. ▶

Example 6 (A non-polynomial ODE with a rational integral)**.** Consider the following ODE

$$\dot{x} = \frac{2x + \alpha}{\left(x-y\right)^{\frac{3}{2}}}, \quad \dot{y} = \frac{2y + \alpha}{\left(x-y\right)^{\frac{3}{2}}} \tag{8.2.16}$$

which has the following two second integrals

$$p_1(\mathbf{x}) = x + y + \alpha, \quad p_2(\mathbf{x}) = x - y \tag{8.2.17}$$

that both correspond to the cofactor

$$c(\mathbf{x}) = \frac{2}{\left(x-y\right)^{3/2}}. \tag{8.2.18}$$

This means that $H(\mathbf{x}) = (x + y + \alpha)/(x - y)$ is a first integral of the ODE. Now apply to this ODE the generic second-order explicit Runge-Kutta map φ_h from example 5. Then φ_h possesses the discrete second integrals $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ with cofactor

$$\tilde{c}(\mathbf{x}) = \frac{\left(h\sqrt{x-y} + a(\mathbf{x})\left(\left((x-y)^{\frac{3}{2}} + 2h\right)\theta - h\right)\right)}{a(\mathbf{x})\theta\left(x-y\right)^{\frac{3}{2}}} \tag{8.2.19}$$

where

$$a(\mathbf{x}) = \sqrt{\left((x-y)^{\frac{1}{2}} + 2h\theta\left(x-y\right)^{-\frac{1}{2}}\right)}, \tag{8.2.20}$$

in agreement with theorem 8.1 hence $H(\mathbf{x})$ is an integral of φ_h. We can also verify by direct computation that φ_h satisfies $H(\varphi_h(\mathbf{x})) = H(\mathbf{x})$. ▶

8.2.1 Constant cofactor case

We now restrict our discussion to the special case of affine Darboux polynomials with constant cofactor

$$\dot{p}(\mathbf{x}) = \lambda p(\mathbf{x}), \tag{8.2.21}$$

where λ is constant. If such a Darboux polynomial exists then we can solve for $p(\mathbf{x})$

$$p(\mathbf{x}) = Ke^{\lambda t}, \tag{8.2.22}$$

for some arbitrary constant K, and therefore

$$H = p(\mathbf{x})e^{-\lambda t} \tag{8.2.23}$$

is a time-dependent integral of the ODE (8.1.1).

In the discrete-time case, $p(\varphi_h(\mathbf{x})) = \tilde{c}p(\mathbf{x})$ implies that

$$p(\mathbf{x}) = K\tilde{c}^k, \tag{8.2.24}$$

where k is the iteration index and

$$\tilde{H} = \tilde{c}^{-k}p(\mathbf{x}), \tag{8.2.25}$$

is an integral of the map $\varphi_h^{\circ k}$. We now consider the case where the map φ_h comes about as a Runge-Kutta method applied to an ODE that possesses one or more Darboux polynomials of the form (8.2.21).

We begin with an example of a Lotka-Volterra system with a time-dependent integral and its discretisation under a Runge-Kutta method.

Example 7 (An ODE with one time-dependent integral)**.** Consider the following Lotka-Volterra system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1\left(a_1x_2 - a_2x_3 + b\right) \\ x_2\left(-a_1x_1 + a_3x_3 + b\right) \\ x_3\left(a_2x_1 - a_3x_2 + b\right) \end{pmatrix} \tag{8.2.26}$$

Then $p_1(x) = x_1 + x_2 + x_3$ is a Darboux polynomial corresponding to the cofactor $c_1 = b$, hence

$$H = p_1(x)e^{-bt} \tag{8.2.27}$$

is a time-dependent integral of the ODE (8.2.26). Now consider discretisation of the above ODE by a Runge-Kutta method φ_h with stability function $R(z)$. Then, $p_1(x)$ is also a discrete Darboux polynomial of a with the cofactor $\tilde{c}_1 = R(bh)$, hence

$$\tilde{H} = R(bh)^m p_1(x) \tag{8.2.28}$$

where m is the iteration index, is a step-dependent integral of φ_h. ►

We note that this example holds for any Runge-Kutta method φ_h. This is due to the following corollary due to theorem 8.1.

Corollary 8.4. If an autonomous polynomial ODE $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ possesses an affine Darboux polynomial $p(\mathbf{x})$ with a constant cofactor λ, then a Runge-Kutta map φ_h with stability function $R(z)$ satisfies

$$p(\varphi_h(\mathbf{x})) = R(\lambda h)p(\mathbf{x}) \tag{8.2.29}$$

where

$$R(z) = 1 + z\boldsymbol{b}^T(I - z\mathscr{A})^{-1}\mathbb{1} \tag{8.2.30}$$

is the (constant) discrete cofactor, $\mathbb{1}$ is the ones vector and I is the $s \times s$ identity matrix.

Proof. This can be seen by setting $c(\mathbf{x}) = \lambda$ in theorem 8.1. However, we note that as Runge-Kutta methods are affinely equivariant [8], we can take $p(x) = x_1$ without loss of generality, then equation (8.2.21) is identical to Dahlquist's famous test ODE [10] and $R(z)$ is identical to the the stability function in the context of A-stability of one-step methods. The rest follows from e.g., [10, Proposition 3.1]. See appendix 8.A for details. □

As the existence of Darboux polynomials with constant cofactors implies the existence of time-dependent integrals of an ODE and therefore iteration-index-dependent integrals of a Runge-Kutta map in the discrete-time case. One can eliminate this time dependence (and iteration index dependence) by taking quotients. This leads to the preservation of non-rational modified integrals by Runge-Kutta methods.

Corollary 8.5. Given an ODE with a first integral given by

$$H(\mathbf{x}) = \frac{p_1(\mathbf{x})^\sigma}{p_2(\mathbf{x})} \tag{8.2.31}$$

where $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ are affine Darboux polynomials with constant cofactors c_1 and c_2 and

$$\sigma = \frac{c_2}{c_1}, \tag{8.2.32}$$

then any Runge-Kutta map φ_h with stability function $R(z)$ preserves the modified integral

$$\tilde{H}(\mathbf{x}) = \frac{p_1(\mathbf{x})^{\tilde{\sigma}}}{p_2(\mathbf{x})} \tag{8.2.33}$$

with

$$\tilde{\sigma} = \frac{\ln(R(hc_2))}{\ln(R(hc_1))}. \tag{8.2.34}$$

Proof. According to corollary 8.4, φ_h preserves the Darboux polynomials $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ with the modified cofactors $R(hc_1)$ and $R(hc_2)$. It follows that $\tilde{H}(\mathbf{x})$ is also a Darboux polynomial of φ_h with cofactor 1, and hence is a first integral of φ_h. □

This is demonstrated by the following example.

Example 8 (An ODE with a non-rational integral)**.** Consider the following ODE in three dimensions

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1x_2 + x_2x_3 + x_1 \\ -x_1x_2 - x_2x_3 + x_2 \\ (\sigma - 1)\left(x_1 + x_2\right) + \sigma x_3 \end{pmatrix} \tag{8.2.35}$$

where $\sigma \in \mathbb{R}$. Discretising the above ODE by a Runge-Kutta method φ_h with stability function $R(z)$, we find the following affine Darboux polynomials and their corresponding discrete constant cofactors

	$p_i(x)$	$\tilde{c}_i(x)$
$i = 1$	$x_1 + x_2$	$R(h)$
$i = 2$	$x_1 + x_2 + x_3$	$R(\sigma h)$

(8.2.36)

Therefore

$$\tilde{H} = \frac{(x_1 + x_2)^{\tilde{\sigma}}}{x_1 + x_2 + x_3} \tag{8.2.37}$$

defines an h-dependent integral of φ_h with

$$\tilde{\sigma} = \frac{\ln\left(R(\sigma h)\right)}{\ln\left(R(h)\right)}. \tag{8.2.38}$$

We can take the continuum limit $\lim_{h\to 0}(\tilde{\sigma}) = \sigma$, which implies that $\lim_{h\to 0}\tilde{H}(x) = H(x)$, where

$$H(x) = \frac{(x_1+x_2)^{\sigma}}{x_1+x_2+x_3}. \tag{8.2.39}$$

Indeed, the irrational integral $H(x)$ is preserved by the flow of the original ODE. ►

8.3 ODEs with affine higher integrals

The notion of first, second, third* and higher integrals can be generalized by considering solutions of the linear system for $\underline{\mathbf{p}}(\mathbf{x}) = (p_1(\mathbf{x}), p_2(\mathbf{x}), ..., p_m(\mathbf{x}))^T \in \mathbb{R}^m$ satisfying [6]

$$\dot{\underline{\mathbf{p}}}(\mathbf{x}) = L(\mathbf{x})\underline{\mathbf{p}}(\mathbf{x}) \tag{8.3.1}$$

where $L(\mathbf{x}) \in \mathbb{R}^{n\times n}$. If $L_{ij} = 0$ for $j = 1, ..., n$ then $p_i(\mathbf{x})$ is a first integral and is preserved from arbitrary initial conditions. If $L_{ii} \neq 0$ and $L_{ij} = 0$ for $j \neq i$ then $p_i(\mathbf{x})$ is a Darboux polynomial with cofactor L_{ii} and is preserved for initial conditions that begin on its zero level sets. If, for example, $L_{ik} \neq 0$ and $L_{ij} = 0$ for $j \neq i, k$ then $p_i(\mathbf{x})$ is preserved for initial conditions starting on the zero level sets of $p_k(\mathbf{x})$. Here, if $p_k(\mathbf{x})$ is a first integral, then $p_i(\mathbf{x})$ is called a third integral. In this sense, one can define the notion of higher integrals for ODE systems. In this section we consider Runge-Kutta methods applied to ODEs with a sub-system of affine higher integrals with constant coefficient matrix L.

Corollary 8.6. Consider an autonomous ODE in n dimensions that possesses a system of $m \leq n$ linearly-independent affine polynomials $\underline{\mathbf{p}}(\mathbf{x}) = (p_1(\mathbf{x}), ..., p_m(\mathbf{x}))^T \in \mathbb{R}^m$ satisfying

$$\dot{\underline{\mathbf{p}}}(\mathbf{x}) = L\underline{\mathbf{p}}(\mathbf{x}) \tag{8.3.2}$$

where L is a $m \times m$ matrix that is independent of $\mathbf{x}$. Then a Runge-Kutta map φ_h satisfies

$$\underline{\mathbf{p}}(\varphi_h(\mathbf{x})) = R(hL)\underline{\mathbf{p}}(\mathbf{x}) \tag{8.3.3}$$

where $R(z)$ is the stability function of φ_h.

Proof. Let $\underline{\mathbf{p}}(\mathbf{x}) = D\mathbf{x}$, where D is an $m \times n$ matrix of rank $m \leq n$. The numerical solution of φ_h applied to a linear ODE in n dimensions $\dot{\underline{\mathbf{p}}} = L\underline{\mathbf{p}}$ is given by (see for example, [7, p. 194])

$$\varphi_h(\underline{\mathbf{p}}) = R(hL)\underline{\mathbf{p}}. \tag{8.3.4}$$

As Runge-Kutta methods commute with linear transformations we have $\varphi_h(\underline{\mathbf{p}}(\mathbf{x})) = \underline{\mathbf{p}}(\varphi_h(\mathbf{x}))$, which yields the desired result. □

*A third integral is a function $K(\mathbf{x})$ that is preserved on a particular level set of a first integral $H(\mathbf{x})$, e.g., $\dot{K}(\mathbf{x}) = c(\mathbf{x})H(\mathbf{x})$

We remark that an ODE in n dimensions cannot possess more than n functionally independent affine Darboux polynomials with constant cofactor. We demonstrate corollary 8.6 in an example where L is given in Jordan form. But first we will will briefly introduce a particular class of Runge-Kutta methods called diagonal Padé Runge-Kutta methods.

Definition 8.1. A *diagonal Padé Runge-Kutta map* is an s-stage Runge-Kutta map φ_h whose stability function $R(z) = \frac{P(z)}{Q(z)}$ has equal degree in the numerator and denominator, that is $\deg(P(z)) = \deg(Q(z))$. Such a stability function is an order s approximation to e^z with numerator and denominator given by

$$P(z) = \sum_{i=0}^{s} a_i z^i \tag{8.3.5}$$

$$Q(z) = \sum_{i=0}^{s} (-1)^i a_i z^i = P(-z) \tag{8.3.6}$$

where a_i are the constants

$$a_i = \frac{s!(2s-i)!}{(2s)!i!(s-i)!}. \tag{8.3.7}$$

See, for example, [9] and references therein. Such a map has a stability function that satisfies $R(-z)R(z) = 1$.

As an example of some well known Diagonal Padé Runge-Kutta methods, consider the Runge-Kutta map $\varphi_h(\mathbf{x}) = \mathbf{x}'$ defined by

$$\frac{\mathbf{x}'-\mathbf{x}}{h} = (1-2\theta)\mathbf{f}(\frac{\mathbf{x}+\mathbf{x}'}{2}) + \theta\mathbf{f}(\mathbf{x}') + \theta\mathbf{f}(\mathbf{x}) \tag{8.3.8}$$

then its stability function is the diagonal Padé approximation

$$R(\lambda h) = \frac{1+\frac{1}{2}\lambda h}{1-\frac{1}{2}\lambda h}. \tag{8.3.9}$$

Note that the three cases $\theta = 0$, $\theta = \frac{1}{2}$ and $\theta = -\frac{1}{2}$ respectively correspond to the midpoint rule, trapezoidal rule and Kahan's method, the latter being when $\mathbf{f}(\mathbf{x})$ is quadratic [4].

Example 9 (A 3D ODE with a 2D linear subsystem in Jordan form)**.** Consider the following quadratic ODE in 3 dimensions

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} x_1x_2 + x_2x_3 + \sigma x_2 \\ -x_1x_2 - x_2x_3 + \sigma x_1 \\ \left(x_1 + x_2\right) + \sigma x_3 \end{pmatrix} \tag{8.3.10}$$

Setting $\underline{\mathbf{p}}(\mathbf{x}) = (x_1 + x_2 + x_3, x_1 + x_2)^T$ then the above ODE has the following linear subsystem

$$\dot{\underline{\mathbf{p}}}(\mathbf{x}) = \begin{pmatrix} \sigma & 1 \\ 0 & \sigma \end{pmatrix} \underline{\mathbf{p}}(\mathbf{x}) := L\underline{\mathbf{p}}(\mathbf{x}) \tag{8.3.11}$$

If we apply Kahan's method then we get

$$\underline{\mathbf{p}}(\varphi_h(\mathbf{x})) = \begin{pmatrix} \frac{-h\sigma-2}{h\sigma-2} & 4\,\frac{h}{(h\sigma-2)^2} \\ 0 & \frac{-h\sigma-2}{h\sigma-2} \end{pmatrix} \underline{\mathbf{p}}(\mathbf{x}) \tag{8.3.12}$$

$$= \begin{pmatrix} R(h\sigma) & R'(h\sigma) \\ 0 & R(h\sigma) \end{pmatrix} \underline{\mathbf{p}}(\mathbf{x}) = R(hL)\underline{\mathbf{p}}(\mathbf{x}), \tag{8.3.13}$$

which is the form prescribed in corollary 8.6.

▶

It is known that when the Kahan map is applied to a Hamiltonian ODE that it preserves a modified Hamiltonian [4] and the midpoint rule preserves quadratic first integrals. However, there exists some cases where the Kahan map (as well as other diagonal Padé Runge-Kutta maps) preserves a quadratic first integral exactly as we will show in the following example.

Example 10 (An ODE with a quadratic integral that is preserved exactly)**.** Consider the quadratic ODE in three dimensions

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2\,x_2 + x_1 + x_3 + x_3\left(x_1 + x_2\right) + {x_1}^2 \\ -x_1 - x_2 - x_3\left(x_1 + x_2\right) - {x_1}^2 \\ x_3\left(x_1 + x_2\right) + {x_1}^2 \end{pmatrix} \tag{8.3.14}$$

Setting $\underline{\mathbf{p}}(\mathbf{x}) = (x_1 + x_2, x_2 + x_3)^T$ then this satisfies

$$\dot{\underline{\mathbf{p}}}(\mathbf{x}) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \underline{\mathbf{p}}(\mathbf{x}) \tag{8.3.15}$$

If we apply the Kahan discretisation to the above ODE we find that $p_1(\mathbf{x}) = \|\underline{\mathbf{p}}(\mathbf{x})\|^2 = (x_1 + x_2)^2 + (x_2 + x_3)^2$ is a quadratic Darboux polynomial of φ_h with cofactor $\tilde{c} = 1$. Therefore $H(\mathbf{x}) = \|\underline{\mathbf{p}}(\mathbf{x})\|^2$ is an integral of the map, i.e., $H(\varphi_h(\mathbf{x})) = H(\mathbf{x})$. Moreover, we find that any diagonal Padé Runge-Kutta map preserves this integral. This is due to the following corollary. ▶

Corollary 8.7. Given an ODE $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, if there exists m linear polynomials $\underline{\mathbf{p}}(x) = (p_1(x), p_2(x), ..., p_m(x))^T$ that satisfy

$$\dot{\underline{\mathbf{p}}} = AS\underline{\mathbf{p}}(\mathbf{x}), \quad \text{where} \quad A = -A^T \quad \text{and} \quad S = S^T \tag{8.3.16}$$

then $H(x) = \underline{\mathbf{p}}(\mathbf{x})^T S\underline{\mathbf{p}}(\mathbf{x})$ is a first integral of the ODE and any diagonal Padé Runge-Kutta map (e.g., one of the form (8.3.8)) preserves this integral exactly.

Proof. Using the fact that $R(z) = P(z)P(-z)^{-1} = P(-z)^{-1}P(z)$, $P(hAS)^T = P(-hSA)$ and $P(hSA)S = SP(hAS)$ then it's straight forward to show that H is preserved under φ_h

$$\begin{aligned} H(\varphi_h(\mathbf{x})) =& \underline{\mathbf{p}}(\mathbf{x})^T R(hAS)^T SR(hAS)\underline{\mathbf{p}}(\mathbf{x}) && (8.3.17)\\ =& \underline{\mathbf{p}}(\mathbf{x})^T P(-hAS)^{-T} P(hAS)^T SP(hAS)P(-hAS)^{-1}\underline{\mathbf{p}}(\mathbf{x}) && (8.3.18)\\ =& \underline{\mathbf{p}}(\mathbf{x})^T P(hSA)^{-1} P(-hSA)SP(hAS)P(-hAS)^{-1}\underline{\mathbf{p}}(\mathbf{x}) && (8.3.19)\\ =& \underline{\mathbf{p}}(\mathbf{x})^T SP(hAS)^{-1} P(-hAS)P(-hAS)^{-1}P(hAS)\underline{\mathbf{p}}(\mathbf{x}) && (8.3.20)\\ =& \underline{\mathbf{p}}(\mathbf{x})^T S\underline{\mathbf{p}}(\mathbf{x}) && (8.3.21)\\ =& H(\mathbf{x}) && (8.3.22) \end{aligned}$$

□

We can therefore make the following statement about linear ODEs with quadratic integrals.

Corollary 8.8. For all linear ODEs with quadratic first integrals, all diagonal Padé Runge-Kutta maps preserve the integral exactly.

Proof. A linear ODE with a quadratic integral $H = \frac{1}{2}\mathbf{x}^T S\mathbf{x}$ can be written in the form

$$\dot{\mathbf{x}} = A\nabla H = AS\mathbf{x} \tag{8.3.23}$$

for $A = -A^T$ and $S = S^T$, which is in the form of (8.3.16) and according to corollary 8.7, all diagonal Padé Runge-Kutta maps preserve the integral H □

This is a slight generalisation of proposition 5 in [3].

Corollary 8.7 also applies to ODEs with multiple quadratic integrals as shown in the following example.

Example 11 (A 5D system with 2 quadratic integrals preserved exactly)**.** Consider the following ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1\\ x_2\\ x_3\\ x_4\\ x_5 \end{pmatrix} = \begin{pmatrix} x_3{}^2+2x_3x_4+2x_3x_5+2x_4{}^2+2x_4x_5+2x_5{}^2-10x_1-17x_2-7x_3-5x_4+4x_5\\ -x_3{}^2-2x_3x_4-2x_3x_5-2x_4{}^2-2x_4x_5-2x_5{}^2+6x_1+11x_2+5x_3+5x_4-3x_5\\ x_3{}^2+2x_3x_4+2x_3x_5+2x_4{}^2+2x_4x_5+2x_5{}^2-x_2-x_3-4x_4+2x_5\\ -x_1-2x_2-x_3+2x_4\\ x_1+3x_2+2x_3+2x_4-2x_5 \end{pmatrix} \tag{8.3.24}$$

Let $\lambda = 2+i$, where $i^2 = -1$ and discretise the ODE using Kahan's method, which is a diagonal Padé Runge-Kutta map. We find the following affine discrete Darboux polynomials with constant cofactor.

j	$p_j(x)$	$\tilde{c}_j$
1	$-ix_4 + x_1 + 2x_2 + x_3$	$R(\lambda h)$
2	$(106+52i)x_2 + 41x_3 - (52-24i)x_5 + (65+52i)x_1 + (12-15i)x_4$	$R(-\lambda h)$
3	$ix_4 + x_1 + 2x_2 + x_3$	$R(\lambda^* h)$
4	$(106-52i)x_2 + 41x_3 - (52+24i)x_5 + (65-52i)x_1 + (12+15i)x_4$	$R(-\lambda^* h)$

(8.3.25)

From the above four linear Darboux polynomials, we can construct the following complex quadratic integrals

$$K_1 = p_1(\mathbf{x})p_2(\mathbf{x}), \quad K_2 = p_3(\mathbf{x})p_4(\mathbf{x}), \tag{8.3.26}$$

however if we observe that $K_1 = K_2^*$ then the following two independent real integrals can be constructed

$$H_1 = K_1 + K_2, \quad H_2 = i(K_1 - K_2). \tag{8.3.27}$$

As these two integrals are independent of h, it follows that the ODE also possesses these integrals. ►

8.3.1 Detecting the existence of affine higher integrals

Many of our results in this section so far rely on the fact that there exist a change of coordinates $\underline{\mathbf{p}}(x) = Dx$ that allow us to find a linear ODE subsystem of higher integrals of the form $\dot{\underline{\mathbf{p}}} = L\underline{\mathbf{p}}$. A logical question to ask is what is the most general class of ODEs where this sub-system exists and how to calculate this transformation. This is now addressed.

Theorem 8.9. *If an ODE in n dimensions can be expressed in the following form*

$$\dot{\mathbf{x}} = A\mathbf{x} + \sum_{i=1}^{k} b_i(\mathbf{x})\mathbf{v}_i \tag{8.3.28}$$

for the scalar functions $b_i(\mathbf{x})$ and the matrix $V := [\mathbf{v}_1, \ldots, \mathbf{v}_k] \in \mathbb{R}^{n\times k}$ has rank $n-m$, where $m < n$, then there exists the linear transformations $\underline{\mathbf{p}} = Q\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} = R\mathbf{x} \in \mathbb{R}^{n-m}$ that decouple the ODE into an m dimensional linear ODE for $\underline{\mathbf{p}}$ and an $n-m$ dimensional non-linear ODE for $\mathbf{y}$

$$\dot{\underline{\mathbf{p}}} = L\underline{\mathbf{p}} \tag{8.3.29}$$

$$\dot{\mathbf{y}} = \mathbf{g}(\underline{\mathbf{p}}, \mathbf{y}) \tag{8.3.30}$$

(8.3.31)

for some matrix L.

Proof. Given an ODE in the form (8.3.28), then choose the linear transformation $\underline{\mathbf{p}} = Q\mathbf{x}$ such that $\mathbf{v}_i \in \ker(Q)$. Multiplying the ODE (8.3.28) by Q gives

$$Q\dot{\mathbf{x}} = QA\mathbf{x} + \sum_{i=1}^{k} b_i(\mathbf{x})Q\mathbf{v}_i = QA\mathbf{x} = \dot{\underline{\mathbf{p}}}, \tag{8.3.32}$$

which gives the form of the decoupled linear part of the ODE (8.3.29). Now we need to construct the $n-m$ non-linear part of the ODE (8.3.30). To do this, we simply choose some matrix $R \in \mathbb{R}^{(n-m)\times n}$ whose rows are independent to the rows in Q. This then defines the one-to-one transformation

$$\mathbf{z} := \begin{pmatrix} \underline{\mathbf{p}} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} Q \\ R \end{pmatrix} \mathbf{x} := G\mathbf{x} \tag{8.3.33}$$

Then the transformed ODE

$$\dot{\mathbf{z}} = G\mathbf{f}(\mathbf{z})$$

is in the desired form. □

A logical question now is how does one find such a linear transformation. We will address this now by presenting a systematic algorithm to transform an ODE given in the form (8.3.28) into its decoupled form (8.3.29) and (8.3.30).

Algorithm 8.1. Given an ODE $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$

1. Let $i = 1$.

2. Solve the condition $\nabla p_i(\mathbf{x})^T \mathbf{f}(\mathbf{x}) = \lambda p_i(\mathbf{x})$ for affine $p_i(\mathbf{x})$ and constant λ_i.

   (a) If there is no solutions, or they have all been used in successive steps, then end the algorithm.

   (b) If there are one or more solutions, pick one, set $i \to i+1$ and move to the next step.

3. Set $p_{i-1}(\mathbf{x}) = 0$ in $\mathbf{f}$ and solve the condition $\nabla p_i(\mathbf{x})^T \mathbf{f}(\mathbf{x})\Big|_{p_{i-1}(\mathbf{x})=0} = \lambda_i p_i(\mathbf{x})$ for $p_i(\mathbf{x})$ and λ_i.

   (a) If there is a solution, then set $i \to i+1$ and repeat step 3

   (b) If there is no solution, go back to step 2 and pick a different Darboux polynomial solution

4. Calculate $Q = \nabla\underline{\mathbf{p}}(\mathbf{x})$

5. Calculate L from $\nabla(Q.f) = LQ$

6. Choose a matrix R s.t.

$$G = \begin{pmatrix} Q \\ R \end{pmatrix} \tag{8.3.34}$$

has full rank. Then the ODE

$$\dot{\mathbf{z}} = G\mathbf{f}(\mathbf{z}) \tag{8.3.35}$$

is in the desired form. End algorithm.

We will now demonstrate this algorithm with an example.

Example 12 (Example of algorithm 8.1)**.** Consider the following vector field in five dimensions

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} -2\,x_1x_3 + 5\,{x_2}^2 + {x_4}^2 - 2\,{x_5}^2 + x_2 + x_5 \\ \left(-2\,x_3 + 15\right)x_1 + 5\,{x_2}^2 + {x_4}^2 - 2\,{x_5}^2 + 6\,x_2 + 12\,x_3 + 8\,x_4 + x_5 \\ \left(2\,x_3 + 2\right)x_1 - 5\,{x_2}^2 - {x_4}^2 + 2\,{x_5}^2 + x_3 + 2\,x_4 - x_5 \\ \left(2\,x_3 - 7\right)x_1 - 5\,{x_2}^2 - {x_4}^2 + 2\,{x_5}^2 - 3\,x_2 - 6\,x_3 - 3\,x_4 - x_5 \\ x_1x_3 - 2\,{x_2}^2 + {x_5}^2 \end{pmatrix} \tag{8.3.36}$$

We will now implement algorithm 1 to decouple this ODE into a set of linear and non-linear equations. First look for an affine Darboux polynomial that has a constant cofactor. We find the following Darboux polynomial

$$p_1 = x_1 + x_2 + 2\,x_4 \tag{8.3.37}$$

with cofactor 1 is the only affine Darboux polynomial that has constant cofactor. Now eliminate a variable from the vector field by setting $p_1 = 0$, for example, by substituting $x_1 = -x_2 - 2x_4$ into $\mathbf{f}$.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \left(-2\,x_3 + 15\right)\left(-x_2 - 2\,x_4\right) + 5\,{x_2}^2 + {x_4}^2 - 2\,{x_5}^2 + 6\,x_2 + 12\,x_3 + 8\,x_4 + x_5 \\ \left(2\,x_3 + 2\right)\left(-x_2 - 2\,x_4\right) - 5\,{x_2}^2 - {x_4}^2 + 2\,{x_5}^2 + x_3 + 2\,x_4 - x_5 \\ \left(2\,x_3 - 7\right)\left(-x_2 - 2\,x_4\right) - 5\,{x_2}^2 - {x_4}^2 + 2\,{x_5}^2 - 3\,x_2 - 6\,x_3 - 3\,x_4 - x_5 \\ \left(-x_2 - 2\,x_4\right)x_3 - 2\,{x_2}^2 + {x_5}^2 \end{pmatrix} \tag{8.3.38}$$

We now look for Darboux polynomials with constant cofactor on the resulting four dimensional vector field and find

$$p_2 = -x_2 + x_3 - 2\,x_4 \tag{8.3.39}$$

also with cofactor 1. We now substitute $x_2 = x_3 - 2x_4$ and compute constant cofactor Darboux polynomials on the resulting three dimensional system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} -\left(2\,x_3 + 2\right)x_3 - 5\left(x_3 - 2\,x_4\right)^2 - {x_4}^2 + 2\,{x_5}^2 + x_3 + 2\,x_4 - x_5 \\ -\left(2\,x_3 - 7\right)x_3 - 5\left(x_3 - 2\,x_4\right)^2 - {x_4}^2 + 2\,{x_5}^2 - 9\,x_3 + 3\,x_4 - x_5 \\ -{x_3}^2 - 2\left(x_3 - 2\,x_4\right)^2 + {x_5}^2 \end{pmatrix} \tag{8.3.40}$$

this vector field has the polynomial $p_3 = x_4 - x_3$ with cofactor 1. Substituting $x_3 = x_4$ into the above three dimensional vector field, we find that there are no more affine Darboux polynomials with constant cofactor. Setting $\underline{\mathbf{p}}(\mathbf{x}) = (p_1(x), p_2(x), p_3(x))^T$, we can write $\underline{\mathbf{p}}(\mathbf{x}) = Q\mathbf{x}$, where

$$Q = \begin{bmatrix} 1 & 1 & 0 & 2 & 0 \\ 0 & -1 & 1 & -2 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{bmatrix} \tag{8.3.41}$$

It must therefore be possible to write the original ODE as the following decoupled set of ODEs

$$\dot{\underline{\mathbf{p}}} = L\underline{\mathbf{p}} \tag{8.3.42}$$

$$\dot{\mathbf{y}} = \mathbf{g}(\mathbf{y}, \underline{\mathbf{p}}) \tag{8.3.43}$$

where L is triangular. Multiplying the ODE $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ by Q, we get the identity $Q\mathbf{f}(\mathbf{x}) = L\underline{\mathbf{p}}$. In other words, $Q\mathbf{f}(\mathbf{x}) = B\mathbf{x}$ must be linear. The coefficients of the matrix L can therefore be easily found by solving the linear problem $B = LQ$. Doing so, we find

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -9 & -6 & 1 \end{bmatrix} \tag{8.3.44}$$

We can now define the linear transformation

$$\begin{pmatrix} p \\ y \end{pmatrix} = \begin{bmatrix} Q \\ R \end{bmatrix} \mathbf{x} \tag{8.3.45}$$

where R is any $2 \times n$ matrix whose rows are independent to the rows of Q. We will choose $R = [0, I_2]$. Hence by the above linear transformation, the ODE reads

$$\dot{\underline{\mathbf{p}}} = L\underline{\mathbf{p}}, \tag{8.3.46}$$

$$\dot{y}_1 = 2p_1p_3 - 5p_2{}^2 - y_1{}^2 + 2y_2{}^2 - 7p_1 - 3p_2 - 6p_3 - 3y_4 - y_5, \tag{8.3.47}$$

$$\dot{y}_2 = p_1p_3 - 2p_2{}^2 + y_2{}^2. \tag{8.3.48}$$

►

Acknowledgments

We thank GRW Quispel for many useful comments and suggestions.

Bibliography

[1] E. Celledoni, C. Evripidou, D. McLaren, B. Owren, G. Quispel, B. Tapley, and P. van der Kamp, *Using discrete darboux polynomials to detect and determine preserved measures and integrals of rational maps*, Journal of Physics A: Mathematical and Theoretical, 52 (2019), p. 31LT01.

[2] E. Celledoni, C. Evripidou, D. McLaren, B. Owren, R. Quispel, and B. Tapley, *Discrete darboux polynomials and the search for preserved measures and integrals of rational maps*, arXiv preprint arXiv:1902.04685, (2019).

[3] E. Celledoni, R. I. McLachlan, D. I. McLaren, B. Owren, G. R. W. Quispel, and W. M. Wright, *Energy-preserving runge-kutta methods*, ESAIM: Mathematical Modelling and Numerical Analysis, 43 (2009), pp. 645–649.

[4] E. Celledoni, R. I. McLachlan, B. Owren, and G. R. W. Quispel, *Geometric properties of kahan's method*, Journal of Physics A: Mathematical and Theoretical, 46 (2012), p. 025201.

[5] C. Collins, *Algebraic conditions for a centre or a focus in some simple systems of arbitrary degree*, Journal of mathematical analysis and applications, 195 (1995), pp. 719–735.

[6] A. Goriely, *Integrability and nonintegrability of dynamical systems*, vol. 19, World Scientific, 2001.

[7] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, vol. 31, Springer Science & Business Media, 2006.

[8] R. I. McLachlan, G. Quispel, et al., *Six lectures on the geometric integration of ODEs*, Citeseer, 1998.

[9] M. F. Reusch, L. Ratzan, N. Pomphrey, and W. Park, *Diagonal padé approximations for initial value problems*, SIAM journal on scientific and statistical computing, 9 (1988), pp. 829–838.

[10] G. Wanner and E. Hairer, *Solving ordinary differential equations II*, Springer Berlin Heidelberg, 1996.

Appendix

8.A Derivation of stability matrix for corollary 8.4

Let φ_h denote the Runge-Kutta map defined by equations (8.1.4), (8.1.5) and (8.1.6). Now let $G = (g_1, g_2, ..., g_s)^T$ denote the $s \times n$ matrix whose i'th row is g_i^T. Then if $f(x) = \lambda x$ as in the case of our test equation (8.1.5) can be written as

$$G = \mathbb{1}x^T + h\lambda\mathscr{A}G = (I - h\lambda\mathscr{A})^{-1}\mathbb{1}x^T \tag{8.A.1}$$

We therefore have

$$g_i = G^T\hat{e}_i = \left(\hat{e}_i{}^T(I - h\lambda\mathscr{A})^{-1}\mathbb{1}\right)x \tag{8.A.2}$$

where $\hat{e}_i \in \mathbb{R}^s$ is the i'th canonical unit basis vector with a 1 in the i'th component and 0 elsewhere. Inserting this into equation (8.1.6) gives

$$\begin{aligned} \varphi_h(x) =& x + h\lambda\sum_{j=1}^{s} b_j\hat{e}_i{}^T(I - h\lambda\mathscr{A})^{-1}\mathbb{1}x & \text{(8.A.3)} \\ =& (1 + h\lambda b^T(I - h\lambda\mathscr{A})^{-1}\mathbb{1})x & \text{(8.A.4)} \\ :=& R(\lambda h)x & \text{(8.A.5)} \end{aligned}$$

Symplectic integration of PDEs using Clebsch variables

Christian Offen, Robert I McLachlan and Benjamin K Tapley

Journal of Computational Dynamics 6.1, pp. 111–130. issn: 2158-2505.
doi: 10.3934/jcd.2019005. arXiv: 1810.01627

Symplectic integration of PDEs using Clebsch variables

Abstract. Many PDEs (Burgers' equation, KdV, Camassa-Holm, Euler's fluid equations,...) can be formulated as infinite-dimensional Lie-Poisson systems. These are Hamiltonian systems on manifolds equipped with Poisson brackets. The Poisson structure is connected to conservation properties and other geometric features of solutions to the PDE and, therefore, of great interest for numerical integration. For the example of Burgers' equations and related PDEs we use Clebsch variables to lift the original system to a collective Hamiltonian system on a symplectic manifold whose structure is related to the original Lie-Poisson structure. On the collective Hamiltonian system a symplectic integrator can be applied. Our numerical examples show excellent conservation properties and indicate that the disadvantage of an increased phase-space dimension can be outweighed by the advantage of symplectic integration.

9.1 Motivation

Partial differential equations (PDEs) often exhibit interesting structure preserving properties, for example conserved quantities. In many examples, a deeper understanding of the structures can be achieved by viewing the PDE as the Lie-Poisson equation associated to an infinite-dimensional Lie group. This means solutions to the PDE correspond to motions of a Hamiltonian system defined on the dual of the Lie-algebra of a Fréchet Lie-group. Examples include Euler's equations for incompressible fluids, Burgers' equation, equations in magnetohydrodynamics, the Korteweg-de Vries equation, the superconductivity equation, charged ideal fluid equations, the Camassa-Holm equation and the Hunter-Saxton equation [16]. Conserved quantities turn out to be related to the fact that the Hamiltonian flow preserves the Lie-Poisson bracket. This makes Lie-Poisson structures interesting for structure preserving integration. We will give a brief review of Hamiltonian systems on Poisson manifolds in section 9.2.

An approach to construct Lie-Poisson integrators, which works universally in the finite-dimensional setting, is to translate the Lie-Poisson system on a Lie-group G to a Hamiltonian system on the tangent bundle TG with a G-invariant Lagrangian. Using a variational integrator one obtains a Poisson-integrator for the original system [12]. These integrators, however, can be extremely complicated [15, p. 1526]. Moreover, the fact that exponential maps do not constitute local diffeomorphisms for infinite-dimensional manifolds restricts the approach to a finite-dimensional setting. Other approaches for energy preserving integration of finite dimensional Poisson systems with good preservation properties, e.g. preservation of linear symmetries or (quadratic) Casimirs, include [1,3,4]. For a recent review article on Lie-Poisson integrators we refer to [5].

Let us return to the infinite-dimensional setting. For numerical computations a PDE needs to be discretised in space. In the Lie-Poisson setting this corresponds to an approximation of the dual of a Lie algebra $\mathfrak{g}^*$ by a finite-dimensional space. The space $\mathfrak{g}^*$ typically corresponds to some space of $\mathbb{R}^k$-valued functions defined on a manifold. The most natural way of discretising $\mathfrak{g}^*$ is to introduce a grid on the manifold and identify a function with the values it takes over the grid. In this way we naturally obtain a finite-dimensional approximation of $\mathfrak{g}^*$. However, the approximation does not inherit a Poisson structure in a natural way, as we will see in the example of the Burgers' equation (remark 9.3). Therefore, finding a spatial discretisation with good structure preserving properties is a challenge.

Lie-Poisson systems $(\mathfrak{g}^*, \{,\}, H)$ can be realised as collective Hamiltonian systems $(M, \Omega, H \circ J)$ on symplectic manifolds, where $J \colon M \to \mathfrak{g}^*$ is a Poisson map. The flow of $(M, \Omega, H \circ J)$ maps fibres of J to fibres of J and is symplectic. Therefore, it decends to a Poisson map on the original system $(\mathfrak{g}^*, \{,\}, H)$. Since the Hamiltonian vector field to $H \circ J$ on (M, Ω) is J-related to the Hamiltonian vector field to H on $(\mathfrak{g}^*, \{,\})$, motions of $(M, \Omega, H \circ J)$ decend to motions of $(\mathfrak{g}^*, \{,\}, H)$.

The reason to consider a collective system for numerical integrations rather than the Lie-Poisson system directly is that the symplectic structure can easily be preserved under spacial discretisations and widely applicable, efficient symplectic integrators are available [6]. The challenge of integrating $(\mathfrak{g}^*, \{,\}, H)$ in a structure preserving way thus shifts to finding a realisation, i.e. (M, ω) and $J \colon M \to \mathfrak{g}^*$, such that all initial conditions of interest lie in the image of J and such that the system $(M, \Omega, H \circ J)$ is practical to work with.

A practical choice for a realisation is where J is a *Clebsch map* [10]: let X be a Riemannian manifold and let $M = T^*\mathscr{C}^\infty(X, \mathbb{R}^k) \cong C^\infty(X, \mathbb{R}^k) \times C^\infty(X, \mathbb{R}^k)^*$, where $C^\infty(X, \mathbb{R}^k)^*$ is identified with $C^\infty(X, \mathbb{R}^k)$ via the L^2 pairing. The vector space M is equipped with the symplectic form

$$\Omega((u_1, u_2), (v_1, v_2)) = \int_X (\langle u_1, v_2 \rangle_{\mathbb{R}^k} - \langle u_2, v_1 \rangle_{\mathbb{R}^k}) \, \mathrm{dvol}_X,$$

where $\langle .,. \rangle_{\mathbb{R}^k}$ denotes the scalar product in $\mathbb{R}^k$. For an element $(f, g) \in M$ we denote the post-composition of f and g by the projection map to the j^{th} component of $\mathbb{R}^k$ by $q^j(f)$ and $p_j(g)$, respectively. In other words, $q^1,\dots,q^k,p_1,\dots,p_k$ are maps $M \to \mathscr{C}^\infty(X,\mathbb{R})$ such that for $x \in X$

$$(f(x),g(x)) = \Big(\big(q^1(f)(x),\dots,q^k(f)(x)\big),\big(p_1(g)(x),\dots,p_k(g)(x)\big)\Big).$$

Identifying tangent spaces of the vector space M with itself, we may write Ω as

$$\Omega = \int_X \left(\sum_{j=1}^k \mathrm{d}q^j \wedge \mathrm{d}p_j\right) \mathrm{dvol}_X = \int_X \langle \mathrm{d}q \wedge \mathrm{d}p\rangle_{\mathbb{R}^k} \,\mathrm{dvol}_X,$$

where $q = (q^1,\dots,q^k)$ and $p = (p_1,\dots,p_k)$*. If $J\colon M \to \mathfrak{g}^*$ is a realisation of a Lie Poisson system $(\mathfrak{g}^*,\{,\})$, then J is called a *Clebsch map* and (q,p) are called *Clebsch variables*. In Clebsch variables Hamilton's equations for $\bar{H} = H \circ J\colon M \to \mathbb{R}$ are in canonical form, i.e.

$$q_t = \frac{\delta \bar{H}}{\delta p}, \qquad p_t = -\frac{\delta \bar{H}}{\delta q},$$

where $\frac{\delta \bar{H}}{\delta q}$ and $\frac{\delta \bar{H}}{\delta p}$ are variational derivatives. The reason why Clebsch variables are a natural choice of coordinates for a structure preserving setting is that if X is discretised using a mesh then the integral in the expression for Ω naturally becomes a (weighted) sum over all mesh points and Hamilton's equations for the discretisation of the collective system $(M,\Omega,H\circ J)$ are in (a scaled version of the) canonical form. This means the system can be integrated using a symplectic integrator like, for instance, the midpoint rule. The setting is summarised in table 9.1.1.

The symplectic system in Clebsch variables has, after spatial discretisation, twice as many variables as the discretisation of the PDE in the original variables. An increase in the amount of variables needs some justification because it does not only lead to more work per integration step but, thinking of multi-step methods versus one-step methods, can also lead to worse stability behaviour [6, XV]. Moreover, integrating a lifted, symplectic system with a symplectic integrator instead of the original system with a non-symplectic integrator is not necessarily of any advantage. If, for instance, we integrate the Hamiltonian system

$$\begin{aligned} \dot{u} = F(u) \quad &= \quad \nabla_p \langle F(u), p\rangle \\ \dot{p} = -\mathrm{D}F(u)^T p &= -\nabla_u \langle F(u), p\rangle \end{aligned}$$

*The notation is natural when considering M as a Fréchet manifold over $C^\infty(X,\mathbb{R})$ or $C^\infty(X,\mathbb{R}^k)$ with coordinates $(q^1,\dots,q^k,p_1,\dots,p_k)$ or (q,p), respectively.

Continuous system	Spatially discretised system
Collective Hamiltonian system on an infinite-dimensional symplectic vector space in Clebsch variables $q_t = \frac{\delta \bar{H}}{\delta p}, \quad p_t = -\frac{\delta \bar{H}}{\delta q}.$ Exact solutions preserve the symplectic structure, the Hamiltonian $\bar{H} = H \circ J$, all quantities related to the Casimirs of the original PDE and the fibres of the Clebsch map $J(q,p) = u$.	Canonical Hamiltonian ODEs in $2N$ variables $\hat{q}_t = \nabla_{\hat{p}} \hat{\bar{H}}, \quad \hat{p}_t = -\nabla_{\hat{q}} \hat{\bar{H}}.$ The exact flow preserves the symplectic structure and the Hamiltonian $\hat{\bar{H}}$. Time-integration with the midpoint rule is symplectic.
Original PDE, interpreted as a Lie-Poisson equation $u_t = \mathrm{ad}^*_{\frac{\delta H}{\delta u}} u.$ Exact solutions preserve the Poisson structure, the Hamiltonian H and all Casimirs.	Non-Hamiltonian ODEs in N variables $\hat{u}_t = K(\hat{u}) \nabla_{\hat{u}} \hat{H}, \qquad K^T = -K.$ Exact solutions conserve $\hat{H}$. Time-integration with the midpoint rule is *not* symplectic.

Table 9.1.1: Overview of the setting.

rather than the system $\dot{u} = F(u)$ directly then preserving the symplectic structure in a numerical computation does not have any effect: in this example the symplectic structure is artificially introduced and not related to the original system. This illustrates that using symplectic integrators is not an end in itself. It is the presence of a Poisson structure and its interplay with the symplecticity of the collective system which can justify doubling the amount of variables as our numerical examples will indicate.

Let us provide examples for the application of Clebsch variables. Euler's equation in hydrodynamics for an ideal incompressible fluid with velocity u and pressure ρ on a 3-dimensional compact, Riemannian manifold X with boundary ∂X or a region $X \subset \mathbb{R}^3$ are given as

$$u_t + u \cdot \nabla u = -\nabla \rho, \qquad \operatorname{div} u = 0, \qquad u|_{\partial X} \text{ is parallel to } \partial X.$$

Elements in the dual of the Lie-algebra χ^*_{vol} to the Fréchet Lie-group of volume preserving diffeomorphisms $\mathscr{D}_{\text{vol}}$ can be considered as 2-forms on X. Using $\nabla \times u \mathrel{\widehat{=}} \mathrm{d}u^\flat$ Euler's equations correspond to motions on the Lie-Poisson system to $\mathscr{D}_{\text{vol}}$ with Hamiltonian $H(\sigma) = \frac{1}{2}\int_X \langle \Delta^{-1}\sigma, \sigma\rangle \mathrm{dvol}_X$, where Δ is the Laplace-DeRham operator and $\langle , \rangle$ the metric pairing of 2-forms [10].

A Clebsch map $J\colon M \to \chi^*_{\text{vol}}$ can be obtained as the momentum map of the cotangent lifted action of the action $(\eta, f) \mapsto f \circ \eta^{-1}$ of $\mathscr{D}_{\text{vol}}$ on $\mathscr{C}^\infty(X, \mathbb{R})$. However, J is not surjective and flows with non-zero hydrodynamical helicity cannot be modelled. To overcome this issue one can consider $M = \mathscr{C}^\infty(X, S^2)$, where S^2 is the 2-sphere. The symplectic form σ_{S^2} on the sphere induces the symplectic form $\Omega = \int_X \sigma_{S^2} \mathrm{dvol}_X$ on M. We can define $J\colon M \to \chi^*_{\text{vol}}$ as $J(s) = s^*\sigma_{S^2}$, where $s^*\sigma_{S^2}$ denotes the pull-back of σ_{S^2} to a 2-form on X which can be interpreted as an element in χ^*_{vol}. The map J is called a *spherical Clebsch map* and initial conditions with non-zero helicity are admissible. However, the helicity remains quantised [9]. Spherical Clebsch maps have been used for computational purposes in [2]: after a discretisation of the domain X, solutions to the (regularised) hydrodynamical equations are approximated by integrating the corresponding set of ODEs on the product $\Pi_{\text{mesh}(X)} S^2$ while preserving the spheres using a projection method (not preserving the symplectic form $\sum_{\text{mesh}(X)} \sigma_{S^2}$, though).

In the case of Hamiltonian ODEs on (finite-dimensional) Poisson spaces $(\mathfrak{g}^*, \{,\})$, no spatial discretisation is necessary. This setting applies to the rigid-body equations, for instance [11]. In the ODE setting, the authors of [15] apply symplectic integrators to the collective systems $(M, \Omega, H \circ J)$ with the property that the discrete flow preserves the fibres of J. Such integrators are called *collective integrators*. Their flow descends to a Poisson map on the original system

$(\mathfrak{g}^*, \{,\}, H)$ such that one obtains a Poisson integrator for $(\mathfrak{g}^*, \{,\}, H)$.

In this paper, we show how the collective integrator idea can be used in the infinite-dimensional setting, i.e. for Lie-Poisson systems to infinite-dimensional Lie-groups. In particular, we will consider the inviscid Burgers' equation

$$u_t + uu_x = 0$$

with $u(t,.) \in \mathscr{C}^\infty(S^1, \mathbb{R})$. The L^2-norm of $u(t,.)$ as well as the quantity

$$\int_{S^1} \sqrt{|u(t,.)|}\mathrm{d}x$$

are conserved quantities. They constitute the Hamiltonian and Casimirs of the Lie-Poisson formulation of the problem. Setting $u = q_x p$ we obtain the following set of PDEs

$$q_t = -\frac{1}{3}q_x^2 p, \qquad p_t = -\frac{1}{3}(q_x p^2)_x$$

with $q(t,\cdot) \in \mathscr{C}^\infty(S^1, S^1)$ and $p(t,\cdot) \in \mathscr{C}^\infty(S^1, \mathbb{R})$ which is the collective system. The variables q, p may be regarded as Clebsch variables (right in the middle between classical and spherical Clebsch variables).
We will also experiment with the following more complicated PDE which fits into the same setting as the inviscid Burgers' equation.

$$u_t = 3uu_x - \frac{9}{4}u^2 u_x - u_x u_{xx} - 3u_x^2 u_{xx} - 2uu_{xxx} - 2uu_x u_{xxx} - 6uu_{xx}^2$$

It has the conserved quantity $H(u) = \int_{S^1}(u^2 + u_x^2 - 1/2u^3 + u_x^3)\mathrm{d}x$ as well as $\int_{S^1}\sqrt{|u|}\mathrm{d}x$ in time. In Clebsch variables we have

$$\begin{aligned} q_t = \frac{\delta \bar{H}}{\delta p} &= q_x\Big(q_x p - \frac{3}{4}(q_x p)^2 - ((q_x p)_x + \frac{3}{2}(q_x p)_x^2)_x\Big) \\ p_t = -\frac{\delta \bar{H}}{\delta p} &= p\Big(\frac{3}{2}(q_x p)^2 - q_x p + \big((q_x p)_x + \frac{3}{2}(q_x p)_x^2\big)_x\Big)_x. \end{aligned}$$

The PDEs are discretised in space by introducing a periodic grid on S^1 and replacing the integral in H by a sum. In this way we obtain a system of Hamiltonian ODEs in canonical form.

Integration using the symplectic midpoint rule yields an integrator with excellent structure preserving properties like bounded energy and Casimir errors, although it does not preserve the fibres of J and therefore does not descend to a Poisson integrator. The good behaviour is linked to the symplecticity of the collective system which is preserved exactly by the midpoint rule. Therefore, the conservation properties survive even when the equation is perturbed within the class of Hamiltonian PDEs. This robustness can be an advantage over more traditional ways of discretising the PDE directly since these make use of structurally simple symmetries of the equation that are immediately destroyed when higher order terms are introduced. Our numerical experiments indicate that the advantage of symplectic integration can outweigh the disadvantage of doubling the variables from u to (q,p).

9.2 Introduction

Let us briefly review the setting of Hamiltonian systems on Poisson manifolds. For details we refer to [13].

Definition 9.1 (Poisson manifold and Poisson bracket)**.** A *Poisson manifold P* is a smooth manifold together with an $\mathbb{R}$-bilinear map

$$\{\cdot,\cdot\}\colon \mathscr{C}^\infty(P)\times\mathscr{C}^\infty(P)\to\mathscr{C}^\infty(P)$$

satisfying

- $\{f,g\}=-\{g,f\}$ (skew-symmetry),
- $\{f,\{g,h\}\}+\{g,\{h,f\}\}+\{h,\{f,g\}\}=0$ (Jacobi identity),
- $\{fg,h\}=f\{g,h\}+g\{f,h\}$ (Leibniz's rule).

The map $\{\cdot,\cdot\}$ is called the *Poisson bracket.*

Example 13. If G is a (Fréchet-) Lie-group with Lie-algebra $\mathfrak{g}$ and dual $\mathfrak{g}^*$ then

$$\{f,g\}(w)=\left\langle w,\left[\frac{\delta f}{\delta w},\frac{\delta g}{\delta w}\right]\right\rangle,\qquad w\in\mathfrak{g}^*,\ f,g\in\mathscr{C}^\infty(\mathfrak{g}^*) \tag{9.2.1}$$

is a (Lie-) Poisson bracket on $\mathfrak{g}^*$, where $\langle\cdot,\cdot\rangle$ denotes the duality pairing of $\mathfrak{g}^*$ and $\mathfrak{g}$, $[\cdot,\cdot]$ denotes the Lie bracket on $\mathfrak{g}$ and $\frac{\delta f}{\delta w}\in\mathfrak{g}$ is defined by

$$\forall v\in\mathfrak{g}^*:\quad \mathrm{D}f|_w(v)=\left\langle v,\frac{\delta f}{\delta w}\right\rangle$$

with Fréchet derivative D. ►

Definition 9.2 (Hamiltonian system and Hamiltonian motion)**.** A *Hamiltonian system* $(P, \{\cdot,\cdot\}, H)$ is a Poisson manifold $(P, \{\cdot,\cdot\})$ together with a smooth map $H\colon P \to \mathbb{R}$. The *Hamiltonian vectorfield* X_H to the system $(P, \{\cdot,\cdot\}, H)$ is defined as the derivation $X_H = \{\cdot, H\}$. If $f\colon P \to \mathbb{R}$ is a smooth function, then the *motion of the system* $(P, \{\cdot,\cdot\}, H)$ in the coordinate f is given by the differential equation $\dot{f} = \{f, H\}$, where the dot denotes a time-derivative.

Example 14. A Hamiltonian system (M, ω, H) on a symplectic manifold (M, ω) constitutes a Hamiltonian system on the Poisson manifold $(M, \{\cdot,\cdot\})$. The Poisson bracket $\{\cdot,\cdot\}$ is defined by $\{f, g\} = \omega(X_f, X_g)$ where the vector fields X_f and X_g are defined by $\mathrm{d}f = \omega(X_f, \cdot)$ and $\mathrm{d}g = \omega(X_g, \cdot)$. If M is $2n$-dimensional with local coordinates $q^1, \ldots, q^n, p_1, \ldots, p_n$ and $\omega = \sum_{j=1}^n \mathrm{d}q^j \wedge \mathrm{d}p_j$ then

$$X_H = \sum_{j=1}^n \frac{\partial H}{\partial p_j}\frac{\partial}{\partial q^j} - \frac{\partial H}{\partial q^j}\frac{\partial}{\partial p_j}.$$

The motions of the system are given by

$$\begin{aligned} \dot{q}^j &= \{q^j, H\} = X_H(q^j) = \frac{\partial H}{\partial p_j}, \\ \dot{p}_j &= \{p_j, H\} = X_H(p_j) = -\frac{\partial H}{\partial q^j}. \end{aligned}$$

with $j = 1, \ldots, n$. ▶

Remark 9.1. For Hamiltonian systems on a finite-dimensional, symplectic manifold, there exist local coordinates such that the motions are given by

$$\dot{z} = S \nabla H(z),$$

for a constant, skew-symmetric, non-degenerate matrix S. The analogue for finite-dimensional Poisson systems is that S is allowed to be z dependent and degenerate (but still skew-symmetric).

Remark 9.2. Like in the symplectic case, the Hamiltonian is a conserved quantity under motions of the corresponding Hamiltonian system on a Poisson manifold. Additionally, the Poisson structure encodes interesting geometric features of Hamiltonian motions. Casimir functions, which are real valued functions f with $\{f, \cdot\} = 0$ are conserved quantities (with no dependence on the Hamiltonian). While the only Casimirs are constants if the Poisson structure is induced by a symplectic structure, non-trivial Casimir functions are admissible in the Poisson case. Moreover, in a Poisson system a motion never leaves the coadjoint orbit in which it was initialised. We refer to [13, Ch.10] for proofs and more properties of Poisson manifolds.

In what follows we will present an integrator for Hamiltonian systems on the dual of the Lie-algebra of the group of diffeomorphisms on the circle. The setting covers, for example, Burgers' equation and perturbations. This shows how to apply the ideas of [15] in the infinite-dimensional setting of Hamiltonian PDEs.

9.3 Lie-Poisson structure on diff$(S^1)^*$

Consider the Fréchet Lie-group $G = \mathrm{Diff}(S^1)$ of orientation preserving diffeomorphisms on the circle S^1. In the following we view S^1 as the quotient $\mathbb{R}/L\mathbb{Z}$ for $L > 0$ with coordinate x obtained from the universal covering $\mathbb{R} \to \mathbb{R}/L$. The Lie-algebra $\mathfrak{g}$ can be identified with the space of smooth vector fields on $\mathscr{S}^1$, where the Lie-bracket is given as the negative of the usual Lie-bracket of vector fields

$$\left[u\frac{\partial}{\partial x}, v\frac{\partial}{\partial x}\right] = (u_x v - v_x u)\frac{\partial}{\partial x}.$$

Here, the prime denotes a derivative with respect to the coordinate x on $S^1 = \mathbb{R}/L\mathbb{Z}$. [8, Thm.43.1] The dual $\mathfrak{g}^*$ of the Lie algebra[†] can be identified with the quadratic differentials on the circle $\Omega^{\otimes 2}(S^1) = \{u \cdot (\mathrm{d}x)^2 \,|\, u \in \mathscr{C}^\infty(S^1, \mathbb{R})\}$. The dual pairing is given by

$$\left\langle u(\mathrm{d}x)^2, v\frac{\partial}{\partial x}\right\rangle = \int_{S^1} u(x)v(x)\mathrm{d}x.$$

[7, Prop. 2.5] The coadjoint action of an element $\phi \in G$ on an element $u(\mathrm{d}x)^2$ is given as

$$\mathrm{Ad}^*_{\phi^{-1}}\left(u(\mathrm{d}x)^2\right) = (u \circ \phi) \cdot \phi'^2 \cdot (\mathrm{d}x)^2 = \phi^*\left(u(\mathrm{d}x)^2\right).$$

We see that the coadjoint action on $u(\mathrm{d}x)^2$ preserves the zeros of u. The map u will have an even number of zeros. Consider two consecutive zeros $a, b \in S^1$. The integral

$$\int_a^b \sqrt{|u(x)|}\mathrm{d}x$$

is constant on the coadjoint orbit through $u(\mathrm{d}x)^2$ since the action corresponds to a diffeomorphic change of the integration variable in the above expression. It follows that the map $\Phi\colon \mathfrak{g}^* \to \mathbb{R}$ with

$$\Phi(u(\mathrm{d}x)^2) = \int_{S^1} \sqrt{|u(x)|}\mathrm{d}x$$

[†] which does not coincide with the functional analytic dual to $\mathfrak{g}$

is a Casimir for the Poisson structure on $\mathfrak{g}^*$. [7] For $H \in \mathscr{C}^\infty(\mathfrak{g}^*, \mathbb{R})$ Hamilton's equations are given as

$$\frac{\mathrm{d}}{\mathrm{d}t} u(t,x)(\mathrm{d}x)^2 = \mathrm{ad}^*_{\frac{\delta H}{\delta u(t,\cdot)(\mathrm{d}x)^2}} \left(u(t,x)(\mathrm{d}x)^2 \right)$$

or, identifying $\mathfrak{g}$ and $\mathfrak{g}^*$ with $\mathscr{C}^\infty(S^1, \mathbb{R})$,

$$u_t = \mathrm{ad}^*_{\frac{\delta H}{\delta u}} u.$$

Here $\frac{\delta H}{\delta u}$ denotes the functional or variational derivative of H and $\mathrm{ad}^*_\eta \colon \mathfrak{g}^* \to \mathfrak{g}^*$ the dual map to $\mathrm{ad}_\eta \colon \mathfrak{g} \to \mathfrak{g}$ given by

$$\mathrm{ad}_\eta(\mu) = [\eta, \mu].$$

[13, Prop. 10.7.1.]

Lemma 9.1. *Hamilton's equations can be rewritten as*

$$u_t = \left(\frac{\partial}{\partial x} u + u \frac{\partial}{\partial x} \right) \frac{\delta H}{\delta u}. \tag{9.3.1}$$

Proof. Let $v \in \mathfrak{g}$, $u \in \mathfrak{g}^*$ (both identified with $\mathscr{C}^\infty(S^1, \mathbb{R})$). Denoting the dual pairing between $\mathfrak{g}$ and $\mathfrak{g}^*$ by $\langle , \rangle$, we obtain

$$\begin{aligned} \left\langle \mathrm{ad}^*_{\frac{\delta H}{\delta u}} u, v \right\rangle &= \left\langle u, \mathrm{ad}_{\frac{\delta H}{\delta u}} v \right\rangle = \left\langle u, \left[\frac{\delta H}{\delta u}, v \right] \right\rangle = \left\langle u, \left(\frac{\delta H}{\delta u} \right)_x \cdot v - \left(\frac{\delta H}{\delta u} \right) \cdot v_x \right\rangle \\ &= \left\langle u \cdot \left(\frac{\delta H}{\delta u} \right)_x, v \right\rangle - \left\langle u \cdot \left(\frac{\delta H}{\delta u} \right), v_x \right\rangle \\ &= \left\langle u \cdot \left(\frac{\delta H}{\delta u} \right)_x, v \right\rangle + \left\langle \left(u \cdot \left(\frac{\delta H}{\delta u} \right) \right)_x, v \right\rangle, \end{aligned}$$

whereas the last equation follows using integration by parts. □

Example 15. On $\mathfrak{g}^*$ consider the Hamiltonian

$$H(u) = \int_{S^1} \mathscr{H}(u^{\mathrm{jet}}(x)) \mathrm{d}x$$

with $\mathscr{H} \colon \mathbb{R}^{K+1} \to \mathbb{R}$ and the K-jet of the map u

$$\begin{aligned} u^{\mathrm{jet}}(x) &:= (u(x), u_x(x), u_{x^2}(x), \ldots, u_{x^K}(x)) \\ &:= \left(u(x), \left. \frac{\partial u}{\partial x} \right|_x, \left. \frac{\partial^2 u}{\partial x^2} \right|_x, \ldots, \left. \frac{\partial^K u}{\partial x^K} \right|_x \right). \end{aligned}$$

By lemma 9.1, Hamilton's equations are given as

$$u_t = \left(\frac{\partial}{\partial x}u + u\frac{\partial}{\partial x}\right)\sum_{j=0}^{K}(-1)^j \frac{\partial}{\partial x^j}\left(\frac{\partial \mathscr{H}}{\partial u_{x^j}}(u^{\mathrm{jet}})\right).$$

For $\mathscr{H}(u) = -\frac{1}{6}u^2$ we obtain the inviscid Burgers' equation $u_t + uu_x = 0$. ▶

Remark 9.3. Using formula (9.2.1) from example 13 identifying $\mathfrak{g} \cong \mathscr{C}^\infty(S^1,\mathbb{R})$ and $\mathfrak{g}^* \cong \mathscr{C}^\infty(S^1,\mathbb{R})$, the Lie-Poisson bracket is given by

$$\{F,G\}(u) = \int_{S^1}\left(\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\delta F}{\delta u}\right)\frac{\delta G}{\delta u} - \frac{\delta F}{\delta u}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\delta G}{\delta u}\right)\right) f\,\mathrm{d}x,$$

where $\frac{\delta F}{\delta u}$ denotes the functional or variational derivative of F at u. Discretising $S^1 \cong \mathbb{R}/\mathbb{Z}$ using a (periodic) grid with N grid-points, we naturally obtain $\mathbb{R}^N$ as a discrete analog of $\mathfrak{g}^*$. However, the above Poisson structure does not pass naturally to $\mathbb{R}^N$.

9.4 The collective system

Let us construct a realisation $J\colon M \to \mathfrak{g}^*$ where M is a symplectic vector space. Consider the left-action of $g \in G = \mathrm{Diff}(S^1)$ on $q \in Q = \mathscr{C}^\infty(S^1,S^1)$ defined by $g.q = q \circ g^{-1}$.

Lemma 9.2. *The vector field $\hat{v}$ generated by the infinitesimal action of an element $v \in \mathfrak{g} \cong \mathfrak{X}(S^1)$ on Q is given by the Lie-derivative $-\mathscr{L}_v$. Interpreting v as an element in $\mathscr{C}^\infty(S^1,\mathbb{R})$, this becomes $\hat{v}_q = -v \cdot q' \in \mathscr{C}^\infty(S^1,\mathbb{R}) \cong T_qQ$.*

Proof. Let $g\colon (-\epsilon,\epsilon) \to \mathrm{Diff}(S^1)$ be a smooth curve with $g_0 = \mathrm{id}$ and $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} g_t = v \in \mathfrak{g} \cong \mathscr{C}^\infty(S^1,\mathbb{R})$. Let $x \in S^1$. Deriving $x = g_t(g_t^{-1}(x))$ w.r.t. t at $t=0$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0} g_t^{-1}(x) = -v(x).$$

Let $q \in Q$. We have

$$\hat{v}_q(x) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0}(g_t.q)(x) = \frac{\mathrm{d}}{\mathrm{d}t}\bigg|_{t=0}\left(q \circ g_t^{-1}\right)(x) = -v(x)q'(x).$$

□

Let M denote the cotangent bundle over Q, which is viewed as $T^*Q \cong Q \times \mathscr{C}^\infty(S^1,\mathbb{R})$. The pairing of $(q,p) \in M$ with an element $v \in T_qQ \cong \mathscr{C}^\infty(S^1,\mathbb{R})$ is given by

$$\langle (q,p), v\rangle = \int_{S^1} p(x)v(x)\mathrm{d}x.$$

A symplectic structure on M is given by

$$\Omega((v^q, v^p), (w^q, w^p)) = \int_{S^1} (w^p v^q - v^p w^q)\mathrm{d}x.$$

For $(q, p) \in M$ and $\bar{H}\colon M \to \mathbb{R}$ the maps $\frac{\delta \bar{H}}{\delta q}$ and $\frac{\delta \bar{H}}{\delta p}$ can be defined by

$$\mathrm{D}\bar{H}|_{(q,p)}(w^q, 0) = \int_{S^1} \frac{\delta \bar{H}}{\delta q} w^q \mathrm{d}x, \quad \mathrm{D}\bar{H}|_{(q,p)}(0, w^p) = \int_{S^1} \frac{\delta \bar{H}}{\delta p} w^p \mathrm{d}x,$$

where D denotes the Gâteaux derivative.[‡] Now

$$\mathrm{D}\bar{H}|_{(q,p)}(w^q, w^p) = \Omega\left(\left(\frac{\delta \bar{H}}{\delta p}, -\frac{\delta \bar{H}}{\delta q}\right), (w^q, w^p)\right)$$

and Hamilton's equations can be written in the familiar looking form

$$q_t = \frac{\delta \bar{H}}{\delta p}, \qquad p_t = -\frac{\delta \bar{H}}{\delta q}. \tag{9.4.1}$$

We consider the cotangent lifted action of the aforementioned action of G on Q to obtain a Hamiltonian group action of G on M given by

$$g.(q, p) = (q \circ g^{-1}, p \circ g^{-1} \cdot (g^{-1})_x).$$

Alternatively, interpreting the fibre component of elements in T^*Q as 1-forms the action is given by $g.(q, p\mathrm{d}x) = \left(q \circ g^{-1}, (g^{-1})^*(p\mathrm{d}x)\right)$.

Proposition 1 *The momentum map $J\colon M \to \mathfrak{g}^*$ of the cotangent lifted action of G on M is given as*

$$J(q, p) = -q_x \cdot p.$$

Proof. Using the formula for the momentum map of cotangent lifted action (see [11, p.283]) we obtain

$$\langle v, J(q,p)\rangle = \langle (q,p), \hat{v}_q\rangle = \langle (q,p), -v \cdot q_x\rangle = -\int_{S^1} v(x) p(x) q_x(x) \mathrm{d}x = \langle v, -q_x \cdot p\rangle$$

as claimed. □

[‡] Each maps $\frac{\delta \bar{H}}{\delta q}$ and $\frac{\delta \bar{H}}{\delta p}$ can depend on both q and p although this is not incorporated in the notation.

The manifold M is equipped with a Poisson structure defined by the symplectic structure Ω. By construction, the momentum map $J\colon M \to \mathfrak{g}^*$ is a Poisson map. It is surjective (take $q = \mathrm{id}$) and therefore called a *full realisation of* $\mathfrak{g}^*$. If H is a Hamiltonian on $\mathfrak{g}^*$ then the Hamiltonian flow of the *collective* system $(M, \Omega, H \circ J)$ maps fibres of J to fibres and descends to the Hamiltonian flow of the system $(\mathfrak{g}^*, \{\cdot,\cdot\}, H)$ because the Hamiltonian vector fields are J-related and J is a Poisson map. More generally, a symplectic map on M that maps fibres to fibres descends to a Poisson map on g^*.

Example 16. As in example 15 we consider the Hamiltonian

$$H(u) = \int_{S^1} \mathscr{H}(u^{\mathrm{jet}}(x))\mathrm{d}x$$

on $\mathfrak{g}^*$. Hamilton's equations of the collective system $(M, \Omega, H \circ J)$ are given as the following system of PDEs

$$\begin{aligned} q_t &= q_x \sum_{j=0}^{K} (-1)^j \frac{\partial^j}{\partial x^j}\left(\frac{\partial \mathscr{H}}{\partial u_{x^j}}(u^{\mathrm{jet}})\right), \\ p_t &= -\frac{\partial}{\partial x}\left(p \sum_{j=0}^{K} (-1)^j \frac{\partial^j}{\partial x^j}\left(\frac{\partial \mathscr{H}}{\partial u_{x^j}}(u^{\mathrm{jet}})\right)\right). \end{aligned}$$

Choosing $\mathscr{H}(u) = -\frac{1}{6}u^2$ (Burgers' equation) yields

$$q_t = -\frac{1}{3} q_x^2 p, \qquad p_t = -\frac{1}{3}(q_x p^2)_x.$$

▶

9.5 Integrator of the collective system

9.5.1 Spatial discretisation

We use a second-order finite-difference method in space to discretise the realisation J and the Hamiltonian H to obtain a system of Hamiltonian ODEs in canonical form: as before, we consider S^1 as the quotient $\mathbb{R}/L\mathbb{Z}$. We introduce a uniform grid $(x_1, \ldots, x_N)$, $x_j = j \cdot \Delta x$, $\Delta x = 1/N$ with N points and periodic boundary conditions. Moreover, we consider the corresponding half-grid $(x_{1/2}, \ldots, x_{N-1/2})$. Both grids are illustrated in figure 9.5.1. In the discretised setting, elements in $Q = \mathscr{C}^\infty(S^1, S^1)$ and $\mathscr{C}^\infty(S^1, \mathbb{R})$ are approximated by their values on the considered grid. This leads to an approximation of $\mathfrak{g}^*$ and Q by the vector space $\mathbb{R}^N$ and an approximation of M by $T^*\mathbb{R}^N \cong \mathbb{R}^{2N}$, which

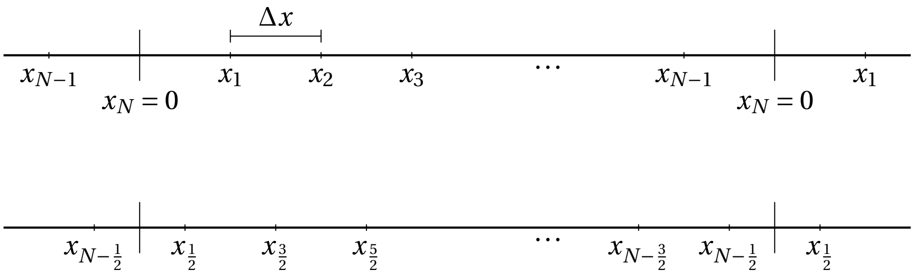


Figure 9.5.1: Uniform periodic grids on $S^1 \cong \mathbb{R}/L\mathbb{Z}$, $L > 0$.

we equip with coordinates $(\hat{q}, \hat{p}) = q^1, \dots, q^N, p_1, \dots, p_n$ in the usual way. Discretising the symplectic structure Ω we obtain

$$\omega = \Delta x \sum_{j=1}^{N} \mathrm{d}q^j \wedge \mathrm{d}p_j,$$

which is the standard symplectic structure up to the factor Δx. For $q \in Q$ we obtain a second-order accurate approximation $D_{\Delta x}(\hat{q})$ of the spatial derivative q_x on the half-grid $(1/2\Delta x, 3/2\Delta x, \dots, (N-1/2)\Delta x)$ using compact central differences as follows:

$$\left(q_x(\tfrac{1}{2}\Delta x), q_x(\tfrac{3}{2}\Delta x), \dots, q_x((N-\tfrac{3}{2})\Delta x), q_x((N-\tfrac{1}{2})\Delta x)\right)^T$$

$$\approx \frac{1}{\Delta x}\left[\underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ -1 & 1 & \dots & 0 & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}}_{=:T}\begin{pmatrix} q(\Delta x) \\ q(2\Delta x) \\ \vdots \\ q((N-1)\Delta x) \\ q(N\Delta x) \end{pmatrix} + \begin{pmatrix} C(q) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}\right].$$

The quantity $C(q)/L$ is the winding number (degree) of the map $q\colon S^1 \to S^1$[§]. The values for q_x are now available on the half-grid. Notice that the quantity $C(q)$ is constant if q evolves smoothly subject to the PDE (9.4.1) because $C(q)$ can only take values in $L\mathbb{Z}$.

[§] Let $\pi\colon \mathbb{R} \to S^1$ denote the universal covering of $S^1 \cong \mathbb{R}/L\mathbb{Z}$ and let $\tilde{q}\colon \mathbb{R} \to \mathbb{R}$ be any lift of the map $\pi \circ q\colon \mathbb{R} \to S^1$ to the covering space. Now $C(q) = \tilde{q}(L) - \tilde{q}(0)$. If, for instance, q is the identity map on S^1 then $C(q) = L$.

A discrete version of the map $J\colon M \to g^*$ is given by $\hat{J}\colon \mathbb{R}^{2N} \to \mathbb{R}^N$ with $\hat{J}(\hat{q},\hat{p}) = D_{\Delta x}\hat{q}.S\hat{p}$. Its values correspond to the half-grid. The matrix S is given as

$$S = \frac{1}{2}\begin{pmatrix} 1 & & & 1 \\ 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{pmatrix}.$$

It averages the values of $\hat{p}$ to obtain second order accurate approximations of p on the half-grid. In this way, we obtain approximations to $u = q_x p$ on the half grid. Approximations for u_x and higher derivatives are obtained by successively applying $T_{\Delta x}$ and $T_{\Delta x}^T$, i.e.

$$\frac{\partial^k u}{\partial x^k} \approx \frac{\partial^k_{\Delta x} u}{\partial_{\Delta x} x^k} := \begin{cases} D_{\Delta x}\hat{q}.S\hat{p} & \text{if } k = 0 \\ -T^T \frac{\partial^{k-1}_{\Delta x} u}{\partial_{\Delta x} x^{k-1}} / \Delta x & \text{if } k \text{ is odd} \\ T \frac{\partial^{k-1}_{\Delta x} u}{\partial_{\Delta x} x^{k-1}} / \Delta x & \text{if } k \text{ is even.} \end{cases} \tag{9.5.1}$$

Here T^T denotes the transpose of the matrix T and . denotes component-wise multiplication. Now all approximations for even derivatives are available on the half-grid and all odd derivatives on the full-grid. A Hamiltonian of the form $\int_{S^1} \mathcal{H}(u, u_x, u_{xx}, \ldots)\mathrm{d}x$ is approximated by the sum

$$\int_{S^1} \mathcal{H}(u, u_x, u_{xx}, \ldots)\mathrm{d}x \approx \Delta x \sum_{j=1}^{N} \mathcal{H}(u(x_{j-1/2}), u_x(x_{j-1/2}), u_{xx}(x_{j-1/2}), \ldots). \tag{9.5.2}$$

To evaluate (9.5.2), all approximations of $\frac{\partial^k u}{\partial x^k}$ where k is odd are multiplied by S such that the approximation of the jet of u is available on the half-grid. The second-order averaging with S can be avoided if $\mathcal{H}$ is of the form

$$\begin{aligned} \mathcal{H}(u^{\text{jet}}(x)) = \mathcal{H}^{\text{even}}&(u(x), u_{xx}(x), u_{xxxx}(x), \ldots) \\ &+ \mathcal{H}^{\text{odd}}(u_x(x), u_{xxx}(x), u_{xxxxx}(x), \ldots). \end{aligned}$$

We can then approximate the Hamiltonian by

$$\begin{aligned} \int_{S^1} \mathcal{H}(u, u_x, u_{xx}, \ldots)\mathrm{d}x \approx \Delta x &\sum_{j=1}^{N} \mathcal{H}^{\text{even}}(u(x_{j-1/2}), u_{xx}(x_{j-1/2}), u_{xxxx}(x_{j-1/2}) \ldots) \\ + \Delta x &\sum_{j=1}^{N} \mathcal{H}^{\text{odd}}(u_x(x_j), u_{xxx}(x_j), u_{xxxxx}(x_j), \ldots). \end{aligned} \tag{9.5.3}$$

Taking into account that the symplectic form ω is the canonical symplectic structure scaled by Δx, defining $\hat{H}$ as

$$\hat{H}(u) = \sum_{j=1}^{N} \mathscr{H}(u(x_{j-1/2}), u_x(x_{j-1/2}), u_{xx}(x_{j-1/2}), \ldots) \tag{9.5.4}$$

or as the corresponding term from (9.5.3) puts Hamilton's equations into the canonical form

$$\dot{\hat{q}} = \nabla_{\hat{p}} \bar{\hat{H}}(\hat{q}, \hat{p}), \qquad \dot{\hat{p}} = -\nabla_{\hat{q}} \bar{\hat{H}}(\hat{q}, \hat{p}) \tag{9.5.5}$$

with collective Hamiltonian $\bar{\hat{H}} = \hat{H} \circ \hat{J} \colon \mathbb{R}^{2N} \to \mathbb{R}$. Here the dot denotes the time-derivative. Finally, (9.5.5) is a 2nd order accurate, spatial discretisation of (9.4.1).

Remark 9.4. An alternative to the described finite-difference discretisation are spectral methods. Notice that $q \in \mathscr{C}^\infty(S^1, S^1)$ can be split into the winding term $C(q)\mathrm{id}$ and the term $q - C(q)\mathrm{id}$ which has winding number zero. In a pseudo-spectral discretisation, the derivative of $q - C(q)\mathrm{id}$ is calculated in a Fourier basis and the winding term $C(q)\mathrm{id}$ is accounted for in the derivative q_x by adding the constant $C(q)/L$ component-wise. The derivatives of $u = q_x p$ can be calculated without complications.
A full spectral discretisation is also possible because embedding $\mathscr{C}^\infty(S^1, S^1)$ and $\mathscr{C}^\infty(S^1, \mathbb{R})$ into the Hilbert space L_2 and choosing any orthonormal basis will lead to a symplectic form ω which is in the standard form (splitting q as above to allow for a Fourier basis). Therefore, Hamilton's equations for the basis coefficients appear in canonical form.

9.5.2 The integration scheme

A numerical solution to the original equation (9.3.1) can now be obtained as follows.

1. Lift an initial condition

$$\hat{u}^{(0)} = (u^{(0)}(x_1), \ldots, (u^{(0)}(x_N))$$

to $(\hat{q}^{(0)}, \hat{p}^{(0)}) \in \hat{J}^{-1}(\hat{u}^{(0)})$, for example by setting

$$\begin{aligned} \hat{q}^{(0)} &= (\Delta x, 2\Delta x, \ldots, N\Delta x), \\ \hat{p}^{(0)} &= \hat{u}^{(0)}, \end{aligned}$$

as we will do in our numerical experiments. Notice that $\hat{q}^{(0)}$ is a discretisation of the identity map on S^1. The exact and discrete derivative is the constant 1 function or vector.

2. The system of Hamiltonian ODEs (9.5.5) can be integrated subject to the initial conditions $(\hat{q}^{(0)}, \hat{p}^{(0)})$ using a symplectic numerical integrator.

3. Approximations to u can be calculated from $(\hat{q}, \hat{p})$ on the half-grid as $D_{\Delta x}\hat{q}.S\hat{p}$.

Remark 9.5. Conservation of $\bar{\hat{H}}$ in (9.5.5) exactly corresponds to conservation of the discretised Hamiltonian $\hat{H}$ (9.5.2) or (9.5.3) because we consistently relate u and (q, p) by (9.5.1). Therefore, using a symplectic integrator to solve the system (9.5.5) of Hamiltonian ODEs we expect excellent energy behaviour of the numerical solution. In the following numerical experiments we will use the symplectic implicit midpoint rule. The arising implicit equations will be solved using Newton iterations.

Remark 9.6. In contrast to the case of Hamiltonian-ODEs on Poisson manifolds, it is hard for a symplectic integrator to maintain the structure fibration on the symplectic manifolds induced by the discretisation $\hat{J}$ of the realisation J. Indeed, the implicit midpoint rule used in our numerical examples fails to do so. This is why we do *not* obtain a (discretisation of a) Poisson integrator in this way. However, the described energy conservation properties of remark 9.5 are independent of this drawback. Moreover, our numerical examples will show that we obtain excellent Casimir behaviour although this has not been forced by this construction.

9.6 Numerical experiments

For the following numerical experiments, we consider Hamiltonian systems $(\mathrm{diff}^*(S^1), \{\cdot,\cdot\}, H)$ with

$$H = \int_{\mathscr{S}^1} \left(C_1 u^2 + C_2 u_x^2 + C_3 u^3 + C_4 u_x^3 \right) \mathrm{d}x. \tag{9.6.1}$$

To gain a sense of the relative performance of the collective integration method from section 9.5 we will now develop a conventional finite-difference approach for comparison that is based on [14].

First, a finite-dimensional discrete Hamiltonian approximation is obtained by

$$\hat{H} = \Delta x \sum_{j=1}^{N} (C_1 \hat{u}_j^2 + C_2 (\hat{u}_x)_j^2 + C_3 \hat{u}_j^3 + C_4 \left(\hat{u}_x\right)_j^3), \tag{9.6.2}$$

where $\hat{u}_x = T\hat{u}/\Delta x$ is a compact finite-difference approximation. The PDE is then written as a set of the Hamiltonian ODEs in skew-gradient form

$$\dot{\hat{u}} = K(\hat{u}) \nabla_{\hat{u}} \hat{H}_{\Delta x}. \tag{9.6.3}$$

Here, $K(\hat{u}) = (UD^{(1)} + D^{(1)}U)$ represents the discrete version of the coadjoint operator in equation (9.3.1), where $U = \mathrm{diag}(\hat{u})$ is a diagonal matrix with $\hat{u}_j$ on the jth diagonal and the matrix $D^{(1)}$ is a centered finite-difference matrix with the stencil $[-\frac{1}{2\Delta x}, 0, \frac{1}{2\Delta x}]$ on the main three diagonals and $-\frac{1}{2\Delta x}$ and $\frac{1}{2\Delta x}$ on the top right and bottom left corners, respectively. This yields a skew-symmetric tri-diagonal matrix $K(\hat{u})$ given as

$$\frac{1}{2\Delta x}\begin{pmatrix} 0 & u_1+u_2 & & & -u_n-u_1 \\ -u_1-u_2 & 0 & u_2+u_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -u_{n-2}-u_{n-1} & 0 & u_{n-1}+u_n \\ u_1+u_n & & & -u_{n-1}-u_n & 0 \end{pmatrix},$$

where the diagonal dots denote the continuation of the stencil $[-u_{i-1}-u_i, 0, u_i+u_{i+1}]$ on the ith row. Note that

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{H} = (\nabla_{\hat{u}}\hat{H})^{\mathrm{T}}\dot{\hat{u}} = (\nabla_{\hat{u}}\hat{H})^{\mathrm{T}}K(\hat{u})\nabla_{\hat{u}}\hat{H} = 0, \tag{9.6.4}$$

hence, $\hat{H}$ is a first integral of this ODE. Finally, equation (9.6.3) is integrated using the implicit midpoint rule, which is solved using Newton iterations. This method will henceforth be referred to as the *conventional method.*

The conventional and collective methods are both order-two in space as shown by figure 9.6.1, which show errors for travelling wave solutions of the cubic Hamiltonian system outlined in section 9.6.2. The Hamiltonian error at time $t = t_n$ is calculated by $(\hat{H}(0) - \hat{H}(t_n))/\hat{H}(0)$ and similarly for the Casimir error. The solution error is

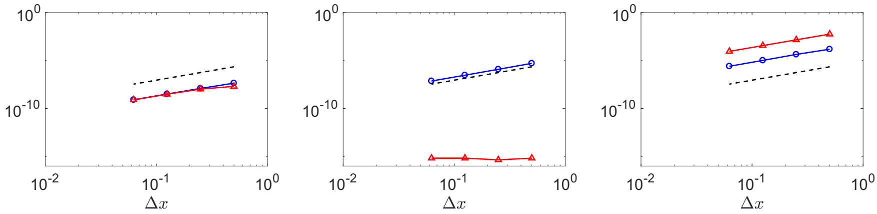
$$\frac{||\hat{u}_n - \hat{u}_e||_2}{||\hat{u}_e||_2},$$

where $\hat{u}_n$ is the numerical solution, $\hat{u}_e$ is the exact solution evaluated on the grid and $||\cdot||_2$ is the discrete L_2-norm. We see from figure 9.6.1b that the collective method preserves the energy up to machine precision for this experiment. We remark that the solution error observed in figure 9.6.1c is largely attributed to phase error and does not reflect the ability of the method to preserve the shape of the travelling wave.

9.6.1 Inviscid Burgers' equation

Setting $C_1 = 1$, $C_2 = 0$, $C_3 = 0$ and $C_4 = 0$ in equation (9.6.1) yields the well-known inviscid Burgers' equation

$$u_t = 6uu_x.$$

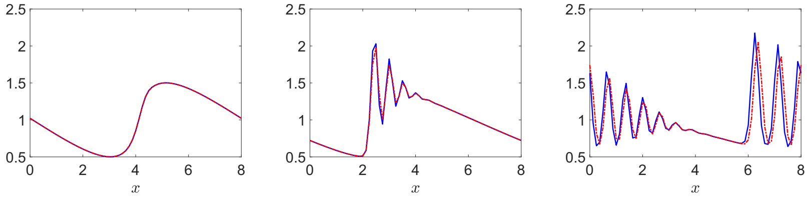


(a) Casimir error

(b) Hamiltonian error

(c) Solution error

Figure 9.6.1: Order-two convergence for the travelling wave solution of the extended Burgers' equation outlined in section 9.6.2. The plots correspond to the conventional solution () and the collective solution () and an order-two reference line (– – –). The error is calculated after 512 timesteps, with $L = 8$, $\Delta t = 2^{-14}$ and $\Delta x = L/2^k$ for $k = 1, 2, 3$ and 4.



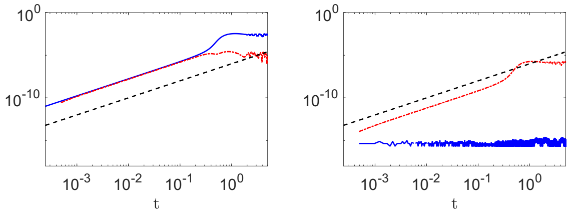
(a) $t = 0.317$

(b) $t = 0.635$

(c) $t = 1.37$

Figure 9.6.2: Inviscid Burgers' equation solutions of the conventional method (———) and collective method (- - - -). The grid parameters are $n_x = 64$, $\Delta x = 0.125$, $L = 8$ and $\Delta t = 2^{-12}$. A shock forms at about $t = 0.4$.

In the following example, the equation is modelled with the initial conditions $u(0,x) = 1 + \frac{1}{2}\cos(2\pi x/L)$, which develops a shock wave at about $t = 0.4$. Figure 9.6.2 shows three snapshots of the conventional and the collective solutions before and after the shock and figure 9.6.3 shows the Casimir and Hamiltonian errors over time. Over the short simulation time, both methods yield qualitatively similar solutions and it is difficult to tell them apart. Due to the presence of shock waves in the inviscid Burgers' equation, it is difficult to gain a sense of the long term behaviour of the methods as no solution exists after a finite time. From figure 9.6.3b we see that the conventional method has exceptional Hamiltonian preservation properties and maintains the error at machine precision throughout the simulation. This can be explained by the fact that the implicit midpoint rule preserves quadratic invariants, that is, $\hat{H}$ is preserved exactly by the conventional method. Otherwise, the errors grow quadratically until the shock develops, after which, they appear bounded. The Hamiltonian error of the collective solution can also be reduced to machine precision by reducing the time step Δt.



(a) Casimir error **(b)** Hamiltonian error

Figure 9.6.3: The errors corresponding to the conventional (——) and collective (- - - -) methods for the inviscid Burgers' equation and $\mathcal{O}(t^2)$ reference lines (— — —).

9.6.2 Extended Burgers' equation

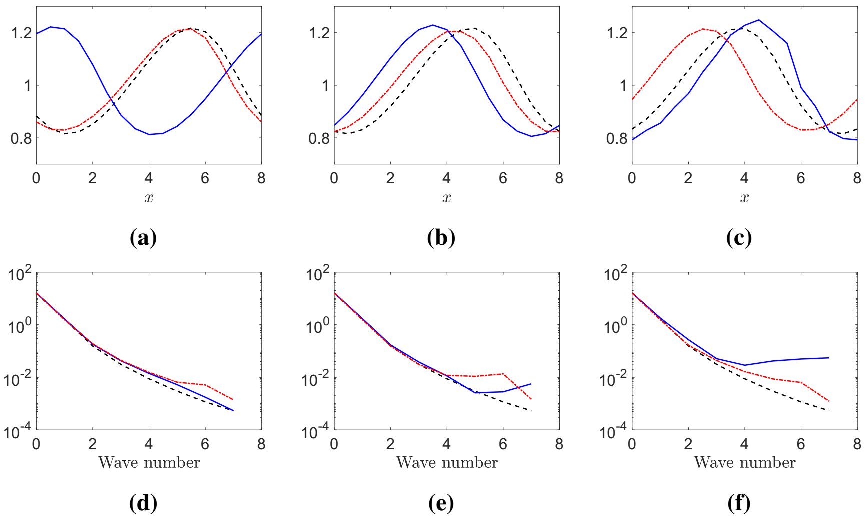
We now focus our attention to a cubic Hamiltonian problem that we have designed to admit non-symmetric travelling wave solutions. The PDE being modelled arises from setting $C_1 = 1/2$, $C_2 = 1/2$, $C_3 = -1/4$ and $C_4 = 1/2$ in equation (9.6.1), which yields

$$u_t = 3uu_x - \frac{9}{4}u^2 u_x - u_x u_{xx} - 3u_x^2 u_{xx} - 2uu_{xxx} - 2uu_x u_{xxx} - 6uu_{xx}^2$$

and is henceforth referred to as the *extended Burgers' equation*.

Travelling wave solutions

In this example, we look for solutions of the form $u(x,t) = f(s)$, where $s = x - ct$ for wave velocity c. This yields an ODE in s, which is solved to a high degree of accuracy on the grid using MATLAB's `ode45`. Figure 9.6.4 shows snapshots of travelling wave solutions to the extended inviscid Burgers' equation and their Fourier transforms and figure 9.6.5 shows the corresponding errors. The main observations concerning these figures is that the errors of the collective solution are bounded whereas the conventional solution errors grow with time. In particular, the high frequency Fourier modes of the conventional solution erroneously drift away from that of the exact solution while the collective solution does a reasonably good job at keeping these modes bounded. These erroneously large high frequency modes can be seen with the naked eye in figure 9.6.4c. This is again highlighted by figure 9.6.5c, which shows that the highest frequency mode (i.e., the mode whose wavelength is equal to the grid spacing Δx) grows exponentially in time. Figures 9.6.5a and 9.6.5b show the behaviour of the Casimir and Hamiltonian errors. This highlights the ability of the collective method to keep the errors bounded, while the errors of the conventional solution grow linearly with time. Towards the end of the simulation, the errors of the conventional solution become so large that the implicit equations



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Figure 9.6.4: Travelling wave solutions of the perturbed Burgers' equation (top row) and the positive Fourier modes (bottom row) at $t = 109$ (left column), $t = 218$ (middle column) and $t = 437$ (right column). The plots correspond to the conventional method (——), collective method (- - - -) and the exact travelling wave solution (— — —). The grid parameters are $n_x = 16$, $\Delta x = 0.5$, $L = 8$ and $\Delta t = 2^{-6}$.

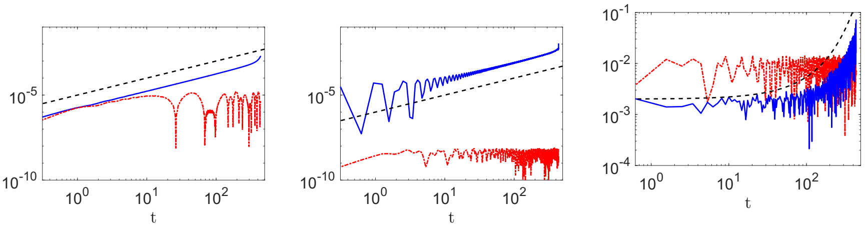
arising from the midpoint rule become too difficult to solve numerically and the Newton iterations fail to converge. The simulation ends with the conventional method errors diverging to infinity.

Periodic bump solutions

In this example, we model solutions to the extended Burgers' equation from the initial condition

$$u(x,0) = 1 + \frac{1}{2}\exp(-\sin^2(\frac{\pi x}{L})).$$

Figure 9.6.6 shows snapshots of the solution and its positive Fourier modes and figure 9.6.7 shows the behaviour of the Casimir and Hamiltonian errors over time. Like the travelling wave example, we see that the high frequency modes of the conventional solution grow with time, which can be seen as rough wiggles in figure 9.6.6c. The conventional solution has bounded Hamiltonian error, despite linear and exponential growth in the Casimir and highest frequency Fourier modes, respectively. In particular, the collective solution has excellent error behaviour, which appears to be bounded over the simulation period for all three plots of figure 9.6.7.



(a) Casimir error

(b) Hamiltonian error

(c) Highest frequency Fourier mode (increased vertical scale)

Figure 9.6.5: The errors corresponding to the conventional (───) and collective (- - -) methods for the travelling wave experiment. The reference lines (– – –) are $\mathcal{O}(t)$ in figures (a) and (b) and exponential in figure (c).

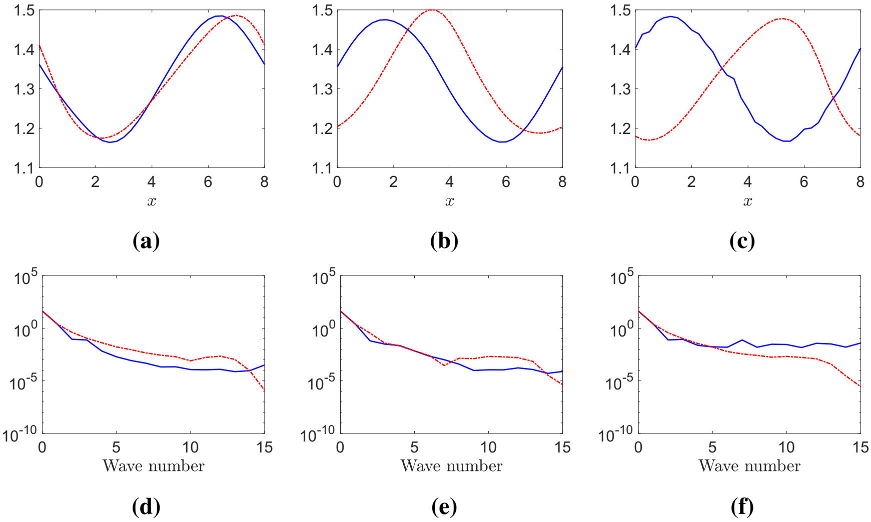
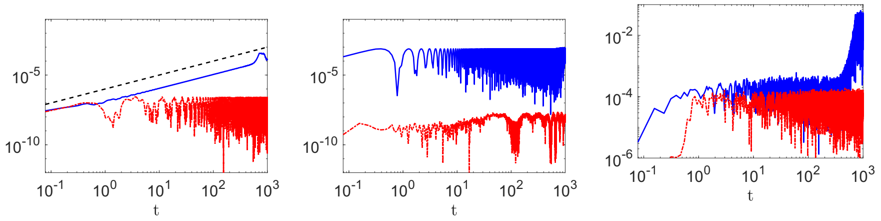


Figure 9.6.6: Periodic bump solutions of the extended Burgers' equation (top row) and the positive Fourier modes (bottom row) at $t = 10$ (left column), $t = 100$ (middle column) and $t = 1000$ (right column). The plots correspond to the conventional method (───) and the collective method (- - -). The grid parameters are $n_x = 32$, $\Delta x = 0.25$, $L = 8$ and $\Delta t = 2^{-8}$.



(a) Casimir error

(b) Hamiltonian error

(c) Highest Fourier frequency mode (increased vertical scale)

Figure 9.6.7: The errors corresponding to the conventional (——) and collective (- - - -) methods for the periodic bump example. The reference line (— — —) in figure (a) is $\mathcal{O}(t)$.

9.7 Conclusion

We have demonstrated that Hamiltonian PDEs on Poisson manifolds can be integrated while maintaining the structure preserving properties of Poisson systems very well. This is achieved by

1. realising the Poisson-Hamiltonian system as an infinite-dimensional, collective Hamiltonian system on a symplectic manifold and lifting the initial condition from the Poisson system to the collective system,

2. discretising the collective system in space to obtain a system of Hamiltonian ODEs, and

3. using a symplectic integrator to solve the system.

The symplectic integrator will, in general, fail to preserve the fibration provided by the realisation. Therefore, the presented integrators for Hamiltonian PDEs cannot be expected to conserve the Poisson structure *exactly*. This is in contrast to the case of Hamiltonian ODEs on Poisson manifolds, where the fibres can be structurally simple for carefully chosen realisations and genuine Poisson integrators can be constructed. Regardless, in the ODE as well as in the PDE case the integrator is guaranteed to inherit the excellent energy behaviour from the symplectic integrator which is applied to the collective system. Moreover, our numerical examples for Hamitonian PDEs show excellent Casimir behaviour as well. Indeed, energy as well as Casimir errors are bounded in long term simulations.

Structure preserving properties of conventional numerical schemes typically rely on the presence of structurally simple symmetries of the differential equation. If the discretisation is invariant under the same symmetry as the equation, then the numerical solution will share all geometric features of the exact solution which are due to the symmetry. The simple form of the symmetries, however, is immediately destroyed when higher order terms in the Hamiltonian are switched on. Although exact solutions still preserve the Hamiltonian, numerical solutions obtained using a traditional scheme fail to show a good energy behaviour. The advantage of the presented integration methods is that their excellent energy behaviour is guaranteed no matter how complicated the Hamiltonian is. Our numerical examples for the extended Burgers' equation demonstrate the importance of structure preservation: while growing energy errors of the conventional solution cause a blow up, there are no signs of instabilities for the collective solution.

Bibliography

[1] L. Brugnano, M. Calvo, J. Montijano and L. Rández, Energy-preserving methods for poisson systems, *Journal of Computational and Applied Mathematics*, **236** (2012), 3890 – 3904, URL http://www.sciencedirect.com/science/article/pii/S0377042712001008, 40 years of numerical analysis: "Is the discrete world an approximation of the continuous one or is it the other way around?".

[2] A. Chern, F. Knöppel, U. Pinkall, P. Schröder and S. Weißmann, Schrödinger's smoke, *ACM Transactions on Graphics (TOG)*, **35** (2016), 77, URL https://doi.org/10.1145/2897824.2925868.

[3] D. Cohen and E. Hairer, Linear energy-preserving integrators for poisson systems, *BIT Numerical Mathematics*, **51** (2011), 91–101, URL https://doi.org/10.1007/s10543-011-0310-z.

[4] M. Dahlby, B. Owren and T. Yaguchi, Preserving multiple first integrals by discrete gradients, *Journal of Physics A: Mathematical and Theoretical*, **44** (2011), 305205, URL http://stacks.iop.org/1751-8121/44/i=30/a=305205.

[5] D. M. de Diego, Lie-poisson integrators, 2018.

[6] E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 2013, URL http://dx.doi.org/10.1007/3-540-30666-8.

[7] B. Khesin and R. Wendt, *Infinite-Dimensional Lie Groups: Their Geometry, Orbits, and Dynamical Systems*, 47–153, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, URL https://doi.org/10.1007/978-3-540-77263-7_2.

[8] A. Kriegl and P. W. Michor, *The Convenient Setting of Global Analysis*, vol. 53, AMS, 1997, URL `http://dx.doi.org/10.1090/surv/053`.

[9] E. Kuznetsov and A. Mikhailov, On the topological meaning of canonical clebsch variables, *Physics Letters A*, **77** (1980), 37 – 38, URL `http://www.sciencedirect.com/science/article/pii/0375960180906271`.

[10] J. Marsden and A. Weinstein, Coadjoint orbits, vortices, and Clebsch variables for incompressible fluids, *Physica D: Nonlinear Phenomena*, **7** (1983), 305 – 323, URL `http://www.sciencedirect.com/science/article/pii/0167278983901343`.

[11] J. E. Marsden and R. Abraham, *Foundations of Mechanics*, 2nd edition, Addison-Wesley Publishing Co., Redwood City, CA., 1978, URL `http://resolver.caltech.edu/CaltechBOOK:1987.001`.

[12] J. E. Marsden, S. Pekarsky and S. Shkoller, Discrete euler-poincaré and lie-poisson equations, *Nonlinearity*, **12** (1999), 1647, URL `http://stacks.iop.org/0951-7715/12/i=6/a=314`.

[13] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*, Springer New York, New York, NY, 1999, URL `http://dx.doi.org/10.1007/978-0-387-21792-5`.

[14] R. I. McLachlan, Spatial discretization of partial differential equations with integrals, *IMA Journal of Numerical Analysis*, **23** (2003), 645–664, URL `http://dx.doi.org/10.1093/imanum/23.4.645`.

[15] R. I. McLachlan, K. Modin and O. Verdier, Collective symplectic integrators, *Nonlinearity*, **27** (2014), 1525, URL `http://stacks.iop.org/0951-7715/27/i=6/a=1525`.

[16] C. Vizman, Geodesic equations on diffeomorphism groups, URL `https://doi.org/10.3842/SIGMA.2008.030`.