Numerical Simulation of Random Solutions to Nonlinear Differential Equations

Moe Küchemann-Scales

This thesis is submitted for the degree of
Doctor of Philosophy

School of Mathematical Sciences
Lancaster University
December 2024

Numerical Simulation of Random Solutions to Nonlinear Differential Equations

Moe Küchemann-Scales

Abstract

This thesis focusses on simulation of weak solutions to nonlinear differential equations. Methods employed include reformulation into stochastic differential equations, and transport of measures via pushforward maps and an action principle. The common thread to approaches discussed is to consider an ensemble of solutions and more generally the distribution that this ensemble will take, rather than specific solutions themselves. A second focus of the work is to restrict solutions to the surface of the sphere $\mathbb{S}^2$, simplifying the use of Fourier series but posing difficulties for the long term stability of numerical algorithms. Numerical simulations of random solutions to the nonlinear Schrödinger equation (NLSE) and the isentropic Euler equations of fluid motion are carried out. For the NLSE, in the case of $\beta = 0$, the empirical distribution is compared with the theoretic distribution of solutions (the Gibbs measure) statistically. Formulating the problem on the sphere involves a Lax pair, one of which is the Frenet-Serret matrix and the second a result of the Hasimoto transform applied to the NLSE. In the case of the Euler equations, the numerical simulation is compared with evolution of a known closed form solution in the one dimensional dam break problem. The proposed algorithm uses optimal transport theory and convexity over a space of absolutely continuous measures in order to allow weak solutions to the differential equations that are distributions in the sense of Schwartz. The simulation is found to closely follow the well established Ritter solutions to the dam break problem.

Contents

Acknowledgements

I would like to thank my supervisors Gordon Blower and Azadeh Khaleghi for the chance to take on this project, and all their help to complete it. Thanks also go to my friends and family for editing the manuscript, even through Christmas. Finally, thanks to Matthew Buck for a number of helpful discussions about geometry.

List of notation

The following is a list of notation used throughout the thesis.

κ *Curvature* defined in Definition 1.2.2.

τ *Torsion* defined in Section 1.2.

κ_n, κ_g The *normal* and *geodesic* curvatures, defined in Definition 2.1.2.

(θ, φ) The first coordinate θ denotes the *longitude*, the angle from the positive x axis. The second coordinate φ is the *colatitude*, the angle from the positive z axis. This is a convention for spherical polar coordinates, see Section 2.2.1.

$C^q(A, B)$ Continuous functions $f : A \to B$ that are q times differentiable.

Ad The *adjoint map*, defined in Definition 2.3.10.

ad The *adjoint representation* defined in Definition 2.3.15.

$M_n(\mathbb{R})$ The set of square $n \times n$ matrices which have entries within $\mathbb{R}$.

$GL_n(\mathbb{R})$ The General Linear group of invertible matrices on $\mathbb{R}$.

$SO(n)$ The Special Orthogonal group of dimension n, see Definition 2.3.8

$\mathfrak{so}(n)$ The Lie algebra of the Lie group $SO(n)$, see Definition 2.3.9

$\mathcal{D}_A(B)$ The *derivative of the matrix exponential map* at A in the direction of B, where A, B are matrices in the Lie algebra, see Equation (2.23).

$\mathcal{E}_A(B)$ The inverse to $\mathcal{D}_A(B)$, as given in Lemma 2.3.22.

$\mathcal{M}(\mathbb{R}^n)$ The set of probability measures on $\mathbb{R}^n$, see above Definition 3.1.6.

$\mathcal{M}_2(\mathbb{R}^n)$ The set of probability measures on $\mathbb{R}^n$ which have finite second moments, see Definition 3.1.6.

$\mathcal{M}_{2,K}(\mathbb{R}^n)$ The set of probability measures on $\mathbb{R}^n$ which have second moments bounded by K, see Definition 3.1.8.

$\mathcal{P}_2(\mathbb{R}^n)$ The subset of $\mathcal{M}_2(\mathbb{R}^n)$ which are absolutely continuous with respect to Lebesgue measure, see Definition 3.1.7.

$\mathcal{P}_{2,K}(\mathbb{R}^n)$ A bounded subset of $\mathcal{P}_2(\mathbb{R}^n)$, with measures having smaller second moments than K, see Definition 3.1.9.

$\mathcal{P}^\gamma(\mathbb{R}^n)$ A set of probability measures which are within $L^\gamma(\mathbb{R}^n)$, see Definition 3.1.10.

$\mathcal{P}^{\gamma,L}(\mathbb{R}^n)$ A bounded subset of $\mathcal{P}^\gamma(\mathbb{R}^n)$, in which their L^γ norm does not exceed L, see Equation (3.5).

$\mathcal{P}^{\gamma,L}_{2,K}(\mathbb{R}^n)$ The intersection $\mathcal{P}^{\gamma,L}(\mathbb{R}^n) \cap \mathcal{P}_{2,K}(\mathbb{R}^n)$, in which the notation allows for more combinations, see Definition 3.1.11.

$\mathcal{P}^G(\mathbb{R}^n)$ A bounded subset of $\mathcal{P}^{\gamma,L}_{2,K}(\mathbb{R}^n)$, in which $G(\rho)$ does not exceed L, see Definition 12.3.4.

$W_q(\mu,\nu)$ The Wasserstein distance between measures μ and ν as defined in Equation (4.4)

$W_q(F,G)$ The Wasserstein distance between one dimensional measures with cumulative distribution functions F and G as discussed in Remark 4.1.6.

σ The integral of τ over space, simplifying the Hasimoto transform, see Equation (5.11).

$\mathcal{U}(\rho)$ The *internal energy* of a fluid as a function of the fluid density ρ.

$U(\rho)$ The *internal energy density*, related to the internal energy as shown in Equation (12.1) and often referred to as the internal energy itself. An explicit definition in the case of Isentropic Euler is given by Definition 6.1.2.

$P(\rho)$ The *pressure* of a fluid as a function of fluid density ρ, defined in terms of internal energy density by Equation (6.4).

$\mathcal{W}$ *Wiener loop measure*, defined in Proposition 8.2.5.

$\nu_{\beta,K}$ The *Gibbs measure*, see Definition 8.3.1.

Introduction

This thesis is an exploration of contemporary methods for numerical simulation of nonlinear differential equations. Specifically, it focusses on two well known dynamic systems, the nonlinear Schrödinger equation (NLSE) and the Euler equations of fluid motion.

Smooth solutions to systems of partial differential equations (PDEs), particularly linear PDEs are well dealt with by standard Runge Kutta methods and finite element analysis. Nonlinear PDEs can have solutions which involve shocks — solutions which are discontinuous on sets of zero measure. The isentropic Euler equations are known to form discontinuities even for smooth initial conditions [16, §5] and as such different approaches are required to deal with them.

This work is differentiated from the larger canon of methods for numerical integration of PDEs in two main ways. One is geometric, the space of interest in this work is the sphere $\mathbb{S}^2$. When searching for solutions to a dynamic system on the surface of the sphere numerical methods designed for an embedding into Euclidean space will produce compounding errors unless they are adapted to the geometry. Instead, employing the theory of Lie algebra's to move from tangent space to base manifold while respecting the geometry of the the space provides substantial benefits to long term stability of the methods [32] [55].

The second way the methods developed here are differentiated is by allowing solutions from a broader space of functions than the continuous and twice differentiable ones required for well known existence theorems. Examples include solutions in the Sobolev space H^1 or stochastic processes for the NLSE, and measures which are absolutely continuous with respect to Lebesgue measure for the isentropic Euler equations. The weak interpretation of the PDEs that extends solutions to these spaces can allow recasting the problem as governing the flow of a probability distribution over time.

This is done via stochastic processes when the dynamical system can be reformulated into something resembling a stochastic differential equation (SDE). Another direction taken is inspired by Hamiltonian dynamics, instead of employing calculus of variations and being left with a set of PDEs, the system is reformulated as a convex optimisation problem with the use of optimal transportation. Solutions to the system are actually pushforward maps between measures describing the evolution of the initial distribution.

Chapters 1 and 2 cover the geometric ideas needed in this thesis, from the Frenet-Serret frame to geodesic and normal curvature. Lie group theory for the sphere is developed and the construction of exponential based solutions to ordinary differential equations (ODEs) in this context is discussed by the Magnus expansion. Chapter 3 introduces the background needed on measure theory, as well as the weak compactness criteria needed for transport of measures in later chapters, and Chapter 4 describes the basics of optimal transport theory, the Monge-Kantorovich problem and Brenier's theorem for quadratic cost. Chapter 5 constructs a matrix Lax pair for the NLSE thanks to the Hasimoto transform. Attention then turns to the Euler equations and exact solutions on the sphere and in the case of a dam break are discussed in Chapters 6-7. Bourgain[10] asserts the measure of the function space of solutions to the periodic NLSE is the Gibbs measure, and this is constructed in Chapter 8. Chapters 9 then gives further details relating to Bourgain's assertion, and shows weak solutions to the Lax pair formulation of the NLSE exist.

Chapter 10 outlines the first numerical method, employing a stochastic differential equation based on the Lax pair developed in Chapter 5. The numerical method developed for this SDE on the sphere is implemented in MATLAB and the distribution of results is compared empirically with the theory. This is done statistically using a large number (10^6) of sample paths of the process.

Chapters 12 and 13 explore optimal transport theory in the case of Wasserstein distance W_2 and then construct a convex structure on a space of measures to allow the theory to be applied in the context of the Euler equations on Euclidean space. Chapter 14 concludes the discussion of the optimal transport based method first developed by Gangbo [27] by applying the method to the dam break problem and evaluating the results.

Overall three numerical methods are built,

- Periodic solutions to the NLSE in three dimensions with initial data on the sphere are simulated via stochastic process, for $\beta = 0$, the distribution of solutions (sample paths) is found to coincide with surface area measure on the sphere.
- A numerical method for simulation of the dynamics of the Fourier modes of a solution to the NLSE is conceived of, and with more computing power could be compared with the Gibbs measure.
- A transport based method for simulating the isentropic Euler equations using discontinuous densities is shown to exhibit realistic behaviour, matching the Ritter solutions for the dam break problem under suitable initial conditions.

Chapter 1

The Frenet-Serret frame

The Frenet-Serret frame is a triplet of orthogonal vectors, and a first order differential equation which governs the motion of this frame thanks to its curvature and torsion. This is one approach to describing the motion of a curve in $\mathbb{R}^3$, where the position of the curve is at $(0, 0, 0)$ with respect to the Frenet-Serret frame. This chapter presents the background needed to formulate the frame and its motion. This formulation will be used in later chapters for both the isentropic Euler equations and the nonlinear Schrödinger equation. The mathematics developed in this section is bookwork which can be found in sources such as Pressley [61].

1.1 Arc length reparametrisation

A curve γ is a smooth function, parameterised by time t,

$$\begin{aligned} \gamma \ : \ [0, \tau] &\to \mathbb{R}^3. \\ t &\mapsto \gamma(t). \end{aligned}$$

Unless defined otherwise it will be assumed it is twice differentiable.

Definition 1.1.1. A *regular* curve is defined as a curve with nowhere zero speed. Symbolically,

$$\|\dot{\gamma}(t)\| \neq 0, \quad \forall t \in [0, \tau].$$

The derivative of the curve, known as the velocity, is a vector quantity pointing

tangentially to (in the direction of) the curve at each point, with magnitude equal to the speed of the curve.

Definition 1.1.2. A *unit-speed* curve is a curve with speed 1 everywhere. $\|\dot{\gamma}(t)\| = 1, \forall t \in [0, \tau]$.

Definition 1.1.3. The *arc-length* of a curve $\gamma(t)$ is the distance an observer will have travelled if following the curve from $t = 0$ up to the present time, t. It is given by the function $s(t)$,

$$s(t) = \int_0^t \|\dot{\gamma}(u)\| du. \tag{1.1}$$

The derivative of the arc length with respect to time gives the speed of the curve

$$\frac{ds}{dt} = \frac{d}{dt} \int_0^t \|\dot{\gamma}(u)\| du = \|\dot{\gamma}(t)\|. \tag{1.2}$$

Proposition 1.1.4 (Pressley)**.** *[60, Cor 1.3.7] Every regular curve $\gamma(t)$ will have a unit-speed parametrisation $\tilde{\gamma}(f(t))$ and the reparametrisation f will be equal to $s(t)$ up to a constant.*

One can express the velocity of a curve with respect to its unit speed parameterisation like so,

$$\tilde{\gamma}'(s) = \frac{d}{ds}\tilde{\gamma}(s(t)) = \frac{1}{\frac{ds}{dt}} \frac{d}{dt}\gamma(t) = \frac{1}{\|\dot{\gamma}(t)\|}\dot{\gamma}(t). \tag{1.3}$$

This gives a unit vector pointing in the tangential direction to the curve, and acts as the starting point for the Frenet-Serret frame.

1.2 The Tangent, Normal and Binormal vectors

In this section, the curve $\gamma(t)$ is assumed to be unit speed.

Definition 1.2.1. The *tangent* to the curve, denoted $\mathbf{t}$ is the unit length vector pointing tangentially to the curve at each point. As such it can be defined in terms of γ as $\mathbf{t} = \dot{\gamma}$.

A unit speed curve $\gamma(t)$ has no acceleration in the tangent direction (otherwise it would speed up!), this can be verified by taking the derivative of the relation $\langle \dot{\gamma}, \dot{\gamma} \rangle = 1$ with respect to t. This gives the relation $\langle \ddot{\gamma}, \dot{\gamma} \rangle = 0$ and therefore the curves acceleration, $\ddot{\gamma}$ lies in the plane perpendicular to $\mathbf{t}$.

Definition 1.2.2. The direction of the acceleration of the unit speed curve $\gamma(t)$ is called the *normal* to the curve, denoted $\mathbf{n}$. The magnitude of this vector is known as the curvature, and is denoted κ. To enforce the unit length requirement, $\mathbf{n} = \ddot{\gamma}/\|\ddot{\gamma}\|$, and the magnitude $\kappa = \|\ddot{\gamma}\|$.

Definition 1.2.3. The orthonormal vectors $\mathbf{t}$ and $\mathbf{n}$ can be extended to a co-moving basis of $\mathbb{R}^3$ with use of the cross product. A choice of orientation is all that is left to define the *binormal* vector $\mathbf{b}$,

$$\mathbf{b} = \mathbf{t} \times \mathbf{n},$$

or in terms of the unit speed curve

$$\mathbf{b} = \frac{\dot{\gamma} \times \ddot{\gamma}}{\|\ddot{\gamma}\|}.$$

Due to the fact both $\|\mathbf{t}\| = 1$ and $\|\mathbf{n}\| = 1$, $\mathbf{b}$ is also unit length and $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$ form an orthonormal basis of $\mathbb{R}^3$.

By the definition, it is clear that the normal is related to the derivative of the tangent via the expression

$$\mathbf{n} = \frac{\dot{\mathbf{t}}}{\kappa}, \tag{1.4}$$

and the derivative of $\mathbf{n}$ can be calculated using its definition in terms of the unit speed curve.

$$\begin{aligned}
\frac{d}{dt}\mathbf{n} &= \frac{\dddot{\gamma}\|\ddot{\gamma}\|^2 - \ddot{\gamma}\langle\ddot{\gamma}, \dddot{\gamma}\rangle}{\|\ddot{\gamma}\|^3}, \\
&= \langle\frac{\dddot{\gamma}}{\|\ddot{\gamma}\|}, \mathbf{t}\rangle\mathbf{t} + \langle\frac{\dddot{\gamma}}{\|\ddot{\gamma}\|}, \mathbf{b}\rangle\mathbf{b}, \\
&= -\kappa\mathbf{t} + \langle\frac{\dddot{\gamma}}{\|\ddot{\gamma}\|}, \mathbf{b}\rangle\mathbf{b},
\end{aligned} \tag{1.5}$$

where the last line follows from taking derivatives of $\langle\ddot{\gamma}, \dot{\gamma}\rangle = 0$.

$$\begin{aligned}
\frac{d}{dt}\langle\ddot{\gamma}, \dot{\gamma}\rangle &= 0, \\
\langle\dddot{\gamma}, \dot{\gamma}\rangle + \|\ddot{\gamma}\|^2 &= 0, \\
\langle\frac{\ddot{\gamma}}{\|\ddot{\gamma}\|}, \dot{\gamma}\rangle &= -\|\ddot{\gamma}\|.
\end{aligned}$$

The final relation needed to build the Frenet-Serret equations for the motion of the $\mathbf{t}, \mathbf{n}, \mathbf{b}$ frame is the derivative of $\mathbf{b}$.

$$\begin{aligned}\frac{d}{dt}\mathbf{b} &= \frac{d}{dt}\mathbf{t} \times \mathbf{n} + \mathbf{t} \times \frac{d}{dt}\mathbf{n}, \\ &= \mathbf{t} \times \dot{\mathbf{n}}.\end{aligned}$$

This shows that $\dot{\mathbf{b}}$ is orthogonal to $\mathbf{t}$, and the fact $\mathbf{b}$ is a unit vector implies $\dot{\mathbf{b}}$ is orthogonal to $\mathbf{b}$. Thus $\dot{\mathbf{b}}$ points in the direction of $\mathbf{n}$, and the magnitude can be defined as the torsion $\dot{\mathbf{b}} = -\tau\mathbf{n}$. This can also be calculated explicitly, using Equation (1.5)

$$\begin{aligned}&= \mathbf{t} \times \left(-\kappa\mathbf{t} + \langle \frac{\dddot{\gamma}}{\|\ddot{\gamma}\|}, \mathbf{b} \rangle \mathbf{b} \right), \\ &= -\langle \frac{\dddot{\gamma}}{\|\ddot{\gamma}\|}, \mathbf{b} \rangle \mathbf{n}.\end{aligned}$$

Thus the torsion can be given by the inner product,

$$\tau = \langle \frac{\dddot{\gamma}}{\|\ddot{\gamma}\|}, \frac{\dot{\gamma} \times \ddot{\gamma}}{\|\ddot{\gamma}\|} \rangle.$$

Combining the results of this section gives the Frenet-Serret frame.

Proposition 1.2.4. *Any regular curve $\gamma \in C^3(\mathbb{R}, \mathbb{R}^3)$ or smoother can be described by the evolution of a frame,*

$$\frac{d}{ds}\begin{bmatrix}\mathbf{t} \\ \mathbf{n} \\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\mathbf{t} \\ \mathbf{n} \\ \mathbf{b}\end{bmatrix}, \tag{1.6}$$

where $\kappa(s) > 0$ and $\tau(s)$ are in $C^3(\mathbb{R}, \mathbb{R})$. The converse is also true, for any $\kappa, \tau \in C^3(\mathbb{R}^3)$ there exists a unique (up to isometries of $\mathbb{R}^3$) curve γ which is specified by its frame $[\mathbf{t}, \mathbf{n}, \mathbf{b}]$.

Proof. The forward direction is an application of the definitions specified in this section. Provided γ is regular and three times differentiable $\mathbf{t}, \mathbf{n}, \mathbf{b}, \kappa, \tau$ are all well defined. The converse direction is an application of classical ODE existence theorems, for example Birkhoff [4]. Provided κ and τ are differentiable, then with sufficient initial conditions

the ODE has a unique solution. The details of the isometry needed on $\mathbb{R}^3$ can be found in [60, Thm 2.3.6], and amount to translating between initial points of the two potential solutions, and then rotating their frames to match. $\square$

Chapter 2

Spherical geometry

Following on from the previous chapter on Frenet-Serret frames, there are directions that discussion can be developed further for curves which are constrained to stay on the unit sphere. Any two dimensional surface embedded in $\mathbb{R}^3$ has a vector normal to the surface within the ambient space, this offers an alternate choice of normal to the one given in the Frenet-Serret frame. The first section of this chapter discusses this idea. Also included in the chapter is background material on Lie groups and algebras which is used to construct solutions to differential equations constrained to the sphere. The Magnus expansion is the prototypical example, and this is mentioned subsequently. However to understand what is meant by the Magnus expansion, the directional derivative of the exponential map within the Lie algebra is carefully defined. Finally, the chapter covers the Rodriguez formula, a closed form expression for the exponential of a three dimensional skew symmetric matrix. Many of the proofs given in this chapter are basic Lie group theory and can be found in sources such as Stillwell [68].

2.1 Geometry of the sphere

For any two dimensional surface (or smooth Riemannian manifold) embedded in $\mathbb{R}^3$ there exists an alternative co-moving frame for curves travelling along this surface. This comes from the fact that the velocity vector always lies in the tangent space of the surface, and there exists a unique (up to sign) unit normal to this plane in $\mathbb{R}^3$. Thus

the old tangent vector of a unit speed curve (as defined in Definition 1.2.1) and the normal to the surface are always orthogonal and one can construct a frame from these vectors. In this case, no longer does the acceleration of a unit speed curve point in the direction of the 'normal' vector, instead it lies in the plane orthogonal to the tangent. The component of the acceleration in the direction of the 'normal' vector is related to the curvature of the surface as will be discussed directly.

Definition 2.1.1. The unit sphere $\mathbb{S}^{n-1}$ is the boundary of the unit ball, ∂B in $\mathbb{R}^n$, which is defined

$$\partial B(a, r) = \{x \in \mathbb{R}^n \mid \|x - a\| = r\}. \tag{2.1}$$

The unit sphere is $\partial B(0, 1)$, and unless specified, it is assumed $n = 3$.

For the unit sphere, $\mathbb{S}^2$, the normal to the surface is given everywhere by the radial vector, denoted $\hat{r}$. Thus if the curve $\gamma(t)$ is unit speed and constrained to the unit sphere, then the vectors $\{\dot{\gamma}, \hat{r}, \dot{\gamma} \times \hat{r}\}$ form an orthogonal basis for $\mathbb{R}^3$ which moves with the curve.

Definition 2.1.2. The acceleration of a unit speed curve constrained to the surface of the unit sphere can be expressed as,

$$\ddot{\gamma} = \kappa_n \hat{r} + \kappa_g \dot{\gamma} \times \hat{r}, \tag{2.2}$$

where κ_n is defined as the *normal curvature* and κ_g is the *geodesic curvature*.

This relation holds for other surfaces too and at its simplest is just expressing $\ddot{\gamma}$ with respect to the frame $\{\dot{\gamma}, \hat{r}, \dot{\gamma} \times \hat{r}\}$. The acceleration has no component tangentially as it is unit speed ($\langle \ddot{\gamma}, \dot{\gamma} \rangle = 0$). The property that makes the sphere unique is the simple expression for $\hat{n}$, a generic normal vector, as $\hat{r}$ –the radial unit vector.

Lemma 2.1.3. *Consider any unit speed curve $\gamma(t)$ which lies on the sphere. Its normal curvature is $\kappa_n = -1$.*

Proof. Any curve on the sphere has $\langle \gamma, \dot{\gamma} \rangle = 0$ otherwise it would move off the sphere. The derivative of this relation is $\langle \ddot{\gamma}, \gamma \rangle = -\|\dot{\gamma}\|^2$, and in the case of the unit sphere, $\gamma(t) = \hat{r}$. Then $\kappa_n = \langle \ddot{\gamma}, \hat{r} \rangle = \langle \ddot{\gamma}, \gamma \rangle = -\|\dot{\gamma}\|^2$, and recall that γ is unit speed. □

2.2 Frenet-Serret on the sphere

Proposition 2.2.1. *For a unit speed curve $\gamma(t)$ which is on the unit sphere, its curvature and torsion satisfy the relation*

$$\frac{\tau}{\kappa} = \frac{d}{dt}\frac{\dot{\kappa}}{\kappa^2\tau}, \tag{2.3}$$

and the converse is also true.

Proof. As $\gamma(t)$ is in $\mathbb{R}^3$ it can be decomposed into its coordinates with respect to the basis $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$, $\gamma(t) = a(t)\mathbf{t} + b(t)\mathbf{n} + c(t)\mathbf{b}$. Now as $\gamma(t)$ is on the unit sphere, $\langle \gamma, \mathbf{t} \rangle = 0$, and so $a(t) = 0$. Then the derivative of this expression is

$$\langle \gamma(t), \dot{\mathbf{t}} \rangle + \langle \mathbf{t}, \mathbf{t} \rangle = 0,$$

making use of the relation $\dot{\gamma}(t) = \mathbf{t}$ because γ is unit speed

$$\begin{aligned} \langle \gamma(t), \kappa\mathbf{n} \rangle &= -\|\mathbf{t}\|^2, \\ \langle \gamma(t), \mathbf{n} \rangle &= -\frac{1}{\kappa}. \end{aligned}$$

This implies that $b(t) = -1/\kappa$. A further derivative will allow $c(t)$ to be calculated,

$$\begin{aligned} \frac{d}{dt}\langle \gamma(t), \mathbf{n} \rangle &= -\frac{\dot{\kappa}}{\kappa^2}, \\ \langle \mathbf{t}, \mathbf{n} \rangle - \langle \gamma(t), \kappa\mathbf{t} \rangle + \langle \gamma(t), \tau\mathbf{b} \rangle &= \frac{\dot{\kappa}}{\kappa^2}, \\ \langle \gamma(t), \mathbf{b} \rangle &= \frac{\dot{\kappa}}{\kappa^2\tau}. \end{aligned}$$

With reference back to the generic expression for $\gamma(t)$, $c(t) = \dot{\kappa}/\kappa^2\tau$ and thus $\gamma(t) = -(1/\kappa)\mathbf{n} + (\dot{\kappa}/\kappa^2\tau)\mathbf{b}$. To finish the proof, take the derivative a final time,

$$\begin{aligned} \frac{d}{dt}\langle \gamma(t), \mathbf{b} \rangle &= \frac{d}{dt}\frac{\dot{\kappa}}{\kappa^2\tau}, \\ \langle \gamma(t), \mathbf{n} \rangle &= -\frac{1}{\tau}\frac{d}{dt}\frac{\dot{\kappa}}{\kappa^2\tau}, \\ \frac{1}{\kappa} &= \frac{1}{\tau}\frac{d}{dt}\frac{\dot{\kappa}}{\kappa^2\tau}. \end{aligned}$$

The converse can be seen from reversing the steps in the proof. □

2.2.1 Local coordinates

One can study the geometry of the sphere as a Riemannian manifold without a general theory of differential geometry thanks to spherical polar coordinates. The sphere is embedded into $\mathbb{R}^3$ and covered with the two natural surface patches with respect to latitude and longitude. The curvature of the manifold is then dealt with using curvatures already mentioned in this work.

The spherical polar coordinates of $\mathbb{R}^3$ are given by (r, θ, φ) where θ denotes co-latitude - angle from the z axis - and φ denotes longitude - angle from the x axis and r is the radial distance. Any point on the surface of the unit sphere has $r = 1$ and so a local set of coordinates $(\theta, \varphi) \in (0, \pi) \times [0, 2\pi)$ maps the majority of the sphere (a second chart is needed to complete the atlas, which can simply be a translation of this surface patch around the sphere). The embedding map which maps the local coordinates to points in $\mathbb{R}^3$ in which the manifold has been embedded is given by

$$\hat{r} = \begin{pmatrix} \sin(\theta)\cos(\varphi) \\ \sin(\theta)\sin(\varphi) \\ \cos(\theta) \end{pmatrix}. \tag{2.4}$$

As is natural in differential geometry of a smooth Riemannian manifold, the basis of the tangent space will be ∂_θ and ∂_φ in which the meaning of both are the partial derivative operators at a point mapping curves passing through that point into the tangent space.

$$\hat{\theta} = \frac{\partial \hat{r}}{\partial \theta} = \begin{pmatrix} \cos(\theta)\cos(\varphi) \\ \cos(\theta)\sin(\varphi) \\ -\sin(\theta) \end{pmatrix}, \qquad \hat{\varphi} = \frac{1}{\sin(\theta)} \frac{\partial \hat{r}}{\partial \varphi} = \begin{pmatrix} -\sin(\varphi) \\ \cos(\varphi) \\ 0 \end{pmatrix}.$$

And the collection $(\hat{r}, \hat{\theta}, \hat{\varphi})$ form an orthogonal basis of $\mathbb{R}^3$. For future ease, the partial

derivatives of the vectors $\hat{\theta}(\theta, \varphi)$ and $\hat{\varphi}(\theta, \varphi)$:

$$\frac{\partial \hat{\varphi}}{\partial \varphi} = -\zeta = \begin{pmatrix} -\cos(\varphi) \\ -\sin(\varphi) \\ 0 \end{pmatrix}, \qquad \frac{\partial \hat{\theta}}{\partial \theta} = -\hat{r},$$

$$\frac{\partial \hat{\theta}}{\partial \varphi} = \cos(\theta)\hat{\varphi}, \qquad \frac{\partial \hat{\varphi}}{\partial \theta} = 0.$$

The ζ component can be decomposed into its constituent parts by:

$$\begin{aligned} \langle \zeta, \hat{r} \rangle &= \sin(\theta)\hat{\varphi}, \\ \langle \zeta, \hat{\varphi} \rangle &= 0, \\ \langle \zeta, \hat{\theta} \rangle &= \cos(\theta), \end{aligned}$$

making the relation $\frac{\partial \hat{\varphi}}{\partial \varphi} = -\sin(\theta)\hat{r} - \cos(\theta)\hat{\theta}$. Thus, when taking the time derivatives of the vectors $(\hat{\theta}, \hat{\varphi}, \hat{r})$ the resulting vectors are,

$$\begin{aligned} \frac{d}{dt}\hat{\theta} &= -\dot{\theta}\hat{r} + \dot{\varphi}\cos(\theta), \\ \frac{d}{dt}\hat{\varphi} &= -\dot{\varphi}\sin(\theta)\hat{r} - \dot{\varphi}\cos(\theta)\hat{\theta}, \\ \frac{d}{dt}\hat{r} &= \dot{\theta}\hat{\theta} + \dot{\varphi}\sin(\theta)\hat{\varphi}. \end{aligned}$$

A hypothetical smooth curve on the surface of the sphere can be parameterised in intrinsic coordinates $(\theta(t), \varphi(t))$, embedding the sphere in $\mathbb{R}^3$. With use of $\hat{r}$ one can express the curve in Cartesian coordinates by

$$\gamma(t) = \hat{r}(\theta(t), \varphi(t)) = (\sin(\theta(t))\cos(\varphi(t)), \sin(\theta(t))\sin(\varphi(t)), \cos(\theta(t))).$$

Taking the derivative twice with respect to time and expressing the result with respect

to the basis $(\hat{r}, \hat{\theta}, \hat{\varphi})$ equals

$$\begin{aligned}
\frac{d}{dt}\gamma(t) &= \dot{\theta}\hat{\theta} + \dot{\varphi}\sin(\theta)\hat{\varphi}, \\
\frac{d^2}{dt^2}\gamma(t) &= \ddot{\theta}\hat{\theta} + \dot{\theta}\frac{d}{dt}\hat{\theta} + \ddot{\varphi}\sin\theta\hat{\varphi} + \dot{\theta}\dot{\varphi}\cos\theta\hat{\varphi} + \dot{\varphi}\sin(\theta)\frac{d}{dt}\hat{\varphi}, \\
\frac{d^2}{dt^2}\gamma(t) &= \ddot{\theta}\hat{\theta} - \dot{\theta}^2\hat{r} + \dot{\theta}\dot{\varphi}\cos\theta\hat{\varphi} + \ddot{\varphi}\sin\theta\hat{\varphi} + \dot{\theta}\dot{\varphi}\cos\theta\hat{\varphi} \\
&\quad + \dot{\varphi}\sin(\theta)\left(-\dot{\varphi}\sin(\theta)\hat{r} - \dot{\varphi}\cos(\theta)\hat{\theta}\right).
\end{aligned} \tag{2.5}$$

Collecting terms,

$$\begin{aligned}
\frac{d^2}{dt^2}\gamma(t) &= \left(\ddot{\theta} - \dot{\varphi}^2\sin(\theta)\cos(\theta)\right)\hat{\theta} - \left(\dot{\theta}^2 + \dot{\varphi}^2\sin^2(\theta)\right)\hat{r} \\
&\quad + \left(2\dot{\theta}\dot{\varphi}\cos\theta + \ddot{\varphi}\sin\theta\right)\hat{\varphi}.
\end{aligned} \tag{2.6}$$

The component of the acceleration in the direction normal to the surface is $-\dot{\theta}^2 - \dot{\varphi}^2\sin^2(\theta(t)) = -\|\dot{\gamma}\|^2$. This is necessary for the curve to stay on the sphere. The velocity of a curve on a surface must lie in the tangent space to the surface. The tangent space to the sphere at the point given by (θ, φ) in local coordinates is the plane perpendicular to the vector $\hat{r}(\theta, \varphi)$ as viewed extrinsically in $\mathbb{R}^3$. In other words, for any curve $X(t)$, using the Euclidean inner product on $\mathbb{R}^3$, $\langle \gamma, \dot{\gamma} \rangle = 0$. Taking the derivative of said relation implies that $\langle \gamma, \ddot{\gamma} \rangle = -\|\dot{\gamma}\|^2$. This is a special case of the covariant derivative, the derivative of a vector field on a manifold is given by the covariant derivative of the vector field along curves on the manifold. As such $d^2/dt^2\gamma = \nabla_\gamma \dot{\gamma} = \ddot{\gamma} - \langle \dot{\gamma}, N \rangle N$, where N is the normal vector to the surface of the manifold when the manifold is embedded in a higher dimensional Euclidean space (and the inner product is with respect to that space). In the case of the sphere the normal vector is simply $\hat{r}$, and the normal component of $\ddot{\gamma}$ becomes an additional constraint for the curve to stay on the sphere.

2.2.2 Geodesic and normal curvature

Returning to the frame outlined in Section 2.1, the normal and geodesic curvatures of a general curve $\gamma(t) = \hat{r}(\theta(t), \varphi(t))$ can now be calculated explicitly.

Proposition 2.2.2. *In the case of a unit sphere, any unit speed curve along* $\mathbb{S}^2$ *has normal curvature* $\kappa_n = -1$ *and geodesic curvature,*

$$\kappa_g = (\ddot{\varphi}\dot{\theta} - \ddot{\theta}\dot{\varphi})\sin(\theta) + 2\dot{\theta}^2\dot{\varphi}^2\cos(\theta) + \dot{\varphi}^3\sin^2(\theta)\cos(\theta) \tag{2.7}$$

where $\gamma(t) = \hat{r}(\theta(t), \varphi(t))$ *is the unit speed curve expressed in local coordinates on the sphere.*

Proof. Consider the curve constructed in the previous section, $\gamma(t) = \hat{r}(\theta(t), \varphi(t))$, with its acceleration given in Equation (2.6). Assuming now that this curve is in fact unit speed, then $\|\dot{\gamma}\|^2 = \dot{\theta}^2 + \dot{\varphi}^2\sin^2(\theta(t)) = 1$, and so from the $\hat{r}$ component of Equation (2.6) we can deduce that $\kappa_n = -1$. The vector $\dot{\gamma}$ can be expressed in terms of $(\hat{\theta}, \hat{\varphi})$ as in Equation (2.5), $\dot{\gamma} = \dot{\theta}\hat{\theta} + \dot{\varphi}\sin(\theta)\hat{\varphi}$. Taking the cross product with $\hat{r}$ will define $\hat{r} \times \dot{\gamma}$, noting that $\hat{r} \times \hat{\theta} = \hat{\varphi}$ allows us to calculate

$$\hat{r} \times \dot{\gamma} = \dot{\theta}\hat{\varphi} - \dot{\varphi}\sin(\theta)\hat{\theta}.$$

Thus all that is left is to calculate $\kappa_g = \langle \ddot{\gamma}, \hat{r} \times \dot{\gamma} \rangle$ and verify that $\langle \ddot{\gamma}, \dot{\gamma} \rangle = 0$.

$$\begin{aligned}
\langle \ddot{\gamma}, \hat{r} \times \dot{\gamma} \rangle &= \dot{\theta}\langle \ddot{\gamma}, \hat{\varphi} \rangle - \dot{\varphi}\sin(\theta)\langle \ddot{\gamma}, \hat{\theta} \rangle \\
&= \dot{\theta}\left(2\dot{\theta}\dot{\varphi}\cos\theta + \ddot{\varphi}\sin\theta\right) - \dot{\varphi}\sin(\theta)\left(\ddot{\theta} - \dot{\varphi}^2\sin(\theta)\cos(\theta)\right) \\
&= (\ddot{\varphi}\dot{\theta} - \ddot{\theta}\dot{\varphi})\sin(\theta) + 2\dot{\theta}^2\dot{\varphi}^2\cos(\theta) + \dot{\varphi}^3\sin^2(\theta)\cos(\theta).
\end{aligned}$$

This gives the proposition. To verify that our curve remains unit speed and thus all acceleration inside of the tangent space to the sphere is purely perpendicular to the motion we must check that $\langle \ddot{\gamma}, \dot{\gamma} \rangle = 0$.

$$\begin{aligned}
\langle \ddot{\gamma}, \dot{\gamma} \rangle &= \dot{\theta}\langle \ddot{\gamma}, \hat{\theta} \rangle + \dot{\varphi}\sin(\theta)\langle \ddot{\gamma}, \hat{\varphi} \rangle \\
&= \dot{\theta}\left(\ddot{\theta} - \dot{\varphi}^2\sin(\theta)\cos(\theta)\right) + \dot{\varphi}\sin(\theta)\left(2\dot{\theta}\dot{\varphi}\cos\theta + \ddot{\varphi}\sin\theta\right) \\
&= \ddot{\theta}\dot{\theta} + \dot{\theta}\dot{\varphi}^2\sin(\theta)\cos(\theta) + \ddot{\varphi}\dot{\varphi}\sin^2(\theta) \\
&= \frac{1}{2}\frac{d}{dt}\left(\dot{\theta}^2 + \dot{\varphi}^2\sin^2(\theta(t))\right) \\
&= 0.
\end{aligned}$$

□

2.3 Lie group theory for the Magnus expansion

In this section an algebraic analysis of the derivative of the exponential of a matrix is carried out. The *Magnus expansion* gives the impetus behind this formalism, to discuss solutions to matrix differential equations which are given by an exponential of a function. The context of interest in this section is that of the differential equation

$$\frac{d}{dt}X(t) = A(t)X(t), \tag{2.8}$$

for $A(t) \in \mathfrak{so}(n)$ or more generally $A(t) \in M_n(\mathbb{R})$ and $X(t) : [0,t] \to M$ where M could be $\mathbb{R}^n$, or a bounded subspace such as the unit sphere $\mathbb{S}^{n-1}$. In the case of $\mathfrak{so}(n)$ the solution can be given by a vector on $\mathbb{S}^{n-1}$ and in turn this can be represented by an initial vector and a rotation in $SO(n)$. This rotation matrix can then be expressed as the exponential of an element of $\mathfrak{so}(n)$ and the details and validity of this statement are discussed, as well as worked out explicitly for $n = 3$. First the subject is motivated by the method of Picard iteration.

Motivating examples

For an intuition of where the Magnus expansion has been developed from consider the following examples.

Example 2.3.1. Consider the initial value problem,

$$\frac{d}{dt}X(t) = AX(t), \qquad X(0) = X_0, \tag{2.9}$$

where $X \in C([0,t], \mathbb{R}^n)$ and A is a constant matrix $A \in M_n(\mathbb{R})$. As an illustrative exercise we will solve this using successive approximations. In integral form the initial value problem becomes

$$X(t) = X_0 + \int_0^t AX(s)ds \coloneqq TX(t). \tag{2.10}$$

which can be represented as the operator T. The solution X is the function such that $X = TX$, and the approximation can be started with X_0

$$TX_0 = (I + tA)X_0 \coloneqq X_1,$$
$$TX_1 = (I + tA + \frac{t^2}{2}A^2)X_0 \coloneqq X_2,$$

From which a clear pattern emerges,

$$TX_n = \sum_{j=0}^{n} \frac{t^j A^j}{j!} X_0, \tag{2.11}$$

and hence $\lim_{n\to\infty} TX_n = \exp(tA)X_0$.

The problem becomes more complex if the matrix A is no longer constant.

Example 2.3.2. Consider the initial value problem,

$$\frac{d}{dt}X(t) = A(t)X(t), \qquad X(0) = X_0,$$

where $X \in C([0,t], \mathbb{R}^n)$ and $A(t) \in C([0,t], M_n(\mathbb{R})$. Also assume that $A(t)$ commutes with $B(t) \coloneqq \int_0^t A(s)ds$ for all t.

Again using the idea of successive approximations, it is clear that $TX_0 = (I + B(t))X_0$. Applying the operator to X_1,

$$TX_1 = X_0 + \int_0^t A(u)duX_0 + \int_0^t A(u)\left(\int_0^t A(s)ds\right)duX_0.$$

This equation can be reduced with use of the assumption that $A(t)$ and $B(t)$ commute, the third term is simplified using integration by parts,

$$\int_0^t A(u)B(u)du = B(t)B(t) - \int_0^t B(u)A(u)du,$$
$$\int_0^t A(u)B(u)du = \frac{1}{2}B(t)B(t).$$

This reduces the expression for X_2 to $X_2 = (I + B(t) + \frac{1}{2}B(t)^2)X_0$ and the method of

integration by parts can be applied again for subsequent terms using the assumption that $A(t)$ and $B(t)$ commute. This suggests the limit

$$\lim_{n\to\infty} TX_n = \exp(B(t))X_0. \tag{2.12}$$

The assumption on $A(t)$ and its integral commuting for all t is not realistic. If it doesn't hold, then the infinite series will become more complicated, involving nested commutator brackets. This offers a motivation for the Magnus expansion, the examples have suggested the form $X(t) = \exp(\Omega(t))X(0)$ for solutions to the initial value problems posed, for this function to solve a general ODE on $SO(3)$, what differential equation will Ω have to satisfy.

These examples also motivate the classical theorems on existence of solutions to ODEs, the operator T defined in Equation (2.10) is the iterator in Picard's iteration theorem [66].

Theorem 2.3.3 (Picard iteration theorem)**.** *Consider $A(t) : [0,\tau] \to M_n(\mathbb{R})$ continuous and Lipschitz in the sense that for all $t \in [0,\tau]$, $\|A(t)X - A(t)Y\| \leq K\|X - Y\|$ for $X, Y \in R \subset \mathbb{R}^n$ and R a closed and bounded subset. Then there exists a solution $X(t) : [0,\tau] \to R$ to the differential equation*

$$\frac{d}{dt}X(t) = A(t)X(t), \tag{2.13}$$

on an interval $t \in [0,h]$ where $h \leq \tau$.

Proof. First constrain our problem to the interval $[0,h]$ where $h \in [0,t]$ is chosen so that $Kh < 1$. Using the functional T given in equation (2.10), iteratively define

$$\begin{aligned}
X_1 &:= T(X_0) = X_0 + \int_0^h A(s)X_0 ds,\\
X_2 &:= T(X_1) = X_0 + \int_0^h A(s)X_1 ds,\\
&\vdots\\
X_n &:= T(X_{n-1}) = X_0 + \int_0^h A(s)X_{n-1} ds.
\end{aligned}$$

By defining $A(t)$ as a bounded linear function of X on R in the assumptions of the theorem it is clear that $\|A(t)X\| \leq K\|X\| \leq K_2$ for all $X \in R$. Thus $\|X_1 - X_0\| = \|\int_0^h A(s)X_0 ds\| \leq hK_2$, which in turn implies a smaller limit for $\|X_2 - X_1\| = \|\int_0^h A(t)(X_1 - X_0)ds\| \leq hK\|X_1 - X_0\| \leq hKhK_2$. Seeing the pattern, the distance $\|X_n - X_{n-1}\| \leq (hK)^{n-1}hK_2$. Next consider the convergence of the series

$$X = X_0 + \sum_{n=1}^{\infty}(X_n - X_{n-1}),$$

in the Euclidean norm on $\mathbb{R}^n$. By applying the triangle inequality, the norm of X is bounded above by

$$\|X\| \leq \|X_0\| + \sum_{n=1}^{\infty}\|X_n - X_{n-1}\| \leq \|X_0\| + hK_2\sum_{n=0}^{\infty}(hK)^n,$$

so the series X converges. Furthermore the series $X = X(t)$ will converge by the same estimate for any $t \in [0, h]$ hence the convergence is uniform. Lastly, the series X is a solution to the differential equation because

$$\begin{aligned}
0 &= X_n - T(X_{n-1}),\\
X - T(X) &= X - T(X) - X_n + T(X_{n-1}),\\
\|X - T(X)\| &\leq \|X - X_n\| + \|T(X) - T(X_{n-1})\|,\\
\|X - T(X)\| &\leq \|X - X_n\| + \|\int_0^h A(s)(X - X_{n-1})ds\|,\\
\|X - T(X)\| &\leq \|X - X_n\| + hK\|X - X_{n-1}\|.
\end{aligned}$$

As $X_n \to X$ as $n \to \infty$ it is clear that $\|X - T(X)\|$ tends to zero hence X is a solution. □

2.3.1 Lie theory

Define M to be a differentiable manifold. That is a Hausdorff space which is second countable, and has a smooth differentiable atlas. Any function on this space is differentiable thanks to the atlas, and thus the tangent space to the manifold can be

defined.

Definition 2.3.4. A Tangent space to the manifold M at the point $p \in M$ is denoted T_pM. Let $\gamma(t) : \mathbb{R} \to M$ denote all smooth curves on M such that $\gamma(0) = p$. The Tangent space is the set of derivatives of the curves $\dot{\gamma}(0)$. For each vector $a \in T_pM$ there are many curves such that $\dot{\gamma}(0) = a$ therefore the definition can be reduced to the set of equivalence classes of derivatives of curves. Consider $P = \{\gamma : \mathbb{R} \to M \ : \gamma(0) = p\}$ then an equivalence relation $\dot{\gamma}_1 \sim \dot{\gamma}_2$ if $\dot{\gamma}_1(0) = \dot{\gamma}_2(0)$ will produce the equivalence classes and the quotient vector space $P/\sim = T_pM$.

With these concepts the Lie group and Lie algebra can be defined.

Definition 2.3.5. A Lie group, G, is a differentiable manifold M which is also a group and the group operation and inversion are smooth.

Definition 2.3.6. A Lie algebra is a vector space $\mathfrak{g}$ with a Lie bracket $[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ which is linear in both arguments, antisymmetric and satisfies the Jacobi identity,

$$[A, [B, C]] + [B, [C, A]] + [C, [A, B]] = 0, \quad \forall A, B, C \in \mathfrak{g}.$$

For this project only Lie groups which can be represented as matrices (so considered as subgroups of $GL_n(\mathbb{R})$ are needed, which simplifies the definitions). The Lie algebra of a matrix Lie group can be defined as follows

Definition 2.3.7. The Lie algebra of a matrix Lie group, G, consists of the matrices X such that $\exp(tX) \in G$ for all t.

The tangent space of a Lie group is a Lie algebra. We will only consider matrix Lie groups and matrix Lie algebras, for which the exponential map

$$\exp : \mathfrak{g} \to G,$$

which maps the algebra to the group, coincides with the matrix exponential,

$$\exp(A) = \sum_{n=0}^{\infty} \frac{A^k}{k!}.$$

The Lie groups which are discussed most in this thesis are the special orthogonal groups $SO(n)$, and specifically $SO(3)$.

Definition 2.3.8. One can define $SO(n)$ via its matrix representation

$$SO(n) = \{A \in GL_n(\mathbb{R}) \ : AA^\top = I, \det(A) = 1\}. \tag{2.14}$$

This set, with the binary operation of matrix multiplication, forms a Lie group.

Definition 2.3.9. The Lie group $SO(n)$ has accompanying Lie algebra $\mathfrak{so}(n)$, with matrix representation

$$\mathfrak{so}(n) = \{X \in GL_n(\mathbb{R}) \ : X + X^\top = 0\}. \tag{2.15}$$

Definition 2.3.10. The adjoint map of a Lie group, if G is a Lie group and $\mathfrak{g}$ is it's algebra, then for all $A \in G$ there is a linear automorphism of $\mathfrak{g}$, $\mathrm{Ad}_A : \mathfrak{g} \to \mathfrak{g}$ given by

$$\mathrm{Ad}_A(X) = AXA^{-1}. \tag{2.16}$$

Recall that $\mathfrak{g}$ is an algebra over a vector space, so linear maps $\mathfrak{g} \to \mathfrak{g}$ themselves form a group $GL(\mathfrak{g})$. Thus the map $\mathrm{Ad} : A \mapsto \mathrm{Ad}_A$ is a group homomorphism $G \to GL(\mathfrak{g})$.

Remark 2.3.11. For a relevant example, consider the Lie group $SO(3)$. If $A \in SO(3)$ and $X \in \mathfrak{so}(3)$ then $\mathrm{Ad}_A(X) = AXA^\top = -AX^\top A^\top = -(AXA^\top)^\top$, hence $AXA^\top$ is in $\mathfrak{so}(3)$ too.

Lemma 2.3.12. *Any matrix $A \in SO(3)$ can be decomposed into a triplet $A = BC_\theta B^{-1}$ in which $B \in SO(3)$, and $C_\theta \in SO(3)$ is of the specific form which rotates the z axis by an angle θ,*

$$C_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.17}$$

Proof. One can prove the statement concerning the rotation around a fixed axis, and the specific decomposition of any matrix $A \in SO(3)$ by studying its characteristic equation. The equation $p_A(x) = \det(xI - A)$ is a cubic monic polynomial and thus must have at least one real root by the intermediate value theorem. What is more, the intercept of

the graph, $p_A(0) = \det(-A) = -1$. As the polynomial is monic,$p_A(x) \to \infty$ as $x \to \infty$, and so the intermediate value theorem again establishes that $p_A(x)$ should have at least one *positive* real root. For any eigenvalue λ and for $\mathbf{v}$ its eigenvector, $(A\mathbf{v})^\top A\mathbf{v} = \mathbf{v}^\top\mathbf{v}$, but $(A\mathbf{v})^\top A\mathbf{v} = |\lambda|^2\mathbf{v}^\top\mathbf{v}$, so $|\lambda| = 1$. Together these statements imply that $\lambda = 1$ must be an eigenvalue of A, and its accompanying (unit) eigenvector $\mathbf{v}$ is the Euler axis of the rotation for C_θ. The rest of the argument resembles diagonalising or the Jordan normal form. The matrix B can be seen as the change of coordinates from (x, y, z) to $(\mathbf{w}_1, \mathbf{w}_2, \mathbf{v})$ where orthonormal $\mathbf{w}_1$ and $\mathbf{w}_2$ span the plane orthogonal to $\mathbf{v}$. With this definition alone $B \in SO(3)$ and so by the closure property of the group, C_θ must be in $SO(3)$ too. Hence, from the currently known information, C_θ must be of the form

$$C_\theta = \begin{pmatrix} & & 0 \\ & U & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

By the condition that $C_\theta C_\theta^\top = I$, the smaller $U \in M_2(\mathbb{R})$ must satisfy $UU^\top = I$. The determinant relation $\det(C_\theta) = 1$ directly implies $\det(U) = 1$ and thus $U \in SO(2)$. A generic element of $SO(2)$ can be expressed via one parameter θ by the top right elements of C_θ in Equation (2.17). □

Lemma 2.3.13. *The action of $SO(3)$ on itself is by conjugation, or the adjoint map Ad. This action is isomorphic to the space of rotations $SO(2)$*

Proof. Consider the group action Ad_U for some matrix $U \in SO(3)$. For $A \in SO(3)$ the action of Ad_U forms an orbit, the orbit of A is the set $\{Ad_U(A) \ : U \in SO(3)\}$. For each A there is a representative member of the set equal to C_θ by Lemma 2.3.12. The set $\{C_\theta \ : \ \theta \in (0, 2\pi]\}$ is isomorphic to $SO(2)$. □

Lemma 2.3.14. *The exponential map* $\exp : \mathfrak{so}(3) \to SO(3)$ *is surjective.*

Proof. To prove the surjectivity of the exponential map, consider the following relation. For any value of $\theta \in [0, 2\pi]$,

$$\exp \begin{pmatrix} 0 & -\theta & 0 \\ \theta & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.18}$$

If D_θ is the matrix specified, so that $\exp(D_\theta) = C_\theta$ then note that $\exp(BD_\theta B^{-1}) = B\exp(D_\theta)B^{-1}$. This follows from the fact that $BB^\top = I$, which follows from the construction of B where each row is an orthonormal vector. Finally, $D_\theta \in \mathfrak{so}(3)$ and as explained in Remark 2.3.11, so is $\mathrm{Ad}_B(D_\theta)$. So exp is surjective, for all $A \in SO(3)$ there exists $\mathrm{Ad}_B(D_\theta) \in \mathfrak{so}(3)$ such that $\exp(\mathrm{Ad}_B(D_\theta)) = A$. □

Definition 2.3.15. The adjoint representation of a Lie algebra is the linear map $\mathrm{ad}_Y : \mathfrak{g} \to \mathfrak{g}$ which, for each $Y \in \mathfrak{g}$ is given by

$$\mathrm{ad}_Y(X) = [Y, X]\,. \tag{2.19}$$

In the same way as the adjoint map of the group can also be seen as a group homomorphism $G \to GL(\mathfrak{g})$, the adjoint representation of the algebra can be viewed as a map $Y \to \mathrm{ad}_Y$ which maps elements of $\mathfrak{g}$ to linear operators on $\mathfrak{g}$.

Theorem 2.3.16. *[33, Thm 3.18] For any given Lie group homomorphism $\phi : G \to H$ between matrix Lie groups, there exists a unique real linear map between their algebras $\tilde{\phi} : \mathfrak{g} \to \mathfrak{h}$ such that,*

$$\phi(e^X) = e^{\tilde{\phi}(X)}. \tag{2.20}$$

This map has the following property,

$$\tilde{\phi}(X) = \frac{\partial}{\partial t}\phi(\exp(tX))|_{t=0}. \tag{2.21}$$

The Adjoint map Ad is a group homomorphism $G \to GL(\mathfrak{g})$ and the adjoint representation, ad, is its corresponding map between the algebras, $\tilde{\mathrm{Ad}} = \mathrm{ad}$. This is shown by, for all $Y \in \mathfrak{g}$,

$$\begin{aligned}
\frac{\partial}{\partial t}\mathrm{Ad}_{e^{tX}}(Y)\bigg|_{t=0} &= \left(\frac{\partial}{\partial t}e^{tX}\right) Y e^{-tX}\bigg|_{t=0} + e^{tX} Y \left(\frac{\partial}{\partial t}e^{-tX}\right)\bigg|_{t=0} \\
&= XY - YX \\
&= \mathrm{ad}_X(Y)
\end{aligned}$$

and therefore by Theorem 2.3.16,

$$\mathrm{Ad}_{e^X} = e^{\mathrm{ad}_X}. \tag{2.22}$$

2.3.2 The derivative of the exponential map

Let $\mathfrak{g}$ be the Lie algebra to the Lie group G, the derivative of a smooth function lying in the group (at the identity) lies in the algebra. The Lie algebra is its own tangent space, therefore the differential of the exponential function at a point $X \in \mathfrak{g}$ will be an endomorphism,

$$\mathcal{D}_X(Y) = (d\exp)|_X(Y) : \mathfrak{g} \to \mathfrak{g}. \tag{2.23}$$

The function,

$$f : t \mapsto \exp(-X)\exp(X + tY) \tag{2.24}$$

is a map from $\mathbb{R}$ to G if $X, Y \in \mathfrak{g}$, and satisfies $f(0) = I$, therefore $\dot{f}(t) \in \mathfrak{g}$. The function $f : Y \mapsto \dot{f}(0)$ can be identified with $(d\exp)|_X(Y)$, they are both functions $\mathfrak{g} \to \mathfrak{g}$, and this identification

$$(d\exp)|_X(Y) = \left(\frac{\partial}{\partial t}\exp(-X)\exp(X + tY)\right)_{t=0} \tag{2.25}$$

is used in Thm 2.14.3, by Varadarajan to establish the definition,

$$\mathcal{D}_X(Y) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!}\mathrm{ad}_X^k(Y). \tag{2.26}$$

The expression in Equation (2.26) can be proven using the method given by Tuynman [72]. Though said reference does not expand on the subtleties of the convergence properties of a certain power series, which are only true thanks to Weierstrass' Double Series theorem for complex analytic functions [38, p.35].

Lemma 2.3.17. *For a complex matrix $X \in M_n(\mathbb{C})$, the following limit holds uniformly on compact sets as $n \to \infty$,*

$$\frac{\exp(X) - 1}{n(\exp(\frac{X}{n}) - 1)} \to \frac{\exp(X) - 1}{X}, \tag{2.27}$$

where the fraction represents a compact notation for a power series, not division with respect to a matrix.

Proof. The partial geometric sum gives a starting point, for $x \in \mathbb{C}$,

$$\frac{x^n - 1}{n(x-1)} = \frac{1}{n} \sum_{k=0}^{n-1} x^k.$$

The association of $x^{\frac{1}{n}} = \exp(z)$ implies the following,

$$\begin{aligned} \frac{\exp(z) - 1}{n(\exp(\frac{z}{n}) - 1)} &= \frac{1}{n} \sum_{k=0}^{n-1} \exp\left(\frac{zk}{n}\right), \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{zk}{n}\right)^j. \end{aligned}$$

In the case of matrices $X \in M_n(\mathbb{C})$ the power series is

$$\frac{\exp(X) - 1}{n(\exp(\frac{X}{n}) - 1)} = \frac{1}{n} \sum_{k=0}^{n-1} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{k}{n}\right)^j X^j. \tag{2.28}$$

The convergence of this series is not taken for granted, and will be established via Weierstrass' double series theorem. First the complex variable case, exchange of summands leaves the power series,

$$F_n(z) = \sum_{j=0}^{\infty} \left(\frac{1}{n} \sum_{k=0}^{n-1} \frac{1}{j!} \left(\frac{k}{n}\right)^j \right) z^j. \tag{2.29}$$

Express $F_n = \sum_{j=0}^{\infty} u_j(z)$, then each $u_j(z)$ is analytic on the disc $\{z \in \mathbb{C} : |z| < R\}$ as on said disc, and for all n,

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} \left(\frac{k}{n}\right)^j \frac{z^j}{j!} \right| \leq \left(\frac{1}{n} \sum_{k=0}^{n-1} \left(\frac{k}{n}\right)^j \right) \left| \frac{z^j}{j!} \right| \leq \frac{R^j}{j!}, \tag{2.30}$$

where the term in brackets is the mean of $\frac{k}{n}$ and each $\frac{k}{n} < 1$. Hence, thanks to Weierstrass' M-test, F_n is absolutely convergent on $\{z \in \mathbb{C} : |z| < R\}$ with the

estimate,

$$\left|\sum_{k=0}^{n-1}\left(\frac{k}{n}\right)^j\frac{z^j}{j!}\right| \leq \sum_{k=0}^{n-1}\frac{R^j}{j!} \leq \exp(R) < \infty. \tag{2.31}$$

Analogously, each term of $F_n(X) = \sum_{j=0}^{\infty} u_j(X)$ is analytic on the disc $\{X \in M_n(\mathbb{C}) | \|X\| < R\}$ because the operator norm is submultiplicative, and hence $\|X^j\| \leq \|X\|^j \leq R^j$. This makes $F_n(X)$ absolutely convergent on $\{X \in M_n(\mathbb{C}) | \|X\| < R\}$ with the estimate,

$$\left\|\sum_{k=0}^{n-1}\left(\frac{k}{n}\right)^j\frac{X^j}{j!}\right\| \leq \sum_{k=0}^{n-1}\frac{R^j}{j!} \leq \exp(R) < \infty. \tag{2.32}$$

As F_n is absolutely convergent on the disc for each n, the sequence $(F_n)_{n\in\mathbb{N}}$ converges to an analytic function denoted F. An analytic function can be represented by its Maclaurin series, and Weierstrass' double series theorem says that one can calculate the derivatives of a function $\sum_{j=0}^{\infty} u_j(z)$ term by term provided each term is analytic on the same disc, as just established.

$$F_n(X) = \sum_{l=0}^{\infty}\frac{F^{(l)}(0)}{l!}X^l,$$

$$F^{(1)}(0) = \frac{1}{n}\sum_{k=0}^{n-1}\left(\frac{k}{n}\right), \qquad F^{(l)}(0) = \frac{1}{n}\frac{l!}{l!}\sum_{k=0}^{n-1}\left(\frac{k}{n}\right)^l.$$

Finally, let $n \to \infty$ and notice the resemblance between $F^{(l)}(0)$ and a Riemann integral, as $\left(\frac{k}{n}\right)_{k=0}^{k=n-1}$ are the left limits of the nth partition of the interval $[0, 1]$.

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\left(\frac{k}{n}\right)^l = \int_0^1 x^l dx = \frac{1}{l+1}. \tag{2.33}$$

Substituting into the Maclaurin series for F_n and taking the limit as $n \to \infty$,

$$F_n(X) \to \sum_{l=0}^{\infty}\frac{z^l}{(l+1)!} = \frac{\exp(X)-1}{X} \tag{2.34}$$

where the last equals sign gives a closed form expression for the power series. □

Theorem 2.3.18 (Tuynman [72]). *For matrices $X, Y \in \mathfrak{g}$ the directional derivative,*

$$\frac{d}{dt} e^{X+tY}\Big|_{t=0} = e^X \left(\frac{I - e^{-ad_X}}{ad_X} \right) (Y). \tag{2.35}$$

Where the expression within the brackets is a formal power series.

Proof. Let us use the expression given in Equation 2.25 to express the derivative of $\exp(\frac{X}{n} + t\frac{Y}{n})^n$. The product rule implies,

$$\begin{aligned}
\frac{d}{dt} \exp(\frac{X}{n} + t\frac{Y}{n})^n &= \sum_{j=0}^{n-1} \exp\left(\frac{X}{n}\right)^{n-1-j} \left(\frac{d}{dt} \exp(\frac{X}{n} + t\frac{Y}{n}) \right) \Big|_{t=0} \exp\left(\frac{X}{n}\right)^{j}, \\
&= \exp(X) \sum_{j=0}^{n-1} \exp\left(\frac{X}{n}\right)^{-j} (\mathcal{C}_{X/n}) \left(\frac{Y}{n}\right) \exp\left(\frac{X}{n}\right)^{j}, \\
&= \exp(X) \sum_{j=0}^{n-1} Ad^j_{\exp(-X/n)} \left((\mathcal{C}_{X/n}) \left(\frac{Y}{n}\right), \right)
\end{aligned}$$

where $\mathcal{C}_X(Y) = e^{-X} \frac{d}{dt} \exp(X + tY)\Big|_{t=0}$ and is a derivative, therefore linear in Y.

$$\begin{aligned}
\frac{d}{dt} \exp(\frac{X}{n} + t\frac{Y}{n})^n &= \frac{e^X}{n} \sum_{j=0}^{n-1} Ad^j_{\exp(-X/n)} \left((\mathcal{C}_{X/n}) (Y) \right), \\
&= e^X \left(\frac{I - Ad_{\exp(-X))}}{n(I - Ad_{\exp(-X/n))}} \right) \left((\mathcal{C}_{X/n}) (Y) \right),
\end{aligned}$$

the expression within the brackets is an operator on $\mathfrak{so}(3)$, the geometric sum of the series Ad up to the n^{th} term. The relationship $Ad_{A^n}(B) = Ad^n_A(B)$ is also exploited. The limit of this expression as $n \to \infty$,

$$\left(\frac{I - Ad_{\exp(-X))}}{n(I - Ad_{\exp(-X/n))}} \right) \longrightarrow \frac{\exp(ad_{-X}) - I}{ad_{-X}} = \frac{I - \exp(-ad_X)}{ad_X}.$$

The identification of linear automorphisms given in Equation (2.22) relates the adjoint map (Ad) to the adjoint representation (ad). Consider the complexification V of $\mathfrak{so}(3)$, and the space of bounded linear operators on V, the operator ad_{-X} can be extended to a linear operator on V by expressing X with respect to a basis of complex matrices.

Lemma 2.3.17 applies to bounded linear operators on $M_n(\mathbb{C})$ in the same way as it does for complex matrices. Finally, the limiting argument is applied to the original approximation for the exponential,

$$\begin{aligned}\lim_{n\to\infty}\frac{d}{dt}\exp(\frac{X}{n}+t\frac{Y}{n})^n &= \exp(X)\lim_{n\to\infty}\left(\frac{I-Ad_{\exp(-X))}}{n(I-Ad_{\exp(-X/n)})}\right)\left(\mathcal{C}_{\frac{X}{n}}(Y)\right),\\ &= \exp(X)\left(\frac{I-\exp(-ad_X)}{ad_X}\right)(Y).\end{aligned}$$

□

The phrase 'formal power series' is used in the theorem to refer to the linear operator $\left(\frac{I-\exp(-ad_X)}{ad_X}\right)$. This is because the operator is used as a succinct representation of the power series

$$\frac{I-\exp(-ad_X)}{ad_X}=\sum_{l=0}^{\infty}\frac{ad_{-X}^l}{(l+1)!}=\sum_{l=0}^{\infty}\frac{(-1)^l\,ad_X^l}{(l+1)!}=\mathcal{D}_X(Y). \tag{2.36}$$

Corollary 2.3.19. *In addition, as outlined by Hall [33, Thm 5.4], an application of the chain rule on* exp *and* $X(t)$*,*

$$\frac{d}{dt}\exp(X(t))|_{t=0}=\exp(X(t))|_{t=0}\left(\frac{I-\exp(-ad_X)}{ad_X}\right)(\dot{X}(t)|_{t=0}). \tag{2.37}$$

Or as a power series rather than a ambiguous operator,

$$\frac{d}{dt}\exp(X(t))=\exp(X(t))\mathcal{D}_{X(t)}(\dot{X}(t)). \tag{2.38}$$

Lemma 2.3.20. *An alternate form for Equation* (2.38) *is,*

$$\frac{d}{dt}\exp(X(t))=\mathcal{D}_{-X(t)}(\dot{X}(t))\exp(X(t)). \tag{2.39}$$

Proof. Take the derivative of $I = \exp(-X(t))\exp(X(t))$ to give,

$$\begin{aligned} 0 &= \frac{d}{dt}\exp(-X(t))\exp(X(t)) + \exp(-X(t))\frac{d}{dt}\exp(X(t)) \\ 0 &= \exp(-X(t))\mathcal{D}_{-X(t)}(-\dot{X}(t))\exp(X(t)) + \exp(-X(t))\frac{d}{dt}\exp(X(t)) \\ \frac{d}{dt}\exp(X(t)) &= \mathcal{D}_{-X(t)}(\dot{X}(t))\exp(X(t)), \end{aligned}$$

where the linearity of $\mathcal{D}_X(Y)$ in Y is used. If this product rule based proof does not convince, then consider the proof of Theorem 2.3.18, the terms of the product rule expansion in the proof can be chosen alternatively as,

$$\begin{aligned} \frac{d}{dt}\exp(\frac{X}{n} + t\frac{Y}{n})^n &= \sum_{j=0}^{n-1} \exp\left(\frac{X}{n}\right)^j \left(\frac{d}{dt}\exp(\frac{X}{n} + t\frac{Y}{n})\right)\Big|_{t=0} \exp\left(\frac{X}{n}\right)^{n-1-j}, \\ &= \sum_{j=0}^{n-1} \exp\left(\frac{X}{n}\right)^j (\mathcal{C}_{X/n})\left(\frac{Y}{n}\right)\exp\left(\frac{X}{n}\right)^{-j}\exp(X), \\ &= \sum_{j=0}^{n-1} Ad^j_{\exp(X/n)}\left((\mathcal{C}_{X/n})\left(\frac{Y}{n}\right)\right)\exp(X). \end{aligned}$$

Where $\mathcal{C}_X(Y) = \left(\frac{d}{dt}\exp(X+tY)\right)\Big|_{t=0}\exp(-X)$ is an alternate valid definition, which is again linear in Y. Following the proof of the theorem through with this equation and the operator ad_X, Equation (2.39) will be recovered. □

Definition 2.3.21. Derived from the reciprocal of the exponential power series, for $|x| > 0$, the series,

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} B_k \frac{x^n}{n!}, \tag{2.40}$$

where B_k represents the k^{th} Bernoulli number.

Using this series and the power series of the exponential we obtain a relation for the

coefficients c_n defined below,

$$\begin{aligned} 1 &= \frac{e^x - 1}{x} \frac{x}{e^x - 1}, \\ &= \left(\sum_{n=0}^{\infty} \frac{x^n}{(n+1)!} \right) \left(\sum_{n=0}^{\infty} B_n \frac{x^n}{n!} \right), \\ &= \sum_{n=0}^{\infty} c_n x^n, \end{aligned}$$

where c_n is given by

$$\begin{aligned} c_n &= \sum_{l,k \,:\, n=l+k} \frac{x^l}{(l+1)!} \frac{B_k}{k!}, \\ &= \sum_{l=0}^{n} \frac{B_{n-l}}{(n-l)!} \frac{1}{(l+1)!}, \\ &= \frac{1}{(n+1)!} \sum_{l=0}^{n} \binom{n+1}{n-l} B_{n-l}. \end{aligned} \tag{2.41}$$

This calculation implies that $c_n = 0$ for all $n \geq 1$.

Lemma 2.3.22. *The inverse mapping to $\mathcal{D}_X$ is given by*

$$\mathcal{E}_X \coloneqq \sum_{n=0}^{\infty} (-1)^n \frac{B_n}{n!} ad_X^n. \tag{2.42}$$

Proof. Recall that ad is linear, for all $Y \in \mathfrak{g}$,

$$\begin{aligned} \mathcal{D}_X \left(\mathcal{E}_X(Y) \right) &= \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!} \mathrm{ad}_X^k (\mathcal{D}_X(Y)), \\ &= \sum_{k=0}^{\infty} \frac{1}{(k+1)!} \mathrm{ad}_{-X}^k \left(\sum_{j=0}^{\infty} (-1)^j \frac{B_j}{j!} \mathrm{ad}_X^j(Y) \right), \\ &= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \frac{1}{(k+1)!} \frac{B_j}{j!} \mathrm{ad}_{-X}^{j+k}(Y), \\ &= \sum_{n=0}^{\infty} c_n \mathrm{ad}_{-X}^n(Y) = Y, \end{aligned}$$

where c_n is defined as in Equation (2.41). As $c_n = 0$ for all $n \geq 1$, the sum reduces to the first term. The linearity of the adjoint operator implies the converse is also true. $\square$

2.3.3 Magnus Expansion

Let us return to the question of existence of solutions to matrix valued differential equations on $SO(n)$ from the beginning of this section,

$$\frac{dX}{dt} = A(t)X(t), \quad X(0) = X_0.$$

For solutions over a bounded interval $t \in [0, t]$ there are classical theorems for the existence and uniqueness of solutions which require the matrix $A(t)$ to be Lipschitz, See Theorem 2.3.3.

Under the assumption that there exists a unique solution X to the initial value problem the question becomes where can that solution be expressed as $X(t) = \exp(\Omega(t))X_0$ for some unique $\Omega(t)$. This is similar to the idea of a monodromy factor, and Birkhoff factorisation[61, §8.2] gives a condition under which a complex analogue of Equation (2.8) can be reduced to $dX = t^{-1}RX$. Consider the complex problem

$$\frac{dv}{dz} = A(z)v(z),$$

for $v : \mathbb{C} \to \mathbb{C}^n$, and matrix valued function $A \in M_n(\mathbb{C})$ which is holomorphic in a neighbourhood around the origin in $\mathbb{C} \setminus \{0\}$ and has a simple pole there too. Provided $A(0)$ has no pair of eigenvalues such that $\lambda_1 - \lambda_2 = 2\pi i\mathbb{Z}$ the differential equation can be reduced to

$$\frac{dv}{dz} = \frac{R}{z}v(z),$$

where R is a constant matrix equal to the residue of $A(z)$ at the simple pole 0 [61, §8.2]. This problem has unique solution up to a monodromy factor, and this solution can be described by an exponential function. The reparameterisation $z = e^t$ coupled with the

chain rule implies,

$$\begin{aligned}\frac{d}{dz}v(z) &= \frac{1}{e^t}\frac{d}{dt}v(e^t),\\ \frac{d}{dt}v(e^t) &= e^t\frac{R}{e^t}v(e^t),\\ v(e^t) &= e^{Rt}v_0.\end{aligned}$$

If the matrix $M = \exp(2\pi iR)$ then the solution $v(e^{t+2\pi i}) = v(e^t)M$ when travelling once around the origin, and M is the monodromy factor. Equally, the function can then be expressed as $v(z) = z^R z_0$ with $t = \log(z)$. The monodromy factor arises here due to function $\log : \mathbb{C} \to \mathbb{C}$ not being injective.

When one returns to the differential equation at the start of the section, Equation (2.8), if the matrix $A(t)$ is Lipschitz with $A(0)$ having no pair of eigenvalues such that $\lambda_1 - \lambda_2 = 2\pi i\mathbb{Z}$, then there exists a local solution to Equation (2.8) on some interval $t \in [0, \tau]$. This solution can be extended to $t \in \mathbb{R}$ thanks to a monodromy factor.

Proposition 2.3.23 (Magnus)**.** *[49, Thm 5] Consider the initial value problem given in Equation* (2.8)*, where the matrix $A(t)$ is Lipschitz with $A(0)$ having no pair of eigenvalues such that $\lambda_1 - \lambda_2 = 2\pi i\mathbb{Z}$. The solution $X(t)$ is equal to $\exp(\Omega(t))X_0$ on the bounded interval $t \in [0, \tau]$, and $\Omega(t)$ satisfies the ODE,*

$$\dot{\Omega}(t) = \mathcal{E}_{-\Omega(t)}(A(t)), \tag{2.43}$$

where the inverse of the exponential is given by the power series,

$$\mathcal{E}_{-\Omega(t)}(A(t)) = \sum_{n=0}^{\infty} \frac{B_k}{n!} ad^k_{\Omega(t)}(A(t)). \tag{2.44}$$

To see the relationship between the differential equations, note that if $\Omega(t)$ is a solution to Equation (2.43) then by applying the operator $\mathcal{D}_{-\Omega(t)}$ to both sides,

$$\mathcal{D}_{-\Omega(t)}(\dot{\Omega}(t)) = A(t). \tag{2.45}$$

The derivative of the exponential map is given in lemma 2.3.20,

$$\dot{X}(t) = \mathcal{D}_{-\Omega(t)}(\dot{\Omega}(t)) \exp(\Omega(t)) X_0 \tag{2.46}$$

$$= \mathcal{D}_{-\Omega(t)}(\dot{\Omega}(t)) X(t), \tag{2.47}$$

substitution of Equation (2.45) into this expression leaves us with Equation (2.8)

$$= A(t)X(t). \tag{2.48}$$

The direct analogy of the monodromy factor in this case results from the fact the exponential map is surjective by Lemma 2.3.14, but it is only injective on a neighbourhood of the origin. For example on $SO(3)$, $\exp(t\Omega) = \exp((2\pi\alpha + t)\Omega)$ for $\Omega \in \mathfrak{so}(3)$ where α is a constant that depends on Ω and can be seen from Rodriguez' formula in the next section.

2.4 Rodriguez' rotation formula

2.4.1 Rotation along geodesic through vector

By employing the relationship between the Lie algebra $\mathfrak{so}(3)$ and the cross product, as well as Rodriguez' rotation formula we show how to specify the element of $SO(3)$ which corresponds to the rotation along the velocity vector V of a distance $h\|V\|$.

Definition 2.4.1. Take a vector $a \in \mathbb{R}^3$ and specify it by $a = (a_1, a_2, a_3)$. Let us denote by M_a the element of $\mathfrak{so}(3)$ which acts by left multiplication such that, for all $x \in \mathbb{R}^3$

$$M_a x = a \times x.$$

This matrix is specified by,

$$M_a = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix},$$

and defines an isomorphism between the two spaces.

Proposition 2.4.2 (Rodriguez' formula)**.** *The exponential of an element of* $\mathfrak{so}(3)$ *can be given in closed form by*

$$\exp(hM_a) = I + \frac{\sin(h\|a\|)}{\|a\|}M_a + \frac{1-\cos(h\|a\|)}{\|a\|^2}M_a^2. \tag{2.49}$$

Proof. This follows from the power series expression for the exponential map,

$$\exp(hM_a) = \sum_{n=0}^{\infty} \frac{(hM_a)^n}{n!}.$$

The matrix M_a multiplies with itself such that $M_a^3 = -\|a\|^2 M_a$, and thus the power series reduces to

$$\begin{aligned}\exp(hM_a) = I &+ \left(h - \frac{h^3}{3!}\|a\|^2 + \frac{h^5}{5!}\|a\|^4 - ...\right) M_a \\ &+ \left(\frac{h^2}{2!} - \frac{h^4}{4!}\|a\|^2 + \frac{h^6}{6!}\|a\|^4 + ...\right) M_a^2.\end{aligned}$$

The resultant power series are of $\sin(h\|a\|)$ and $\frac{1-\cos(h\|a\|)}{\|a\|^2}$ respectively. □

Proposition 2.4.3. *Let* x *denote a unit vector in* $\mathbb{R}^3$ *orthogonal to the aforementioned vector* a*. The rotation matrix* $\exp(hM_a)$ *applied to* x *causes a rotation about the axis* a *of an angle* $h\|a\|$*.*

Proof. Due to the orthogonality of x and a, $a \times (a \times x) = -\|a\|^2 x$.

$$\begin{aligned}\exp(hM_a)x &= x + \frac{\sin(h\|a\|)}{\|a\|}M_a x + \frac{1-\cos(h\|a\|)}{\|a\|^2}M_a^2 x \\ &= x + \frac{\sin(h\|a\|)}{\|a\|}a \times x + \frac{1-\cos(h\|a\|)}{\|a\|^2}a \times (a \times x) \\ &= \frac{\sin(h\|a\|)}{\|a\|}a \times x + \cos(h\|a\|)x.\end{aligned}$$

Observe that $\exp(hM_a)x$ is in the plane perpendicular to a as it lies in the plane spanned by x and $a \times x$. □

Proposition 2.4.4. *If the geodesic curve* $\gamma(t)$ *passes through the point* $X \in \mathbb{S}^2$ *(* $\gamma(0) =$

X) with a velocity $V \in \mathbb{R}^3$ ($\dot{\gamma}(0) = V$) then the point $\gamma(h)$ is given by

$$\gamma(h) = \exp(hM_{X\times V})X. \tag{2.50}$$

Proof. Expand out the exponential and compare with the geodesic map given by Absil [1]. As X is unit length and orthogonal to V, $\|X \times V\| = \|V\|$,

$$\exp(hM_{X\times V})X = \frac{\sin(h\|V\|)}{\|V\|}(X \times V) \times X + \cos(h\|V\|)X. \tag{2.51}$$

Then by properties of the cross product,

$$\begin{aligned}(X \times V) \times X &= \langle X, X\rangle V - \langle V, X\rangle X, \\ &= \|X\|^2 V.\end{aligned}$$

□

Chapter 3

Relevant measure theory

Ideas from measure theory in the context of optimal transport are important to understand later chapters, therefore this chapter outlines these concepts. This includes establishing notation. The reason for the numerous definitions is that generally optimal transport deals with probability measures with finite variance. However, for the fluid problem tackled in this thesis it is not physically realistic for mass to pile onto sets of measure zero. Therefore an additional constraint on the sets of measures under discussion must be added. Furthermore, in order to establish the existence and uniqueness of a minimising measure within a prescribed set of measures, the set is required to be compact. Compact subsets of the sets of measures are thus defined and, in the third section of the chapter, it is proven that they are compact with respect to the weak topology specified. The fundamentals explained in this section are developed in more detail in books on measure theory [41] [3].

3.1 Spaces of measures

To define a measure space, first a σ-algebra is required. As the scope of this work requires working with Borel σ-algebras, those will be the focus. To this end, the base space Ω will always be a complete and separable metric space, and for most intents and purposes, either $\mathbb{R}^n$ or $\mathbb{S}^2$.

Definition 3.1.1 (σ-algebra)**.** Consider a total space Ω, a σ-algebra on Ω is a collection of subsets of Ω, denoted $\mathcal{F}$, satisfying:

(i) $\Omega \in \mathcal{F}$.

(ii) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

(iii) $\mathcal{F}$ is closed under countable intersections (and unions by de Morgan's law [41, p.1]).

Definition 3.1.2 (Borel σ-algebra)**.** Consider a Euclidean space $\Omega = \mathbb{R}^n$ or a compact subset. The topology of Ω is defined by the Euclidean metric, and the smallest σ-algebra that contains all of the open sets on Ω is called the *Borel σ-algebra* and denoted $\mathcal{B}(M)$.

Definition 3.1.3. A measure is a set function $\mu : \mathcal{F} \to \mathbb{R}_+ \cap \{\infty\}$, which is countably additive on any collection of disjoint elements of the σ-algebra and assigns zero value to the empty set. The triplet $(M, \mathcal{F}, \mu)$ is a measure space.

Definition 3.1.4. The set of all bounded measures on Ω is denoted $M_b(\Omega)$.

Definition 3.1.5. Consider two measure spaces, $(M_1, \mathcal{F}_1, \rho_1)$ and $(M_2, \mathcal{F}_2, \rho_2)$. A $\mathcal{F}_1/\mathcal{F}_2$-measurable function is a function $f : M_1 \to M_2$ such that the preimage of any measurable set is measurable, $f^{-1}(F_2) = F_1$ where $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$.

If both σ-algebras are Borel then any continuous function $f : M_1 \to M_2$ is measurable (though the class of measurable functions includes non-continuous ones too). If $M_2 = \mathbb{R}^n$ then it is common to omit this term from the -measurable nomenclature.

From the idea of a measurable function, a general definition of integration with respect to a measure can be developed [3]. Consider the measure space $(\Omega, \mathcal{F}_1, \rho)$, for any $\mathcal{F}_1/\mathcal{B}(\mathbb{R})$-measurable, bounded function $f : \Omega \to \mathbb{R}$. The integral

$$\int_A f(x)\rho(dx), \quad \forall A \in \mathcal{F}_1$$

is a well defined object taking a value between $(-\infty, \infty)$. Details of the construction are common in books on measure theory [3], as is the construction of Lebesgue measure, μ. Lebesgue measure makes $(\Omega, \mathcal{B}(\Omega), \mu)$ a measure space, when Ω is $\mathbb{R}^n$ or a compact subset.

A probability measure is a measure which assigns the measure 1 to the total space Ω, and non-negative values to every element of $\mathcal{F}$. The set of all Borel probability measures on Ω is denoted $\mathcal{M}(\Omega)$.

Definition 3.1.6. The set $\mathcal{M}_2(\mathbb{R}^n)$ is the set of Borel probability measures on $\mathbb{R}^n$ which have finite variance. In which case $\rho \in \mathcal{M}_2(\mathbb{R}^n)$ if and only if

$$\int_{\mathbb{R}^n} |x|^2 d\rho(x) < \infty. \tag{3.1}$$

Definition 3.1.7. The set $\mathcal{P}(\mathbb{R}^n)$ denotes the set of probability measures which are absolutely continuous with respect to Lebesgue measure. Thus, $\mathcal{P}_2(\mathbb{R}^n)$ denotes the measures belonging to $\mathcal{P}(\mathbb{R}^n)$ with finite variance.

In the case of $\rho \in \mathcal{P}(\mathbb{R}^n)$, the measure ρ has a probability density, which is a positive function in L^1 which we can also denote ρ without ambiguity as

$$\int_{\mathbb{R}^n} \rho(dx) = \int_{\mathbb{R}^n} \rho(x)\mu(dx),$$

in which the former use of ρ is as a measure, and the latter as a density. μ is Lebesgue measure unless defined otherwise.

Definition 3.1.8. A subset of $\mathcal{M}_2(\mathbb{R}^n)$ in which each probability measure has a smaller variance than some $K > 0$ is denoted,

$$\mathcal{M}_{2,K}(\mathbb{R}^n) = \{\rho \in \mathcal{M}_2(\mathbb{R}^n) \mid \int_{\mathbb{R}^n} |x|^2 \rho(x)dx < K\}. \tag{3.2}$$

Definition 3.1.9. A subset of $\mathcal{P}_2(\mathbb{R}^n)$ in which each probability measure has a smaller variance than some $K > 0$ is denoted,

$$\mathcal{P}_{2,K}(\mathbb{R}^n) = \{\rho \in \mathcal{P}_2(\mathbb{R}^n) \mid \int_{\mathbb{R}^n} |x|^2 \rho(x)dx < K\}. \tag{3.3}$$

Definition 3.1.10. For $\gamma \in (1, 2]$ the set $\mathcal{P}^\gamma(\mathbb{R}^n)$ is defined as the set of probabilities in $\mathcal{P}(\mathbb{R}^n)$ which also satisfy

$$\int_{\mathbb{R}^n} \rho^\gamma(x)dx < \infty. \tag{3.4}$$

In other words they have finite L^γ norm, where L^p spaces will be defined in the following section. Analogously to $\mathcal{P}_{2,K}(\mathbb{R}^n)$, the subset $\mathcal{P}^{\gamma,L}(\mathbb{R}^n)$ is defined

$$\mathcal{P}^{\gamma,L}(\mathbb{R}^n) = \{\rho \in \mathcal{P}^\gamma(\mathbb{R}^n) \mid \int_{\mathbb{R}^n} \rho^\gamma(x)dx < L\}. \tag{3.5}$$

Definition 3.1.11. The set $\mathcal{P}^{\gamma,L}_{2,K}(\mathbb{R}^n)$ is defined as the intersection $\mathcal{P}_{2,K}(\mathbb{R}^n) \cap \mathcal{P}^{\gamma,L}(\mathbb{R}^n)$.

3.2 Function spaces

First, general function spaces used within the thesis are defined, then special attention is given to the spaces H^1 and $L^2(\mathbb{T})$, which have particular relevance to Chapter 8 on constructing the Gibbs measure.

Definition 3.2.1. The $L^p(\mathbb{R}^n, \mu)$ spaces are Banach spaces for $p \in [1, \infty)$, and also measure spaces with the Borel σ-algebra. For a function $f : \mathbb{R}^n \to \mathbb{R}$ to belong to $L^p(\mathbb{R}^n, \mu)$,

$$\|f\|^p_{L^p} := \left(\int_{\mathbb{R}^n} |f|^p \mu(dx) \right)^{\frac{1}{p}} < \infty.$$

Note that this definition easily extends to $L^p(\mathbb{R}^n, \rho)$ where $\rho \in \mathcal{P}_2(\mathbb{R}^n)$ by replacing Lebesgue measure with the new measure ρ: $\rho(dx) = \rho(x)\mu(dx)$.

It is well known that the L^2 spaces are Hilbert spaces, the principle example discussed in this work is $L^2(\mathbb{T})$ — the space of functions $f : [0, 2\pi) \to \mathbb{R}$ which have finite L^p-norm. In addition, the L^p spaces can be combined very naturally, so a 'vector' function $f : \mathbb{R}^n \to \mathbb{R}^n$ can lie in the function space $L^p(\mathbb{R}^n; \mathbb{R}^n, \rho)$ if each component f_i lies in $L^p(\mathbb{R}^n, \rho)$, and the norm on $L^p(\mathbb{R}^n; \mathbb{R}^n, \rho)$ is defined

$$\|f\|^p_{L^p} := \left(\int_{\mathbb{R}^n} \|f\|^p_p \mu(dx) \right)^{\frac{1}{p}}, \tag{3.6}$$

where $\|f\|_p = (\sum_i^n f_i^p)^{\frac{1}{p}}$ denotes the p-norm on $\mathbb{R}^n$.

Definition 3.2.2. The set of bounded continuous functions $f : M \to \mathbb{R}$ from a Riemannian manifold (usually $\mathbb{R}^n$) is denoted $C_b(M)$.

Definition 3.2.3. The set of smooth (infinitely differentiable) continuous functions $f : M \to \mathbb{R}$ is defined $C_\infty(\mathbb{R})$.

3.2.1 L^2 and Fourier series

The Gibbs measure will be constructed using primarily the space $L^2(\mathbb{T}, \mathbb{C}) = L^2(\mathbb{T}) \times L^2(\mathbb{T})$, this section first defines $L^2(\mathbb{T}, \mathbb{C})$ and why Fourier series are useful on this space.

The Sobolev space H^1 and its dual are then defined, followed by a discussion of the way each of these spaces can be embedded continuously within each other. Solutions to the NLS exist in H^1 but the Gibbs measure is defined on a larger space.

Definition 3.2.4. $L^2(\mathbb{T}, \mathbb{C})$ will denote the Hilbert space of square, Lebesgue integrable functions $f : \mathbb{T} \to \mathbb{C}$, with finite L^2 norm,

$$\|f\|_{L^2}^2 = \frac{1}{2\pi} \int_0^{2\pi} |f|^2 dx < \infty.$$

This norm is defined from the inner product on $L^2(\mathbb{T}, \mathbb{C})$,

$$\langle f(x), g(x) \rangle_{L^2} = \frac{1}{2\pi} \int_0^{2\pi} f(x)\overline{g(x)} dx.$$

It is worth noting the distinction, $L^2(\mathbb{T}, \mathbb{C})$ is the set of complex valued functions with finite L^2 norm. The real functions $L^2(\mathbb{T}) \subset L^2(\mathbb{T}, \mathbb{C})$ and the constructions in this section are done on $L^2(\mathbb{T}, \mathbb{C})$ and the structure is inherited by the subset.

Lemma 3.2.5. *The sequence* $(\exp(inx))_{n\in\mathbb{N}}$ *is orthonormal in* $L^2(\mathbb{T}, \mathbb{C})$.

Proof. Consider two integers $n, m \in \mathbb{N}$,

$$\langle e^{inx}, e^{imx} \rangle_{L^2} = \frac{1}{2\pi} \int_0^{2\pi} e^{i(n-m)x} dx.$$

If $n = m$ then the integrand is equal to 1, implying each vector $\exp(inx)$ is unit length. If $n \neq m$ then $\exp(i(n-m)x$ is a holomorphic function for $x \in \mathbb{R}$. Furthermore, the function is periodic and traces out a closed loop in the complex plane as x ranges between $[0, 2\pi]$, as for any $n - m \in \mathbb{N}$, $\exp(i(n-m)2\pi) = 1$. As a result, the integral

$$\oint_0^{2\pi} e^{i(n-m)x} dx = 0,$$

and the sequence is orthogonal. □

Let ℓ^2 refer to the space of square summable sequences — $(a_n)_{n\in\mathbb{N}} \in \ell^2$ if and only if $\sum_0^\infty |a_n|^2 < \infty$. To establish a representation of a function $f \in L^2(\mathbb{T}, \mathbb{C})$ as

components with respect to the orthonormal functions $(\exp(inx))_{n\in\mathbb{N}}$ introduce the truncation function,

$$S_N f(x) = \sum_{n=-N}^{N} a_n e^{inx}.$$

Then one can employ the Riesz-Fischer theorem.

Theorem 3.2.1 (Riesz-Fischer)**.** *The function f lies in $L^2(\mathbb{T},\mathbb{C})$ if and only if there exists $(a_n)_{n\in\mathbb{N}} \in \ell^2$ such that $S_N f$ converges to f in L^2-norm as $N \to \infty$.*

Note that a_n is the Fourier transform of f, $\hat{f}(n) = a_n$,

$$a_n = \int_0^{2\pi} f(x) e^{-inx} dx.$$

The Reisz-Fischer theorem can be used to prove Parseval's identity,

$$\frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2 ds = \sum_{n=-\infty}^{\infty} |a_n|^2. \tag{3.7}$$

Two copies of $L^2(\mathbb{T})$ mean a function on $L^2(\mathbb{T},\mathbb{C}) = L^2(\mathbb{T}) \times L^2(\mathbb{T})$ can be represented by $\psi = P + iQ = \sum_{n=-\infty}^{\infty}(a_n + ib_n)e^{inx}$ for $P, Q \in L^2(\mathbb{T})$. More importantly, this then allows $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ to be real, which is not the case for a general function $f \in L^2(\mathbb{T})$ — it may have complex Fourier coefficients.

3.2.2 The Sobolev space H^1

Definition 3.2.6. [26] Recall $\mathbb{T} = \mathbb{R} \setminus 2\pi\mathbb{Z}$, meaning a complex funtion $g : \mathbb{T} \to \mathbb{C}$ may be identified with a 2π-periodic function from $\mathbb{R}$. Let $f \in L^1(\mathbb{T},\mathbb{C})$; then f is said to have weak derivative $h \in L^2(\mathbb{T},\mathbb{C})$ if

$$\int_{\mathbb{T}} f(x)\overline{g'(x)}dx + \int_{\mathbb{T}} h(x)\overline{g(x)} = 0 \tag{3.8}$$

holds for all $g \in C^\infty(\mathbb{T},\mathbb{C})$. Thus the function h behaves as the derivative of f would on any Lebesgue measurable set, when integrated against a test function.

For any function in $L^2(\mathbb{T},\mathbb{C}) \cap L^1(\mathbb{T},\mathbb{C})$ its weak derivative is defined accordingly.

Also note the definition of a weak derivative implies a way to describe weak solutions to certain simple differential equations.

Definition 3.2.7. H^1 will denote the Sobolev space of functions $f \in L^2(\mathbb{T}, \mathbb{C})$ which have weak derivatives, and the weak derivatives are also in $L^2(\mathbb{T}, \mathbb{C})$. As such, the norm is defined:

$$\|f\|_{H^1}^2 = \|f\|_{L^2}^2 + \|f'\|_{L^2}^2. \tag{3.9}$$

So H^1 consists of functions with finite H^1 norm.

Lemma 3.2.8. *One can define the space H^1 in terms of Fourier series by,*

$$H^1 = \{f(x) = \sum_{n=-\infty}^{\infty} a_n e^{inx} \in L^2(\mathbb{T}, \mathbb{C}) : \sum_{n=-\infty}^{\infty} (1+n^2)|a_n|^2 < \infty\}.$$

Proof. Denote $f \in L^2(\mathbb{T}, \mathbb{C}) \cap L^1(\mathbb{T}, \mathbb{C})$ by Fourier series, $f = \sum_{n=-\infty}^{\infty} a_n e^{inx}$ and under the assumption that its weak derivative is also in L^2, let $h = \sum_{n=-\infty}^{\infty} c_n e^{inx}$. The test functions $g \in C^\infty(\mathbb{T}, \mathbb{C})$ are in $L^2(\mathbb{T}, \mathbb{C})$ as $\mathbb{T}$ is compact, and thus can be expressed by Fourier series as well, $g = \sum_{n=-\infty}^{\infty} b_n e^{inx}$. The derivative of g, $g' = \sum_{n=-\infty}^{\infty} inb_n e^{inx}$ and its complex conjugate is $\overline{g'} = \sum_{n=-\infty}^{\infty} in\overline{b_n} e^{inx}$. Thus if $f \in H^1$ and has weak derivative $f' = h$ then by Equation (3.8),

$$\sum_{n=-\infty}^{\infty} c_n \overline{b_n} - \sum_{n=-\infty}^{\infty} ina_n \overline{b_n} = 0.$$

As this holds for any choice of g, the equation implies that $c_n = ina_n$. So the weak derivative of f is given by $h = \sum_{n=-\infty}^{\infty} ina_n e^{inx}$. Then $h \in L^2(\mathbb{T}, \mathbb{C})$ if and only if $\sum_{n=-\infty}^{\infty} n^2|a_n|^2$ converges. Hence, $f \in L^2 \cap L^1$ has weak derivative in L^2 if and only if the Fourier coefficients of f have

$$\sum_{n=-\infty}^{\infty} (1+n^2)|a_n|^2 < \infty,$$

and this sum is equivalent to the norm on H^1, $\|f\|_{H^1}^2$. □

H^{-1} will denote the continuous dual space of H^1, i.e. Continuous linear functionals $F : H^1 \to \mathbb{C}$. Riesz's representation theorem on $L^2(\mathbb{T}, \mathbb{C})$ says that for each linear

functional $F : L^2(\mathbb{T}, \mathbb{C}) \to \mathbb{C}$, there exists $g \in L^2(\mathbb{T}, \mathbb{C})$ such that $F(f) = \langle f, g \rangle$ for all $f \in L^2(\mathbb{T}, \mathbb{C})$. Treating H^1 as a subset of $L^2(\mathbb{T}, \mathbb{C})$, the same conditions hold for $F : H^1 \to \mathbb{C}$. However, for $f \in H^1$ the inner product $\langle f, g \rangle$ is finite for any function with (possibly divergent) Fourier series $g = \sum_{n=-\infty}^{\infty} b_n e^{inx}$ such that $\sum_{n=-\infty}^{\infty} \frac{b_n^2}{n^2+1} < \infty$. Each of these functions g has an associated element of H^{-1} therefore H^{-1} can be represented by the set

$$H^{-1} = \{g(x) = \sum_{n=-\infty}^{\infty} b_n e^{inx} : \sum_{n=-\infty}^{\infty} \frac{1}{n^2+1} |b_n|^2 < \infty\}. \tag{3.10}$$

Thus H^{-1} will have the dual norm defined,

$$\|F\|_{H^{-1}} = \sup_{g \in H^1} \{ |\langle g, f \rangle| \ : \ \|g\|_{H^1} \leq 1 \}. \tag{3.11}$$

3.2.3 A chain of continuous embeddings

Definition 3.2.9. A continuous embedding between two normed vector spaces, $X \hookrightarrow Y$, exists iff the inclusion map $i : X \to Y; x \mapsto x$ is continuous, that is if $\exists C > 0$ such that $\|x\|_Y \leq C\|x\|_X$.

Lemma 3.2.10. $H^1 \hookrightarrow C([0, 2\pi])$.

Proof. Take the Fourier series representation $P(s) = \sum_{n=-\infty}^{\infty} a_n e^{ins} \in H^1$, the zero-th Fourier term is finite because the definition of H^1 implies $|a_0|^2 < \infty$. For the non-zero Fourier terms, P is absolutely convergent by Cauchy-Schwartz:

$$\begin{aligned} \sum_{n=-\infty}^{\infty} |a_n| &= \left\langle \sum_{n=-\infty}^{\infty} \frac{e^{in\theta}}{\sqrt{n^2+1}}, \sum_{n=-\infty}^{\infty} \sqrt{n^2+1}\, |a_n| e^{in\theta} \right\rangle_{L^2} \\ &\leq \left(\sum_{n=-\infty}^{\infty} \frac{1}{n^2+1} \right)^{\frac{1}{2}} \left(\sum_{n=-\infty}^{\infty} (n^2+1)|a_n|^2 \right)^{\frac{1}{2}} < \infty. \end{aligned} \tag{3.12}$$

Then note that $\|P\|_\infty = \sup_s(|P(s)|) \leq \sum_{n=-\infty}^{\infty} |a_n|$, and so equation (3.12) implies that

$$\|P\|_\infty \leq \left(\sum_{n=-\infty}^{\infty} \frac{1}{n^2+1} \right)^{\frac{1}{2}} \|P\|_{H^1}. \tag{3.13}$$

□

Lemma 3.2.11. $C([0,2\pi]) \hookrightarrow L^2$

Proof. For any $f \in C$,

$$\|f\|_{L^2}^2 = \int_0^{2\pi} |f|^2 dx \leq \left(\sup_{x\in[0,2\pi]} |f(x)| \right)^2 \int_0^{2\pi} dx = 2\pi \|f\|_\infty^2.$$

□

Lemma 3.2.12. $H^1 \hookrightarrow L^2$

Proof. For any $P \in H^1$ with Fourier series $P(x) = \sum_{n=-\infty}^{\infty} a_n e^{inx}$,

$$\|P\|_{L^2}^2 = \sum_{n=-\infty}^{\infty} |a_n|^2 \leq \sum_{n=-\infty}^{\infty} (1+n^2)|a_n|^2 = \|P\|_{H^1}^2$$

□

Lemma 3.2.13. $C \hookrightarrow H^{-1}$

Proof. Let $F \in H^{-1}$, and recall the definition of the norm of H^{-1}, and the function f both given in Equation (3.10). With use of Cauchy-Schwartz, and both Lemma 3.2.11 and Lemma 3.2.12,

$$\|F\|_{H^{-1}} \leq |\langle g, f\rangle| \leq \|g\|_{L^2}\|f\|_{L^2} \leq C\|g\|_{H^1}\|f\|_\infty. \tag{3.14}$$

Thus $\|F\|_{H^{-1}} \leq C\|f\|_\infty$. □

Hence, By Lemma 3.2.10 and Lemma 3.2.13 there are continuous linear inclusions $H^1 \hookrightarrow C \hookrightarrow H^{-1}$.

3.2.4 Clarkson's inequality

An inequality used in later sections known as Clarkson's inequality is proven in this section.

Definition 3.2.14 (Holder Conjugates)**.** The pair (p, p') of real numbers are said to be *Holder conjugates* of each other if

$$\frac{1}{p} + \frac{1}{p'} = 1.$$

In the case of $0 < p < 1$, p' will be negative. In light of this, note that the space $L^{p'}(\mathbb{R}^n, \mu)$, while no longer a Banach space, can be defined as the set of Lebesgue measurable functions, f, which do not equal zero except on sets of measure zero. As such,

$$\|f\|_{p'} := \left(\int_{\mathbb{R}^n} \frac{1}{|f|^{\frac{p}{1-p}}} \mu(dx) \right)^{\frac{p-1}{p}} < \infty,$$

though $\|f\|_{p'}$ is symbolic, rather than a norm.

Lemma 3.2.15. *If p denotes the index of an L^p space, and $0 < p < 1$, then Holder's inequality is reversed.*

Proof. This proof is based on the proof of Equation 2.8.4 of Ref. [34]. If p is defined as above, let $\ell = \frac{1}{p}$ then $1 < \ell < \infty$, and the Holder inequality says

$$\sum u_j v_j \leq \left(\sum u_j^{\ell} \right)^{\frac{1}{\ell}} \left(\sum v_j^{\ell'} \right)^{\frac{1}{\ell'}}. \tag{3.15}$$

If $u_j = (a_j b_j)^p$ and $v_j = b_j^{-p}$, then Holder's equation is now

$$\begin{aligned}
\sum a_j^p &\leq \left(\sum a_j b_j \right)^{\frac{1}{\ell}} \left(\sum b_j^{p'} \right)^{\frac{1}{\ell'}}, \\
\sum a_j^p &\leq \left(\sum a_j b_j \right)^{p} \left(\sum b_j^{p'} \right)^{1-p}, \\
\left(\sum b_j^{p'} \right)^{p-1} \sum a_j^p &\leq \left(\sum a_j b_j \right)^{p}, \\
\left(\sum a_j^p \right)^{\frac{1}{p}} \left(\sum b_j^{p'} \right)^{\frac{1}{p'}} &\leq \sum a_j b_j.
\end{aligned}$$

Proving that for $0 < p < 1$ we get a reverse Holder's inequality. □

Lemma 3.2.16. *For L^p spaces of index $0 < p < 1$, the triangle inequality is reversed.*

Proof. Let $s_j = a_j + b_j$, and consider the sum,

$$\begin{aligned}
\sum q_j s_j^p &= \sum q_j a_j s_j^{p-1} + \sum q_j b_j s_j^{p-1} \\
&= \sum q_j^{\frac{1}{p}} a_j (q_j^{\frac{1}{p}} s_j)^{p-1} + \sum q_j^{\frac{1}{p}} b_j (q_j^{\frac{1}{p}} s_j)^{p-1}
\end{aligned}$$

Apply the reverse Holder's inequality to each term,

$$\begin{aligned}
\sum q_j s_j^p &\geq \left(\sum q_j a_j^p\right)^{\frac{1}{p}} \left(\sum (q_j^{\frac{1}{p}} s_j)^{(p-1)p'}\right)^{\frac{1}{p'}} + \left(\sum q_j b_j^p\right)^{\frac{1}{p}} \left(\sum (q_j^{\frac{1}{p}} s_j)^{(p-1)p'}\right)^{\frac{1}{p'}} \\
&= \left[\left(\sum q_j a_j^p\right)^{\frac{1}{p}} + \left(\sum q_j b_j^p\right)^{\frac{1}{p}}\right] \left(\sum q_j s_j^p\right)^{\frac{1}{p'}}, \\
\left(\sum q_j s_j^p\right)^{1-\frac{1}{p'}} &\geq \left(\sum q_j a_j^p\right)^{\frac{1}{p}} + \left(\sum q_j b_j^p\right)^{\frac{1}{p}}.
\end{aligned}$$

And the identity $\frac{1}{p} = 1 - \frac{1}{p'}$ finishes the proof. □

Lemma 3.2.17 (Clarkson's inequality)**.** *For* $1 < \gamma < 2$*,*

$$\left\|\frac{f+g}{2}\right\|_\gamma^{\gamma'} + \left\|\frac{f-g}{2}\right\|_\gamma^{\gamma'} \leq \left(\frac{1}{2}\|f\|_\gamma^\gamma + \|g\|_\gamma^\gamma\right)^{\frac{\gamma'}{\gamma}}. \tag{3.16}$$

Proof. First note a property of the L^γ and $L^{\gamma-1}$ norms for Holder conjugate indices (γ, γ'):

$$\begin{aligned}
\|f\|_\gamma^{\gamma'} &= \left(\int |f|^\gamma\right)^{\frac{\gamma'}{\gamma}} \\
&= \left(\int |f|^{\gamma'(\gamma-1)}\right)^{\frac{1}{\gamma-1}} \\
&= \||f|^{\gamma'}\|_{\gamma-1}.
\end{aligned}$$

Starting with the left hand side of Clarkson's inequality, apply this property and then, as $\gamma - 1 < 1$, the reverse triangle inequality can be applied.

$$\begin{aligned}
\left\|\frac{f+g}{2}\right\|_\gamma^{\gamma'} + \left\|\frac{f-g}{2}\right\|_\gamma^{\gamma'} &= \left\|\left|\frac{f+g}{2}\right|^{\gamma'}\right\|_{\gamma-1} + \left\|\left|\frac{f-g}{2}\right|^{\gamma'}\right\|_{\gamma-1} \\
&\leq \left\|\left|\frac{f+g}{2}\right|^{\gamma'} + \left|\frac{f-g}{2}\right|^{\gamma'}\right\|_{\gamma-1}.
\end{aligned} \tag{3.17}$$

The following step relies on a pointwise inequality for a convex function. The function

$x \mapsto |x|^\gamma$ is convex for $1 < \gamma < 2$, and thus

$$\left|\frac{f+g}{2}\right|^{\gamma'} = \left|\frac{f+g}{2}\right|^{\gamma(\gamma'-1)} \leq \left(\frac{1}{2}\left|\frac{f}{2}\right|^{\gamma} + \frac{1}{2}\left|\frac{g}{2}\right|^{\gamma}\right)^{\gamma'-1}.$$

Therefore,

$$\begin{aligned}\left|\frac{f+g}{2}\right|^{\gamma'} + \left|\frac{f-g}{2}\right|^{\gamma'} &\leq 2\left(\frac{1}{2}\left|\frac{f}{2}\right|^{\gamma} + \frac{1}{2}\left|\frac{g}{2}\right|^{\gamma}\right)^{\gamma'-1} \\ &= \frac{1}{2^{\gamma'-1}}\left(\frac{1}{2}|f|^{\gamma} + \frac{1}{2}|g|^{\gamma}\right)^{\gamma'-1} \\ &\leq \left(\frac{1}{2}|f|^{\gamma} + \frac{1}{2}|g|^{\gamma}\right)^{\gamma'-1}.\end{aligned}$$

Apply this pointwise inequality to Equation (3.17), and manipulate the norm relation discussed at the start of the proof using $(\gamma-1)(\gamma'-1) = 1$ to get Clarkson's inequality.

$$\left\|\left|\frac{f+g}{2}\right|^{\gamma'} + \left|\frac{f-g}{2}\right|^{\gamma'}\right\|_{\gamma-1} \leq \left\|\left(\frac{1}{2}|f|^{\gamma} + \frac{1}{2}|g|^{\gamma}\right)^{\gamma'-1}\right\|_{\gamma-1} \tag{3.18}$$

$$= \left(\int \left(\frac{1}{2}|f|^{\gamma} + \frac{1}{2}|g|^{\gamma}\right)^{1} dx\right)^{\frac{1}{\gamma-1}} \tag{3.19}$$

$$= \left(\frac{1}{2}\|f\|_{\gamma}^{\gamma} + \frac{1}{2}\|g\|_{\gamma}^{\gamma}\right)^{\gamma'-1}. \tag{3.20}$$

□

3.3 Weak compactness

Further into the thesis the compactness of certain spaces of probability measures will be required to achieve convergence of our numerical methods. Ideally we would simply show the compactness of $\mathcal{P}_2(\mathbb{R}^n)$ and $\mathcal{P}_2(\mathbb{S}^2)$, but these spaces are not themselves compact without extra conditions.

The definition of a Borel probability measure implies that any bounded continuous function can be integrated with respect to this probability measure and the integral is bounded. This fact, along with Riesz' representation theorem on L^2, suggests there

could be a relationship between the dual space of $C_b(\Omega)$ and $\mathcal{M}(\Omega)$.

Definition 3.3.1. The dual space of $C_b(\Omega)$ is the space of Bounded Linear operators, $F : C_b(\Omega) \to \mathbb{R}$. It is equipped with the operator norm,

$$\|F\|_{op} = \sup\{|F(f)| \ : \|f\|_\infty \leq 1\}, \tag{3.21}$$

where $\|\cdot\|_\infty$ is the supremum norm.

Theorem 3.3.1 (Riesz' representation of a measure)**.** *[7, Thm 1.1.3] Consider the dual of $C_b(\Omega)$ for some compact metric space Ω, that is, the set of linear functionals $G : C_b(\Omega) \to \mathbb{R}$. Then every $G \in C_b(\Omega)^*$ has a corresponding unique real valued measure, ν on Ω, such that there exists an isomorphism between the spaces,*

$$G(f) = \int_\Omega f(x)\nu(dx).$$

Weak convergence of a sequence of measures in $M_b(\Omega)$ is defined by this correspondence, $\nu_n \to \nu$ as $n \to \infty$ *weakly* if,

$$\int f \ \nu_n(dx) \to \int f \ \nu(dx), \qquad \forall f \in C_b(\Omega).$$

The weak topology on $\mathcal{M}(\Omega)$ is induced by the map $\nu \mapsto \int_\Omega f(x)\nu(dx)$. It can be generated by a *sub basic* collection of open sets defined, for $0 < b$ as $C_{f,b} = \{\nu \in \mathcal{M}(\Omega) \ : \int_\Omega f(x)\nu(dx) < b\}$ then the collection $\{C_{f,b} \ : b \in \mathbb{R}^+, f \in C_b(\Omega)\}$ generates the topology through finite intersections and arbitrary unions. The standard topology on $\mathbb{R}$ has a base formed by the open intervals (with rational endpoints). Analogously, the base for the weak topology on $\mathcal{M}(\Omega)$ is the collection of sets, for $a, b, \in (0, 1]$ defined by $\{\nu \in \mathcal{M}(\Omega) \ : a < \int_\Omega f_j(x)\nu(dx) < b, \ j = 1, \ldots, n\}$. As a base, any open set in the topology is a union of (possibly infinite) sets of this form.

Lemma 3.3.2. *If Ω is compact, then $\mathcal{M}(\Omega)$ is compact with respect to the weak topology.*

Proof. Let B denote the closed unit ball in $C_b(\Omega)$, $B = \{f \in C_b(\Omega) : \|f\|_\infty \leq 1\}$, and $[-1, 1]^B$ denote an uncountably infinite product space. Consider the map $\mathcal{M}(\Omega) \to [-1, 1]^B; \nu \mapsto \left(\int_\Omega f(x)\nu(dx)\right)_{f\in B}$, and the pre-image under this map of the

space $[-1,1]^B$. The space $\mathcal{M}(\Omega)$ is compact with respect to its weak topology by the compactness of $[-1,1]^B$, as demonstrated by Tychonoff [67]. This construction is due to Blower [7]. □

As a result of this Lemma, the set $\mathcal{M}(\mathbb{S}^2)$ is compact. For spaces such as $\mathbb{R}^n$ which are only locally compact, the concept of tightness is needed to define a metric with respect to which sets can be considered compact.

Definition 3.3.3 (Tightness)**.** A collection of probability measures A is *tight* in Ω if for all $\epsilon > 0$ there exists a compact subset $K_\epsilon \subset \Omega$ such that for all $\rho \in A$,

$$\rho(\Omega \setminus K_\epsilon) < \epsilon. \tag{3.22}$$

Lemma 3.3.4. *The collection $\mathcal{M}_{2,L}(\mathbb{R}^n)$ is tight for $L > 0$.*

Proof. By the assumptions of $\mathcal{M}_{2,L}(\mathbb{R}^n)$ given in Definition 3.1.8, the second moment of all ρ are bounded. If this bound is $L > 0$ then for all $\varepsilon > 0$ define $K_\varepsilon = \{x \in \mathbb{R}^n : \|x\|^2 \leq \kappa(\varepsilon)\}$. Then the estimate

$$L > \int_{\mathbb{R}^n} \|x\|^2 \rho(dx) > \int_{\mathbb{R}^n \setminus K_\epsilon} \|x\|^2 \rho(dx) > \kappa(\varepsilon)\rho(\mathbb{R}^n \setminus K_\varepsilon).$$

Hence, if $\kappa(\varepsilon) = L/\varepsilon$ then $\rho(\mathbb{R}^n \setminus K_\epsilon) < \epsilon$ and the collection is tight. □

Theorem 3.3.2 (Prokhorov)**.** *[3, Thm. 5.1] On a separable metric space Ω, a collection $A \subset \mathcal{M}(\Omega)$ is tight if and only if it is sequentially compact with respect to the weak topology on $\mathcal{M}(\Omega)$.*

By Prokhorov's theorem and Lemma 3.3.4 it follows that $\mathcal{M}_{2,L}(\mathbb{R}^n)$ is sequentially compact with respect to the weak topology. Thus, having established that $\mathcal{M}_2(\mathbb{R}^n)$ and $\mathcal{M}_2(\mathbb{S}^2)$ are sequentially compact, the question remains as to what extra conditions are required for $\mathcal{P}_2(\mathbb{R}^n)$ and $\mathcal{P}_2(\mathbb{S}^2)$ to be compact as well. Under the weak topology, $\mathcal{P}_2(\mathbb{R}^n)$ is not compact due to the existence of Dirac delta measures. These measures are not themselves in $\mathcal{P}_2(\mathbb{R}^n)$, but are limits of sequences of measures in $\mathcal{P}_2(\mathbb{R}^n)$.

Let $ca(\mathbb{R})$ denote the set of countably additive probability measures defined on the Borel sets of $\mathbb{R}$.

Theorem 3.3.5 (Sequential precompactness)**.** *[22, p. IV.9.2] A subset $P \subset ca(\mathbb{R})$ is weakly sequentially precompact if and only if it is bounded, tight, and there exists some $\mu \in ca(\mathbb{R})$ such that*

$$\lim_{\mu(E_n)\to 0} \rho(E_n) = 0, \tag{3.23}$$

and this limit is uniform with respect to ρ in P.

Definition 3.3.6. Let $\mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n)$ denote the set of measures $\rho \in \mathcal{P}_{2,L}(\mathbb{R}^n) \cap \mathcal{P}^{\gamma,K}(\mathbb{R}^n)$, in other words

(i) $\int_{\mathbb{R}^n} \rho^\gamma dx \leq K$,

(ii) $\int_{\mathbb{R}^n} \|x\|^2 \rho(x) dx \leq L$.

Some properties of $\mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n)$. For any $\rho \in \mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n)$, and any $M > 0$ the estimate,

$$M^2 \int_{\|x\|>M} \rho(x) dx \leq \int_{\|x\|>M} \|x\|^2 \rho(x) dx \leq L, \tag{3.24}$$

shows that the measure of the set $\{\|x\| \leq M\}$ is at least L/M^2.

Secondly, we have the estimate, for any $R > 0$

$$R^{\gamma-1} \int_{\rho(x)>R} \rho(x) dx \leq \int_{\rho(x)>R} \rho(x)^\gamma dx < K, \tag{3.25}$$

which implies that $\rho(x)$ is bounded above almost everywhere. If coupled with the absolute continuity of ρ with respect to Lebesgue measure, this fact asserts ρ is bounded.

Lemma 3.3.7. *The set $\mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n)$ is sequentially precompact.*

Proof. The set $\mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n) \subset \mathcal{M}_{2,L}(\mathbb{R}^n)$ and is therefore tight by the argument of Lemma 3.3.4. On sets of non-zero Lebesgue measure, elements of $\mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n)$ are bounded as demonstrated by Equation (3.25). Sets of zero Lebesgue measure are evaluated using the condition in Theorem 3.3.5. Let μ denote Lebesgue measure, $\mu \in ca(\mathbb{R})$. Then for

any $\rho \in \mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n)$, any set $E \in \mathcal{B}(\mathbb{R}^n)$, and any $R > 0$,

$$\int_E \rho(x)dx = \int_{E\cap\{\rho(x)\leq L\}} \rho(x)dx + \int_{E\cap\{\rho(x)>R\}} \rho(x)dx, \tag{3.26}$$

$$\leq R\int_E dx + \frac{K}{R^{\gamma-1}}, \tag{3.27}$$

which follows from the fact that $R^{\gamma-1}\rho(x) \leq \rho(x)^\gamma$ on the set $\{\rho(x) > R\}$ and Equation (3.25). The inequality holds for any $R > 0$ thus, taking $R = \sqrt{\mu(E)}$ will bound the integral of ρ by a function of $\mu(E)$. Therefore, if $\mu(E) \to 0$, then $\int_E \rho(x)dx \to 0$ for all ρ uniformly, and thus the conditions for Theorem 3.3.5 hold, and so $\mathcal{P}_{2,L}^{\gamma,K}(\mathbb{R}^n)$ is sequentially precompact.

□

3.4 Weak solutions to PDEs

Having introduced the concept of a weak derivative and discussed dual spaces to spaces of measures, the tools are available now to introduce the general idea of a weak solution to a differential equation, additional detail on this topic is given in Evans [26]. Let L denote an operator and $Lu = 0$ represent a partial differential equation searching for a solution $u : \Omega \to \mathbb{C}$ where Ω is compact. Then the function v is a weak solution to the PDE if

$$\int_\Omega (Lv)\overline{\phi}dx = 0, \qquad \forall \phi \in C^\infty(\Omega, \mathbb{C}).$$

This integral is not necessarily well defined, as a solution v may not be regular enough to have a second derivative for example. However, as with the definition of a weak derivative, compactness of Ω implies that any f is zero on the boundary and integration by parts can be used to define the integral. Furthermore, if L is simple enough to express the integral in the form

$$\int_\Omega (Lv)\overline{\phi}dx = -\int_\Omega v\overline{L_2\phi}dx,$$

for example if $v \in L^2(\Omega, \mathbb{C})$ then L_2 would be the adjoint of L. Then one can see how the integral condition could be well defined on a space of v such that the integral is finite. In addition, the integral could be satified by a measure with density $v(x)$ and

so the concepts of weak solution and measure valued solution to a PDE are in this sense the same. In general, a weak solution means a solution that is less regular than the PDE requires — for example a C^1 function could be a weak solution to a second order PDE. But the integral formulation allows for the potential of discontinuous and nondifferentiable solutions and that is the type of weak solution considered in this piece of work.

Chapter 4

Optimal transport

In this chapter, the topic of optimal transport will be introduced. In the subsequent chapter it will be applied to find weak solutions to PDEs, but this chapter focusses on the theory. Here, the setting of optimal transport is the space $\mathcal{M}_2(\mathbb{R}^n)$, and its subspace of absolutely continuous measures $\mathcal{P}_2(\mathbb{R}^n)$.

To establish the context for later chapters, the PDE under consideration will be Euler's equations of fluid motion. The absolutely continuous measures are the class of distributions inside which solutions to the PDE will be sought, where their densities will represent the distribution of mass of the fluid. The Wasserstein distance will represent a metric for the kinetic energy needed to move between states. The Wasserstein distance and optimal transport give structure to the space.

The basic problem of optimal transport is the Kantorovich mass transportation problem. Optimal transport maps are discussed, and the conditions under which they are monotone and invertible. The most important result stated in this chapter is Brenier's theorem, which establishes the existence of a convex function whose gradient is the optimal map between two measures in $\mathcal{P}_2(\mathbb{R}^n)$. The chapter ends with a discussion of the basics of convex functions, which will be extended in the next chapter to convexity on spaces of probability measures.

4.1 Wasserstein distance

The natural metric for $\mathcal{M}_2(\mathbb{R}^n)$ is the quadratic Wasserstein distance, this is a standard example of a cost function used to search for an optimal plan to transport between two measures. This section covers the basic concepts of optimal transport theory. The 'principle of least action' guides us to minimise the change in energy of the system, and for that we need a measure of the distance between states.

The broadest problem in optimal transportation is the following.

Definition 4.1.1 (The Kantorovich mass transportation problem)**.** Let (X, μ) and (Y, ν) denote probability spaces, and $c(x, y)$ a cost function between them. The set of probability measures on the product set $X \times Y$ with marginals $\mu \in \mathcal{M}_2(X)$ and $\nu \in \mathcal{M}_2(Y)$ is denoted $\Pi(\mu, \nu)$ and the Kantorovich mass transportation problem is to find the infimum,

$$\inf \left\{ \int_{X\times Y} c(x, y) d\pi(x, y) \ : \ \pi \in \Pi(\mu, \nu) \right\}. \tag{4.1}$$

The product measure, π which minimises this integral is known as the transport plan.

The set $\Pi(\mu, \nu)$ is defined equivalently as the set of probability measures π on $X \times Y$ such that for all test functions $(\varphi, \phi) \in L^1(d\mu) \times L^1(d\nu)$,

$$\int_{X\times Y} \varphi(x) + \phi(y) \, d\pi(x, y) = \int_X \varphi(x) d\mu(x) + \int_Y \phi(y) d\nu(y).$$

If X, Y are locally compact Polish spaces (such as $\mathbb{R}^n$), then the space of test functions $L^1(d\mu) \times L^1(d\nu)$ can be reduced to $C_b(X) \times C_b(Y)$, bounded functions on the original spaces.

Definition 4.1.2. The pushforward of a measure is a measurable map ψ between two measures $\rho_1, \rho_2 \in \mathcal{M}_2(\mathbb{R}^n)$ often denoted $\rho_2 = \psi \# \rho_1$. The map ψ must satisfy

$$\int_{\mathbb{R}^n} g(y)\rho_2(y)dy = \int_{\mathbb{R}^n} g(\psi(x))\rho_1(x)dx, \tag{4.2}$$

for any measurable map $g \in L^1(\mathbb{R}, \rho_2)$.

This definition invites a slight reformulation of Kantorovich's problem which is known as Monge's formulation.

Definition 4.1.3. Let (X, μ) and (Y, ν) denote probability spaces, and $c(x, y)$ a cost function between them. Let $T : X \to Y$ denote any transport map which pushes forward μ to ν, then the Monge problem is to find the transport map which realises the infimum,

$$\inf_{T\#\mu=\nu} \left\{ \int_X c(x, T(x)) d\mu(x) \right\}. \tag{4.3}$$

Monge's problem is related to the Kantorovich problem, let $S(x) = (x, T(x))$ and then Kantorovich's joint measure π is defined by the pushforward $\pi = S\#\mu$. This formulation has a limitation — in the general setting in which Kantorovich's problem has a minimiser, Monge's problem may not.

Definition 4.1.4 (Wasserstein distance)**.** The Kantorovich problem on a compact and seperable metric space, in which the cost function is the metric on that space is known as the Wasserstein distance between the two measures considered.

In this work the Wasserstein distance $W_q(\mu, \nu)$ is defined on $\mathcal{M}_2(\mathbb{R}^n)$ with the metric given by the q-norm $\|x\|_q = (\sum_{i=0}^n |x_i|^q)^{1/q}$,

$$W_q(\mu, \nu)^q = \inf \left\{ \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_q^q \, d\pi(x, y) \mid \pi \in \Pi(\mu, \nu) \right\}. \tag{4.4}$$

The distance W_2 is the metric usually considered, thanks to the conclusions of Brenier's theorem in the next section. The Euclidean norm $\|x\|_2$ has the subscript surpressed due to its frequency of use.

Proposition 4.1.5. *Let $\mu, \nu \in \mathcal{M}_2(\mathbb{R}^n)$, there exists a transport plan $\pi \in \Pi(\mu, \nu)$ which minimises the Wasserstein distance $W_2(\mu, \nu)$.*

Proof. As discussed in Villani [74, p.51], let M_μ denote the bound on the variance of the measure $\mu \in \mathcal{M}_2(\mathbb{R}^n)$, and likewise define M_ν for $\mu \in \mathcal{M}_2(\mathbb{R}^n)$. By the triangle inequality,

$$\begin{aligned} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 d\pi(x, y) &\leq \int_{\mathbb{R}^n} \|x\|^2 d\pi(x, y) + \int_{\mathbb{R}^n} \|y\|^2 d\pi(x, y), \\ &= \int_{\mathbb{R}^n} \|x\|^2 d\mu + \int_{\mathbb{R}^n} \|y\|^2 d\nu, \\ &\leq M_\mu + M_\nu. \end{aligned}$$

The properties of $\Pi(\mu,\nu)$ impose the condition that the marginals of π are μ and ν and hence the above calculation. Then the measure μ satisfies the conditions of tightness in $\mathcal{M}_2(\mathbb{R}^n)$ on the set $K_\epsilon^\mu := \{x \in \mathbb{R}^n \mid \|x\|^2 \leq \frac{M_\mu}{\epsilon}\}$ as shown in the proof of Lemma 3.3.4. Define K_ϵ^ν in the same way and then,

$$\begin{aligned}\pi[(\mathbb{R}^n \times \mathbb{R}^n) \setminus (K_\epsilon^\mu \times K_\epsilon^\nu)] &\leq \pi[(\mathbb{R}^n \setminus K_\epsilon^\mu) \times \mathbb{R}^n] + \pi[\mathbb{R}^n \times (\mathbb{R}^n \setminus K_\epsilon^\nu)], \\ &\leq \mu[\mathbb{R}^n \setminus K_\epsilon^\mu] + \nu[\mathbb{R}^n \setminus K_\epsilon^\nu] \leq 2\epsilon.\end{aligned}$$

Hence the set $\Pi(\mu,\nu)$ is tight, and by Prokhorov's theorem (Theorem 3.3.2) it is compact. □

Remark 4.1.6. In one dimension, any Borel measurable probability distribution can be defined according to its cumulative distibution function, often denoted by a capital letter. The Wasserstein distance can be expressed in terms of the probability measure's cumulative distribution function, and if capital letters are used as arguments of a Wasserstien distance, $W_q(F,G)$, then the two measures under consideration are defined by their cumulative distribution functions F and G. Further details are discussed in the Kantorovich-Rubenstein theorem [21, Thm. 11.8.2]

4.1.1 The dual problem

The Kantorovich problem can be reformulated into a dual problem.

Definition 4.1.7 (Kantorovich Duality)**.** [74, p.19] As in the Kantorovich problem, let (X,μ) and (Y,ν) denote probability spaces, and $c(x,y)$ a cost function between them. Introduce the set Φ_c as the set of all measurable pairs $(\varphi,\phi) \in L^1(d\mu) \times L^1(d\nu)$ such that $\varphi(x) + \phi(y) \leq c(x,y)$ for μ-almost all x and ν-almost all y. Then

$$\inf_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{X\times Y} c(x,y) d\pi(x,y) \right\} = \sup_{(\varphi,\phi)\in\Phi_c} \left\{ \int_X \varphi(x) d\mu(x) + \int_Y \phi(y) d\nu(y) \right\}. \quad (4.5)$$

This introduces the dual problem. In the specific case in which we are working, the cost function is quadratic and the pairs (φ,ϕ) turn out to be convex conjugates of each other in the sense of the Legendre dual.

Definition 4.1.8. The Legendre dual of a function $\varphi(x)$ is the convex function denoted

φ^* which is defined as

$$\varphi^*(y) := \sup_x \left(\langle x, y \rangle - \varphi(x) \right), \tag{4.6}$$

and will be referred to as the convex conjugate of φ.

Lemma 4.1.9. *For the Kantorovich dual problem with quadratic cost ($c(x,y) = \|x - y\|^2$), Equation (4.5) can be expressed alternatively by*

$$\sup_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{X \times Y} \langle x, y \rangle d\pi(x,y) \right\} = \inf_{(\vartheta,\theta) \in \Theta_c} \left\{ \int_X \vartheta(x) d\mu(x) + \int_Y \theta(y) d\nu(y) \right\}, \tag{4.7}$$

where Θ_c is defined as the set of all measurable pairs $(\vartheta, \theta) \in L^1(d\mu) \times L^1(d\nu)$ such that $\vartheta(x) + \theta(y) \geq \langle x, y \rangle$ for μ-almost all x and ν-almost all y.

Proof. Considering the standard Kantorovich problem for quadratic cost, the set Φ_c includes all measurable pairs $(\varphi, \phi) \in L^1(d\mu) \times L^1(d\nu)$ such that $\varphi(x) + \phi(y) \leq \|x - y\|^2$ for μ-almost all x and ν-almost all y. This condition is developed by expanding out the norm,

$$2\langle x, y \rangle \leq \left(\|x\|^2 - \varphi(x) \right) + \left(\|y\|^2 - \phi(y) \right).$$

Then defining $\vartheta(x) = (\|x\|^2 - \varphi(x))/2$ and likewise $\theta(y) = (\|y\|^2 - \phi(y))/2$, and for any choice of $(\varphi, \phi) \in L^1(d\mu) \times L^1(d\nu)$, the pair (ϑ, θ) belong to $L^1(d\mu) \times L^1(d\nu)$ too because $\mu, \nu \in \mathcal{M}_2(\mathbb{R}^n)$ have finite variance. Explicitly, there exists a $K \in \mathbb{R}$ such that

$$\int_X \|x\|^2 d\mu(x) + \int_Y \|y\|^2 d\nu(y) = K.$$

The right hand side of Equation (4.5) is expanded in this context to

$$\begin{aligned} \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{X \times Y} \|x - y\|^2 d\pi(x,y) \right\} &= \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \begin{array}{c} \int_X \|x\|^2 d\mu(x) + \int_Y \|y\|^2 d\nu(y) \\ -2 \int_{X \times Y} \langle x, y \rangle d\pi(x,y) \end{array} \right\} \\ &= K - 2 \sup_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{X \times Y} \langle x, y \rangle d\pi(x,y) \right\}. \end{aligned} \tag{4.8}$$

And the left side of Equation (4.5) is also expanded similarly,

$$\sup_{(\varphi,\phi)\in\Phi_c}\left\{\int\varphi(x)d\mu(x)+\int\phi(y)d\nu(y)\right\}= \\ K-2\inf_{(\vartheta,\theta)\in\Theta_c}\left\{\int\vartheta(x)d\mu(x)+\int\theta(y)d\nu(y)\right\}. \tag{4.9}$$

Equating these two expressions provides the desired result. □

Lemma 4.1.10 (Double convexification)**.** *Assume there exists a pair $(\vartheta,\theta)\in\Theta_c$ which realise the infimum in Equation* (4.7)*. Then it can also be realised by the pair of convex conjugate functions $(\vartheta^{**},\vartheta^{*})\in L^1(d\mu)\times L^1(d\nu)$ defined in terms of the original ϑ which also live in Θ_c.*

Proof. All that is required is to establish the inequality,

$$\inf_{(\vartheta^{**},\vartheta^{*})\in\Theta_c}\left\{\int_X\vartheta(x)^{**}d\mu(x)+\int_Y\vartheta^{*}(y)d\nu(y)\right\}\leq\inf_{(\vartheta,\theta)\in\Theta_c}\left\{\int_X\vartheta(x)d\mu(x)+\int_Y\theta(y)d\nu(y)\right\}. \tag{4.10}$$

The definition of Θ_c implies that $\vartheta(x)+\theta(y)\geq\langle x,y\rangle$ on sets of non-negligible measure. The definition of the Legendre dual is for all x [74, Rem. 2.2]. Thus the inequality $\theta(y)\geq\sup_x(\langle x,y\rangle-\vartheta(x))=\vartheta^{*}(y)$ holds for ν-almost all y, implying

$$\int_X\vartheta(x)d\mu(x)+\int_Y\vartheta^{*}(y)d\nu(y)\leq\int_X\vartheta(x)d\mu(x)+\int_Y\theta(y)d\nu(y). \tag{4.11}$$

Now, purely from the definition of the Legendre dual, $\vartheta(x)+\vartheta^{*}(y)\leq\langle x,y\rangle$, and by the taking the dual of the dual, $\vartheta^{**}(x)=\sup_y(\langle x,y\rangle-\vartheta^{*}(y))$. Thus, $\vartheta^{**}(x)\leq\vartheta(x)$ for μ–almost all x, thus Equation (4.11) can be extended,

$$\int_X\vartheta^{**}(x)d\mu(x)+\int_Y\vartheta^{*}(y)d\nu(y)\leq\int_X\vartheta(x)d\mu(x)+\int_Y\theta(y)d\nu(y), \tag{4.12}$$

and Equation (4.10) holds. □

Theorem 4.1.11. *[74, Thm. 2.9] If $\mu,\nu\in\mathcal{M}_2(\mathbb{R}^n)$ then the infimum in Equation* (4.7) *is realised by a pair of conjugate convex functions (ψ^{**},ψ^{*}).*

Theorem 4.1.12 (Brenier's theorem)**.** *[74, Thm 2.12] Let $\rho_1, \rho_2 \in \mathcal{M}_2(\mathbb{R}^n)$, and consider the Monge-Kantorovich problem for a distance between these two measures using the quadratic cost function $\|x-y\|^2_{\mathbb{R}^n}$.*

1. ***Knott-Smith optimality criterion*** *As in the formulation of the dual problem in Definition 4.1.7, $\pi \in \Pi(\rho_1, \rho_2)$ is optimal if and only if there exists ψ a convex function which minimises,*

$$\inf_{(\psi,\psi^*)\in\Theta_c} \left\{ \int_X \psi(x) d\rho_1(x) + \int_Y \psi^*(y) d\rho_2(y) \right\}.$$

   where Θ_c is defined as in Lemma 4.1.9. In addition, ψ must have subdifferential $\partial\psi(x)$ in which y is an element of $\partial\psi(x)$ for each point $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$ of the support of π.

2. ***Brenier's theorem*** *Under the additional assumption that ρ_1 is absolutely continuous with respect to Lebesgue measure, π is the unique optimal probability measure in $\Pi(\rho_1, \rho_2)$ if*

$$\pi = (Id \times \nabla\psi)\#\rho_1, \tag{4.13}$$

   where $\nabla\psi$ is the uniquely determined ρ_1 almost everywhere gradient of a convex function, and $\nabla\psi\#\rho_1 = \rho_2$.

3. *Following the assumptions of part (2), $\nabla\psi$ is the unique solution to the Monge problem.*

4. *If ρ_2 is also absolutely continuous with respect to Lebesgue, then $\nabla\psi$ is invertible almost everywhere.*

4.1.2 Convexity

The idea of convex optimisation is used extensively in many areas of applied mathematics, including machine learning. In general, convexity is an important property when trying to determine the minimisers of a function. Hamilton's principle of least action for example illustrates the importance of extremals when searching for solutions to a dynamic system. Convexity is also important in the present context of transport maps

on spaces of probability functions [2]. Here the basic properties of convex functions are outlined, in a later section convexity in the context of $\mathcal{P}_2(\mathbb{R}^n)$ is discussed too.

Definition 4.1.13. A convex function is a function $f : \mathbb{R} \to \mathbb{R}$ which has a graph that is convex. In other words, for any two points on the graph, a straight line between them will be above the curve. For all $x_1, x_2 \in \mathbb{R}$ and $t \in [0, 1]$,

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2). \tag{4.14}$$

The definition can be extended simply to functions $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 4.1.14. A convex function on $\mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ has equivalent definitions below.

(i) For each $x \in \mathbb{R}^n$ at which $f(x)$ is differentiable, for all $y \in \mathbb{R}^n$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

(ii) For $x, y \in \mathbb{R}$ at which f is differentiable, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$. In other words, if its gradient is monotone then it is convex.

Lemma 4.1.15. *[30, p.4] If $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable, then f is convex iff $f''(x) \geq 0$ for all $x \in \mathbb{R}$.*

Proof. Take the second definition of convexity in one dimension and define a new function, $g(y) \coloneqq f(y) + f'(x)(x - y)$. The function $g(y)$ is convex because a straight line is both convex and concave (just not strictly). Take the derivative of $g(y)$, $g'(y) = f'(y) - f'(x)$. Hence x is an extremal of $g(y)$, this makes it a minimiser because g is convex. A minimiser of a function has positive second derivative, hence $g''(x) \geq 0$, but by the definition of g, $g''(x) = f''(x)$. □

Lemma 4.1.16. *For a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, f is convex if and only if Hess $f \succeq 0$, in other words, Hess f is positive semi-definite.*

The proof follows from Lemma 4.1.15 and can be found in books on convex optimisation [11, Ex 3.8].

4.1.3 Jacobian change of variables

This section follows on from the discussion of convexity and is important to mention for application in later chapters.

Definition 4.1.17 (Monotone function)**.** A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is monotone if and only if, for all $x, y \in \mathbb{R}^n$,

$$\langle f(x) - f(y), x - y \rangle \geq 0. \tag{4.15}$$

It is strictly monotone if also not equal to zero.

The Jacobian change of variables formula can be extended to measures in $\mathcal{P}_2(\mathbb{R}^n)$.

Lemma 4.1.18. *For measures $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R}^n)$, implying they are both absolutely continuous with respect to Lebesgue measure and ψ a pushforward map, $\psi \# \rho_1 = \rho_2$ which is invertible and monotone increasing, the change of variables formula is*

$$\rho_2(\psi(x))|J_\psi(x)| = \rho_1(x). \tag{4.16}$$

where J is the Jacobian of $\psi = (\psi_1, \psi_2, \ldots, \psi_n)$,

$$J_\psi(x) = \begin{pmatrix} \frac{\partial \psi_1}{\partial x_1} & \frac{\partial \psi_1}{\partial x_2} & \cdots \\ \frac{\partial \psi_2}{\partial x_1} & \frac{\partial \psi_2}{\partial x_2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{4.17}$$

and $|\quad|$ denotes the determinant.

In 1D this is,

$$\rho_2(\psi(x)) \left| \frac{d\psi(x)}{dx} \right| = \rho_1(x). \tag{4.18}$$

Proof. The definition of a pushforward measure means, for any $f \in L_1(\mathbb{R}, \rho_1)$

$$\int_{\mathbb{R}^n} f(y)\rho_2(y)dy = \int_{\mathbb{R}^n} f(\psi(x))\rho_1(x)dx. \tag{4.19}$$

Then the Jacobian change of variables formula for the change of variables $y = \psi(x)$ applied to the left hand side

$$\int_{\mathbb{R}^n} f(y)\rho_2(y)dy = \int_{\mathbb{R}^n} f(\psi(x))|J_\psi(x)|\rho_2(\psi(x))dx, \tag{4.20}$$

and therefore, this implies

$$\rho_2(\psi(x))|J_\psi(x)| = \rho_1(x).$$

□

4.2 Otto's interpretation

Otto made the first steps in linking optimal transport to Riemannian geometry in the context of partial differential equations. He was describing weak solutions to the porous medium equation and conceptualised the problem geometrically. In fluid dynamics the system is commonly described by a density $\rho : \mathbb{R}^n \to \mathbb{R}_+$ and an accompanying velocity field $v : \mathbb{R}^n \to \mathbb{R}^n$.

This construction works for any smooth Riemannian manifold but $\mathbb{R}^n$ is the application in mind. A Riemannian metric can be defined in terms of geodesics. Let $\gamma : [0, 1] \to \mathbb{R}^n$ represent a smooth curve on the manifold, then the infimum,

$$d(x, y)^2 = \inf \left\{ \int_0^1 \|\dot{\gamma}(s)\|^2 ds \mid \gamma(0) = x, \gamma(1) = y \right\}, \tag{4.21}$$

defines the Riemannian metric. The Wasserstein metric can be defined on this manifold [57] by

$$W_2(\rho_1, \rho_2)^2 = \inf \left\{ \int_{\mathbb{R}^n \times \mathbb{R}^n} d(x, y)^2 \, d\pi(\rho_1, \rho_2) \mid \pi \in \Pi(\rho_1, \rho_2) \right\}. \tag{4.22}$$

The present context for ρ is as a weak solution to a PDE coupled with a velocity field v which describes the flow of ρ over time, as such they must satisfy the continuity equation

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho v).$$

Geometrically, v should lie in the tangent space to $\mathcal{M}_2(\mathbb{R}^n)$. Thus the tangent space at ρ should include probability densities of the form $-\nabla \cdot (\rho v)$ where $v \in L^2(\mathbb{R}^n, \rho)$, implying the fluid has finite kinetic energy. Otto defines the norm

$$\left\| \frac{\partial \rho}{\partial t} \right\|_\rho^2 = \inf \left\{ \int \|v\|^2 \rho dx \mid \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0, \, v \in L^2(\mathbb{R}^n, \rho) \right\} \tag{4.23}$$

ensuring that the velocity chosen to represent the gradient of the curve satisfies the

continuity equation and minimises the kinetic energy. With a choice of element in the tangent space at each point defined, take a curve in $\mathcal{M}_2(\mathbb{R}^n)$ given by ρ_t that is defined for $1 \leq t \leq 2$. One can return to the definition of the geodesic distance to measure the distance between measures as the Wasserstein distance

$$W_2(\rho_1, \rho_2)^2 = \inf \left\{ \int_1^2 \left\| \frac{\partial \rho_t}{\partial t} \right\|_\rho^2 dt \mid \rho \in \mathcal{M}_2(\mathbb{R}^n) \right\}. \tag{4.24}$$

The Euler equations of fluid dynamics is another system of PDEs where this construction is relevant. These PDEs are dicsussed in Chapter 13 and an adaptation of his approach is used to solve the Euler equations in Chapter 13

Chapter 5

The Lax pair for NLSE via the Hasimoto transform

In this chapter the nonlinear Schrödinger equation (NLSE) is discussed, one of the two integrable systems of interest in this thesis. In the first section the *Hasimoto transform* is used to show the equivalence between periodic solutions to the NLSE, and the curvature and torsion of a smooth curve in $\mathbb{R}^3$. From this relation, a Lax pair is formulated for the dynamics of the Frenet-Serret frame of the previously mentioned curve. The concept of a Lax pair is a powerful tool in recasting a PDE as a coupled set of ODEs along with a consistency condition. Geometrically it can be understood as producing a curvature free connection on the solution manifold for the problem.

5.1 The Hamiltonian system of the NLSE

Definition 5.1.1. The nonlinear Schrödinger equation[26] is the nonlinear extension of Schrödingers well known wave equation. In this work it is defined as,

$$\frac{1}{i}\frac{\partial \psi}{\partial t} = \frac{\partial^2 \psi}{\partial x^2} + \beta|\psi|^2\psi. \tag{5.1}$$

Where $\beta < 0$ and $\beta > 0$ give the focussing and defocussing versions respectively.

Definition 5.1.2. If P, Q are real functions of period 2π on the space $H^1(\mathbb{T}) \times H^1(\mathbb{T})$,

the Hamiltonian

$$H(P,Q) = \frac{\beta}{4}\int_{\mathbb{T}}(P^2+Q^2)^2dx + \frac{1}{2}\int_{\mathbb{T}}(P')^2dx + \frac{1}{2}\int_{\mathbb{T}}(Q')^2dx, \tag{5.2}$$

gives rise to the nonlinear Schrödinger equation for $\psi = P + iQ$, through its canonical equations of motion.

The canonical equations of motion for this Hamiltonian are

$$\begin{aligned} -\frac{\partial P}{\partial t} &= \frac{\partial^2 Q}{\partial x^2} + \beta(P^2+Q^2)Q, \\ \frac{\partial Q}{\partial t} &= \frac{\partial^2 P}{\partial x^2} + \beta(P^2+Q^2)P. \end{aligned}$$

They can be combined into a PDE in terms of ψ,

$$\frac{1}{i}\frac{\partial \psi}{\partial t} = \frac{\partial^2 \psi}{\partial x^2} + \beta|\psi|^2\psi,$$

which is the nonlinear Schrödinger equation as given in Equation (5.1).

5.1.1 Hasimoto's curve

It has been shown by Hasimoto [35] that the curvature and torsion of an isolated thin vortex filament in a ideal fluid can be described by the focussing nonlinear Schrödinger (with $\beta = -1/2$). The relation given by Hasimoto is referred to as the Hasimoto transform.

Definition 5.1.3. Define the Hasimoto transform,

$$\psi(x,t) = \kappa(x,t)\exp\left(i\int_0^x \tau(u,t)\,du\right).$$

Hasimoto states that a solution $\psi \in C^2(\mathbb{T})$ of the NLS is associated with a curve γ and the curvature and torsion of its Frenet-Serret frame $\{\mathbf{t},\mathbf{n},\mathbf{b}\}$ via the Hasimoto transform. This statement is established later in Proposition 5.1.7, but the approach taken by Hasimoto starts with the curve defined below.

Introduce a twice differentiable vector valued function $\gamma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^3$ where $\gamma = \gamma(x,t)$, which satisfies the differential equation:

$$\dot{\gamma} = \kappa \mathbf{b}, \tag{5.3}$$

where a dot denotes $\frac{\partial}{\partial t}$ and a prime will denote $\frac{\partial}{\partial x}$ throughout. γ is defined to be unit speed with respect to the variable x, and as discussed in Section 1.2, the derivative with respect to x can therefore be described by the Frenet-Serret equations,

$$\frac{\partial}{\partial x}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}.$$

The frame $[\mathbf{t}, \mathbf{n}, \mathbf{b}]$ is defined with respect to the curve $\gamma(x,t)$ where the t coordinate has been fixed. The curvature $\kappa(x,t)$ and torsion $\tau(x,t)$ vary smoothly due to the assumptions on γ as shown in Proposition 1.2.4. From Equation (5.3) and the Frenet-Serret frame, the evolution of the frame with respect to time can be expressed in terms of the curvature and torsion.

Proposition 5.1.4. *The time derivatives of the tangent, normal and binormal of γ are,*

$$\frac{\partial}{\partial t}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & -\tau\kappa & \kappa'\\ \tau\kappa & 0 & -\mu\\ -\kappa' & \mu & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}. \tag{5.4}$$

Proof. The second order partial derivatives of a function are equal, hence

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{t} &= \frac{\partial}{\partial x}\kappa\mathbf{b}\\ \dot{\mathbf{t}} &= \kappa'\mathbf{b} + \kappa\mathbf{b}'\\ \dot{\mathbf{t}} &= \kappa'\mathbf{b} - \kappa\tau\mathbf{n}\end{aligned} \tag{5.5}$$

The vectors $\mathbf{t}$, $\mathbf{n}$ and $\mathbf{b}$ are orthonormal and so the vectors $\dot{\mathbf{n}}$ and $\dot{\mathbf{b}}$ can be expressed

as a linear combination of them:

$$\dot{\mathbf{n}} = \alpha\mathbf{t} + \beta\mathbf{n} + \eta\mathbf{b}, \qquad \dot{\mathbf{b}} = \lambda\mathbf{t} + \mu\mathbf{n} + v\mathbf{b}. \tag{5.6}$$

These coefficients will be found below.

Find μ and η The vectors $\mathbf{n}$ and $\mathbf{b}$ are orthogonal, hence $\mathbf{n} \cdot \mathbf{b} = 0$. This means the derivative:

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{n} \cdot \mathbf{b} &= 0, \\ \dot{\mathbf{b}} \cdot \mathbf{n} + \dot{\mathbf{n}} \cdot \mathbf{b} &= 0, \\ \lambda\mathbf{t} \cdot \mathbf{n} + \mu\mathbf{n} \cdot \mathbf{n} + v\mathbf{b} \cdot \mathbf{n} + \alpha\mathbf{t} \cdot \mathbf{b} + \beta\mathbf{n} \cdot \mathbf{b} + \eta\mathbf{b} \cdot \mathbf{b} &= 0, \\ \mu + \eta &= 0. \end{aligned}$$

So the constraints on μ and η is just that $\eta = -\mu$.

Find β and v Both $\mathbf{n}$ and $\mathbf{b}$ are unit length, therefore the same process applies to each. Unit length implies $\mathbf{n} \cdot \mathbf{n} = 1$, hence the derivative:

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{n} \cdot \mathbf{n} &= 0, \\ 2\dot{\mathbf{n}} \cdot \mathbf{n} &= 0, \\ \alpha\mathbf{t} \cdot \mathbf{n} + \beta\mathbf{n} \cdot \mathbf{n} + \eta\mathbf{b} \cdot \mathbf{n} &= 0, \\ \beta &= 0. \end{aligned}$$

The same calculation holds for $\dot{\mathbf{b}}$ and shows that $v = 0$.

Find α Equation (5.5) gives $\dot{\mathbf{t}}$, this can be used with the fact $\mathbf{n}$ and $\mathbf{t}$ are orthogonal, to derive α.

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{n} \cdot \mathbf{t} &= 0, \\ \dot{\mathbf{n}} \cdot \mathbf{t} + \dot{\mathbf{t}} \cdot \mathbf{n} &= 0, \\ \dot{\mathbf{n}} \cdot \mathbf{t} &= -\kappa'\mathbf{b} \cdot \mathbf{n} + \kappa\tau\mathbf{n} \cdot \mathbf{n} \\ \dot{\mathbf{n}} \cdot \mathbf{t} &= \kappa\tau. \end{aligned}$$

Hence $\alpha = \kappa\tau$.

Find λ Analogously to finding α:

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbf{b}\cdot\mathbf{t} &= 0,\\
\dot{\mathbf{b}}\cdot\mathbf{t} + \dot{\mathbf{t}}\cdot\mathbf{b} &= 0,\\
\dot{\mathbf{b}}\cdot\mathbf{t} &= -\kappa'\mathbf{b}\cdot\mathbf{b} + \kappa\tau\mathbf{n}\cdot\mathbf{b}\\
\dot{\mathbf{b}}\cdot\mathbf{t} &= -\kappa'.
\end{aligned}$$

Hence $\lambda = -\kappa'$

□

5.1.2 The Lax pair condition

Denote the matrix derived in the last section and the Frenet-Serret Matrix,

$$\frac{\partial}{\partial x}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}, \qquad \frac{\partial}{\partial t}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & -\tau\kappa & \kappa'\\ \tau\kappa & 0 & -\mu\\ -\kappa' & \mu & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}, \tag{5.7}$$

by $\partial_t X = \Omega_2 X$ and $\partial_x X = \Omega_1 X$ respectively.

Proposition 5.1.5 (Consistency condition)**.** *The equations $\partial_t X = \Omega_2 X$ and $\partial_x X = \Omega_1 X$ are coupled differential equations. To be consistent they must satisfy:*

$$\frac{\partial\Omega_2}{\partial x} - \frac{\partial\Omega_1}{\partial t} = [\Omega_1, \Omega_2]\,. \tag{5.8}$$

Proof. As second order mixed partial derivatives must agree:

$$\begin{aligned}
\partial_x \partial_t X &= \partial_t \partial_x X, \\
\partial_x \Omega_2 X &= \partial_t \Omega_1 X, \\
\frac{\partial \Omega_2}{\partial x} X + \Omega_2 \frac{\partial X}{\partial x} &= \frac{\partial \Omega_1}{\partial t} X + \Omega_1 \frac{\partial X}{\partial t}, \\
\frac{\partial \Omega_2}{\partial x} X + \Omega_2 \Omega_1 X &= \frac{\partial \Omega_1}{\partial t} X + \Omega_1 \Omega_2 X.
\end{aligned}$$

Therefore,

$$\frac{\partial \Omega_2}{\partial x} - \frac{\partial \Omega_1}{\partial t} = [\Omega_1, \Omega_2]\,. \tag{5.9}$$

□

When the full matrices are substituted into this relation it will return some constraints on κ and τ, and will allow the elimination of μ.

$$\begin{gathered}
\frac{\partial}{\partial x} \begin{bmatrix} 0 & -\tau\kappa & \kappa' \\ \tau\kappa & 0 & -\mu \\ -\kappa' & \mu & 0 \end{bmatrix} - \frac{\partial}{\partial t} \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} = [\Omega_1, \Omega_2]\,, \\
\begin{bmatrix} 0 & -\tau'\kappa - \tau\kappa' - \dot{\kappa} & \kappa'' \\ \tau'\kappa + \tau\kappa' + \dot{\kappa} & 0 & -\mu' - \dot{\tau} \\ -\kappa'' & \mu' + \dot{\tau} & 0 \end{bmatrix} = [\Omega_1, \Omega_2]\,.
\end{gathered}$$

Then deal with the bracket.

$$\begin{aligned}
[\Omega_1, \Omega_2] &= \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} 0 & -\tau\kappa & \kappa' \\ \tau\kappa & 0 & -\mu \\ -\kappa' & \mu & 0 \end{bmatrix} - \begin{bmatrix} 0 & -\tau\kappa & \kappa' \\ \tau\kappa & 0 & -\mu \\ -\kappa' & \mu & 0 \end{bmatrix} \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \\
&= \begin{bmatrix} \tau\kappa^2 & 0 & -\kappa\mu \\ -\tau\kappa' & \tau\kappa^2 + \tau\mu & -\kappa\kappa' \\ -\tau^2\kappa & 0 & \tau\mu \end{bmatrix} - \begin{bmatrix} \tau\kappa^2 & -\tau\kappa' & -\tau^2\kappa \\ 0 & \tau\kappa^2 + \mu\tau & 0 \\ -\kappa\mu & -\kappa'\kappa & \tau\mu \end{bmatrix} \\
&= \begin{bmatrix} 0 & \tau\kappa' & -\kappa\mu + \tau^2\kappa \\ -\tau\kappa' & 0 & -\kappa\kappa' \\ -\tau^2\kappa + \kappa\mu & \kappa\kappa' & 0 \end{bmatrix}
\end{aligned}$$

This leads to

$$\begin{bmatrix} 0 & -\tau'\kappa - \tau\kappa' - \dot{\kappa} & \kappa'' \\ \tau'\kappa + \tau\kappa' + \dot{\kappa} & 0 & -\mu' - \dot{\tau} \\ -\kappa'' & \mu' + \dot{\tau} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \tau\kappa' & -\kappa\mu + \tau^2\kappa \\ -\tau\kappa' & 0 & -\kappa\kappa' \\ -\tau^2\kappa + \kappa\mu & \kappa\kappa' & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & -\tau'\kappa - 2\tau\kappa' - \dot{\kappa} & \kappa'' + \kappa\mu - \tau^2\kappa \\ \tau'\kappa + 2\tau\kappa' + \dot{\kappa} & 0 & \mu' + \dot{\tau} + \kappa\kappa' \\ -\kappa'' + \tau^2\kappa - \kappa\mu & -\mu' - \dot{\tau} - \kappa\kappa' & 0 \end{bmatrix} = 0 \tag{5.10}$$

5.1.3 Equivalent representations

In this section the equivalence between the Lax pair representation and the nonlinear Schrödinger equation will be derived. The function σ will be used to denote the integral of the torsion $\tau(x,t)$ with respect to x:

$$\sigma(x,t) := \int_0^x \tau(u,t)\, du. \tag{5.11}$$

Lemma 5.1.6. *The matrix given in Equation* (5.10)*, which represents the consistency condition of the Lax pair, is equivalent to a pair of coupled differential equations*

$$\begin{aligned} \kappa\dot{\sigma} &= -\tau^2\kappa + \kappa'' - \frac{1}{2}\kappa^3 + A(t)\kappa, \\ \dot{\kappa} &= -(\tau'\kappa + 2\tau\kappa'), \end{aligned}$$

where A is a real function of t.

Proof. Let κ, τ and μ satisfy Equation (5.10). Take the differential equation $\mu' + \dot{\tau} + \kappa\kappa' = 0$ given in Equation (5.10) and integrate with respect to s.

$$\begin{aligned} \int \mu' ds &= -\int \dot{\tau} + \kappa\kappa'\, ds, \\ \mu &= -\dot{\sigma} - \frac{1}{2}\kappa^2 + A(t), \end{aligned} \tag{5.12}$$

Where $A(t)$ is a constant of integration with respect to s so can be a real function of t. This equation can be combined with the second term in the matrix $\kappa'' + \kappa\mu - \tau^2\kappa = 0$

to give:

$$\kappa\dot{\sigma} = -\tau^2\kappa + \kappa'' - \frac{1}{2}\kappa^3 + A(t)\kappa. \tag{5.13}$$

In addition, this equation is coupled with the third differential equation from the matrix given in Equation (5.10)

$$\dot{\kappa} = -(\tau'\kappa + 2\tau\kappa'), \tag{5.14}$$

giving the desired coupled equations.

The converse is also true, let κ, τ and A satisfy Equation (5.13) and Equation (5.14) then Equation (5.13) can be rearranged

$$-\frac{\kappa''}{\kappa} + \tau^2 = -\dot{\sigma} - \frac{1}{2}\kappa^2 + A(t).$$

A new function can be defined $\mu := -\frac{\kappa''}{\kappa} + \tau^2$, which will be well defined as $\kappa > 0$, and coincides with one differential equation in Equation (5.10). With the new function μ, Equation (5.13) becomes

$$\mu = -\dot{\sigma} - \frac{1}{2}\kappa^2 + A(t).$$

Then the partial derivative of μ with respect to s must satisfy

$$\mu' = \dot{\tau} + \kappa\kappa'.$$

This forms the 3 differential equations given by Equation (5.10). □

Proposition 5.1.7. *The Lax pair satisfies the consistency condition given in Proposition 5.1.5 if and only if the Hasimoto transform ψ satisfies a variant of the nonlinear Schrödinger equation,*

$$\frac{1}{i}\frac{\partial\psi}{\partial t} = \frac{\partial^2\psi}{\partial x^2} - (\frac{1}{2}|\psi|^2 - A)\psi,$$

where A is a real valued function given in Equation (5.12).

Proof. First note that by Lemma 5.1.6, the Lax pair satisfies the consistency condition if and only if the functions κ and τ satisfy the coupled differential equations given in Equation (5.13) and Equation (5.14). To prove the forward implication, take the partial

derivative of the Hasimoto transform ψ with respect to t,

$$\frac{1}{i}\frac{\partial\psi}{\partial t} = \frac{1}{i}(\dot{\kappa} + i\kappa\dot{\sigma})\exp(i\sigma).$$

Substitute in the consistency conditions for κ and τ given in Equation (5.14) and (5.13) respectively,

$$\begin{aligned}\frac{1}{i}\frac{\partial\psi}{\partial t} &= \frac{1}{i}\left(-(\tau'\kappa + 2\tau\kappa') + i(-\tau^2\kappa + \kappa'' - \frac{1}{2}\kappa^3 + A(t)\kappa\right)\exp(i\sigma),\\ &= \big(i(\tau'\kappa + 2\tau\kappa') + (\kappa'' - \tau^2\kappa)\big)\exp(i\sigma) - (\frac{1}{2}|\psi|^2 - A)\psi.\end{aligned}$$

Lastly, note that the second order partial derivative of ψ is equal to the first term,

$$= \frac{\partial^2\psi}{\partial x^2} - (\frac{1}{2}|\psi|^2 - A)\psi.$$

This is the required equation.

The converse implication can be proven simply by running the above calculation backwards. Let the Hasimoto transform satisfy the Schrödinger equation and expand in terms of κ and τ,

$$\begin{aligned}\frac{1}{i}\frac{\partial\psi}{\partial t} &= \frac{\partial^2\psi}{\partial x^2} - (\frac{1}{2}|\psi|^2 - A)\psi,\\ \frac{1}{i}(\dot{\kappa} + i\kappa\dot{\sigma})\mathrm{e}^{i\sigma} &= \big(i(\tau'\kappa + 2\tau\kappa') + (\kappa'' - \tau^2\kappa)\big)\,\mathrm{e}^{i\sigma} - (\frac{1}{2}|\psi|^2 - A)\psi,\\ \frac{1}{i}(\dot{\kappa} + i\kappa\dot{\sigma})\mathrm{e}^{i\sigma} &= \frac{1}{i}\left(-(\tau'\kappa + 2\tau\kappa') + i(-\tau^2\kappa + \kappa'' - \frac{1}{2}\kappa^3 + A(t)\kappa\right)\mathrm{e}^{i\sigma}.\end{aligned}$$

Separate into real and imaginary parts,

$$\begin{aligned}\kappa\dot{\sigma} &= -\tau^2\kappa + \kappa'' - \frac{1}{2}\kappa^3 + A(t)\kappa,\\ \dot{\kappa} &= -(\tau'\kappa + 2\tau\kappa'),\end{aligned}$$

these are Equation (5.13) and Equation (5.14) respectively. □

Proposition 5.1.8. *If Θ is a solution to the standard defocussing nonlinear Schrödinger*

equation,

$$\frac{1}{i}\frac{\partial \Theta}{\partial t} = \frac{\partial^2 \Theta}{\partial x^2} - \frac{1}{2}|\Theta|^2\Theta,$$

then $\psi = \Theta \exp(i \int A\ dt)$ *is a solution to the form of the nonlinear Schrödinger equation given in Proposition 5.1.7*

5.2 Hasimoto transform

To summarise the reformulation of the NLSE into the evolution of a frame this section is taken from an earlier work [8]. We recall the Hasimoto [35] transform, which associates with a solution $\psi \in C^2$ of the nonlinear Schrödinger equation (Equation 5.1) a space curve in $\mathbb{R}^3$ with moving frame $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$; Hasimoto considered the case $\beta = -1/2$. In the present context, ψ is associated with the space derivative of a tangent vector $\mathbf{t}$ to a unit speed space curve, so the curvature is $\kappa = \|\frac{\partial \mathbf{t}}{\partial x}\|$. We have a polar decomposition $\psi = \kappa e^{i\sigma}$ where $\sigma(x,t) = \int_0^x \tau(y,t)dy$ and τ is the torsion. Then the Serret–Frenet formula is

$$\frac{\partial}{\partial x}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}, \tag{5.15}$$

so the frame develops along the space curve. Let $X = [\mathbf{t}; \mathbf{n}; \mathbf{b}] \in SO(3)$, and $\Omega_1(x,t)$ the matrix in Equation (5.15). When $\Omega_1(\cdot, t) \in C(\mathbb{T}; so(3))$, the solution $X(\cdot, t) \in C([0, 2\pi]; SO(3))$ to Equation (5.15) is 2π periodic up to a multiplicative monodromy factor $U(t) \in SO(3)$ such that $X(x + 2\pi, t) = X(x,t)U(t)$.

The frame also evolves with respect to time, so that with $\mu = -\frac{\partial \sigma}{\partial t} - \beta\kappa^2$, we have

$$\frac{\partial}{\partial t}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & -\tau\kappa & \frac{\partial \kappa}{\partial x}\\ \tau\kappa & 0 & -\mu\\ -\frac{\partial \kappa}{\partial x} & \mu & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}. \tag{5.16}$$

Let Ω_2 denote the matrix in Equation (5.16). For a pair of coupled ODE $dX/dx - \Omega_1 X = 0$ and $dX/dt - \Omega_2 X = 0$, the corresponding Lax pair is

$$\frac{\partial \Omega_1}{\partial t} - \frac{\partial \Omega_2}{\partial x} + \big[\Omega_1, \Omega_2\big] = 0.$$

Lemma 5.2.1. *(Hasimoto) If ψ is a C^2 function that satisfies the nonlinear Schrödinger equation, then the coupled pair of differential equations is consistent in the sense that there exists a local solution of the pair of ODE, and there exists a local solution of Lax pair.*

Thus the frame $X \in SO(3)$ evolves along the solution $P + iQ \in B_K$ of NLS, and we can regard $d/dx - \Omega_1$ and $d/dt - \Omega_2$ as connections for this evolution. Both of the coefficient matrices are real and skew symmetric. One can check that a solution of the integral equation

$$X(x,t) = X_0(x) + t\Omega_2(0,0)X_0(0) + \int_0^x \int_0^t \Big(\frac{\partial \Omega_1(y,s)}{\partial t} + \Omega_1(y,s)\Omega_2(y,s)\Big)X(y,s)dsdy \tag{5.17}$$

satisfies

$$X(x,0) = X_0(x), \qquad \frac{\partial X(x,0)}{\partial t} = \Omega_2(x,0)X_0(x),$$
$$\frac{\partial^2 X(x,t)}{\partial x \partial t} = \Big(\frac{\partial \Omega_1(x,t)}{\partial t} + \Omega_1(x,t)\Omega_2(x,t)\Big)X(x,t),$$

so smooth solutions are given in terms of an integral equation.

From the Serret–Frenet formulas the components of the acceleration along the space curve satisfy

$$\Big\|\mathbf{t} \times \frac{\partial^2 \mathbf{t}}{\partial x^2}\Big\|^2 = \Big(\frac{\partial \kappa}{\partial x}\Big)^2 + \kappa^2\tau^2 = \Big(\frac{\partial Q}{\partial x}\Big)^2 + \Big(\frac{\partial P}{\partial x}\Big)^2,$$
$$\Big(\mathbf{t} \cdot \frac{\partial^2 \mathbf{t}}{\partial x^2}\Big)^2 = \kappa^4 = \big(P^2 + Q^2\big)^2. \tag{5.18}$$

The total curvature of the space curve is

$$\int_{\mathbb{T}} \kappa(x)^2 dx = \int_{\mathbb{T}} (P^2 + Q^2)dx = H_1(P,Q), \tag{5.19}$$

which is an invariant under the flow associated with the NLS.

Chapter 6

The Euler equations

This chapter focusses on the isentropic Euler equations of fluid motion, which govern the motion of an compressible inviscid fluid. The dynamics of this motion are governed by pressure forces as a result of the varying density throughout the fluid. Some of the assumptions made in deriving these equations, and justifications for their relevance to fluid motion are outlined in the first section. Following this, an alternative form of the Euler equations which depend on the internal energy of the fluid instead of the pressure is derived. Then the coordinate frame for the problem is shifted to Lagrangian coordinates, and the differential equation for the evolution of the Lagrangian velocity is defined.

In the second section a standard existence and uniqueness theorem for smooth solutions on a bounded interval is given. This is then applied in the third section to discuss solutions constrained to the sphere. Some necessary conditions on the potential are derived so that solutions may exist on the sphere, and the co-moving frame attached to a given solution curve is discussed in detail. Lastly, solutions to the Lagrangian ODE are found to be consistent with the linear transport equation (or continuity equation) and an interval for which this consistency remains is estimated.

The Euler equations are further discussed in Chapter 7, in which they are compared to the Saint-Venant equations. Then later in Chapter 13 transport-based methods for numerical solutions to the Euler equations are discussed. This method is capable of modelling densities of measures absolutely continuous with respect to Lebesgue measure, which can be non-continuous functions.

6.1 The Euler equations

The isentropic Euler equations describe the motion of a compressible inviscid fluid due to the variations in density throughout the fluid. How the variations of the density produce pressure forces is approximated using the thermodynamic potential . The word 'Isentropic' refers to the system being modelled as a reversible, adiabatic process. An adiabatic process is one in which no heat or mass is transferred from the system to its surroundings, and to be adiabatic and reversible is to maintain a constant entropy. The term 'Euler equation' is used for a wide variety of different equations from many fields of mathematics, in this thesis however we reserve the term specifically for the isentropic Euler equations presented here.

Definition 6.1.1. Let $\rho(x,t) : M \times [0,\tau] \to \mathbb{R}_+$ denote a density and $u(x,t) : M \times [0,\tau] \to TM$ denote a velocity field on the Riemannian manifold M and its tangent space TM. The Euler equations are

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{6.1}$$

$$\partial_t (\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \nabla P(\rho) = 0. \tag{6.2}$$

The function $P(\rho)$ denotes the pressure of the fluid and is related to the internal energy. The explicit form the internal energy takes differs depending on the fluid modelled, and the thermodynamic assumptions made, and in this work is given in Definition 6.1.2.

Definition 6.1.2. The internal energy for an adiabatic ideal gas is given by

$$U(\rho) = \frac{\kappa \rho^\gamma}{\gamma - 1}, \tag{6.3}$$

where κ is a generic constant and γ denotes the adiabatic index which is equal to the ratio of the heat capacities, or degrees of freedom. The pressure is related to the internal energy by [74, p.156],

$$P(\rho) = U'(\rho)\rho - U(\rho), \tag{6.4}$$

and so the pressure forces within an ideal gas undergoing adiabatic dynamics is given by $P(\rho) = \kappa \rho^\gamma$.

Definition 6.1.3. The adiabatic index, γ, which is also known as the ratio of specific

heats [45, p.20] is

$$\gamma = \frac{C_P}{C_V} = \frac{f+2}{f}, \tag{6.5}$$

where C_P is the heat capacity of the gas at constant pressure, and C_V the equivalent at constant volume, and f simply denotes the degrees of freedom of the gas.

A monatomic gas has 3 degrees of freedom $f = 3$ corresponding to the 3 translations in $\mathbb{R}^3$, it has no rotational degrees of freedom on account of its spherical symmetry. A diatomic gas is visualised as two spheres linked together, as such it has one axis of rotation in which the shape remains invariant, the symmetry along this axis of rotation removes one possible rotational degree of freedom, leaving two out of the possible three rotations in $\mathbb{R}^3$ (think Euler angles) and the original three translations. A monatomic gas has $\gamma = 5/3$ and a diatomic gas has $\gamma = 7/5$.

Remark 6.1.4. The Euler equations are constructed to describe the motion of a fluid and to retain this useful symmetry with an observable physical system, the system must not deviate outside realistic values. For example, Equation (6.1) is simply derived from a continuum hypothesis, that is, the fluid - a collection of randomly moving particles of random velocities distributed according to a Maxwell-Boltzmann distribution - acts locally like a continuous deformable volume under the collective motion of a phase velocity u. This is a well founded assumption under which the field of fluid dynamics is separated from statistical mechanics. But it is only an approximation, and the approximation only remains valid when the length scale of the problem (L) is large compared to the molecular mean free path of the fluid in question (l). The molecular mean free path of a fluid is the average distance a particle travels before it collides with another. It is a microscopic quality of the gas but it relates to the macroscopically observed phenomena of momentum diffusion, heat capacity and species (type of gas). The ratio of mean free path to length scale is known as the Knudsen's number $K_n = L/l$, and the continuum hypothesis is empirically observed to be valid for $K_n << 1$. For a gas such as air at room temperature and pressure the mean free path is $l = 1 \times 10^{-6} m$, but for rarefied gases in the upper atmosphere the mean free path is much larger and this could pose an issue [45, p.6].

6.1.1 Equivalent definitions of Euler System

Consider the Euler equations as given in Definition 6.1.1. Assume the manifold in question is $\mathbb{R}^3$ or embedded in $\mathbb{R}^3$, the highest dimensional space we will be working within this chapter.

Proposition 6.1.5. *The pair of functions (ρ, u) as in Definition 6.1.1 are solutions to the Euler equations if and only if they are solutions to the following,*

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{6.6}$$

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla U'(\rho) = 0. \tag{6.7}$$

Proof. The continuity equation is unchanged. Equation (6.7) is equivalent to Equation (6.2) thanks to the following identities. First,

$$\nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) = \rho(\mathbf{u} \cdot \nabla)\mathbf{u} + \mathbf{u}\, \nabla \cdot (\rho \mathbf{u})$$

is a vector algebraic identity for a higher order tensor. And secondly, the gradient of the thermodynamic equation of pressure given in Equation (6.4) is,

$$\nabla P(\rho) = \rho \nabla U'(\rho) + U'(\rho)\nabla\rho - U'(\rho)\nabla\rho \tag{6.8}$$

$$= \rho \nabla U'(\rho). \tag{6.9}$$

Therefore,

$$\begin{aligned} \partial_t(\rho\mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \nabla P(\rho) = 0 \qquad &\iff \\ \rho \partial_t \mathbf{u} + \mathbf{u}\partial_t\rho + \rho(\mathbf{u} \cdot \nabla)\mathbf{u} + \mathbf{u}\nabla \cdot (\rho\mathbf{u}) + \rho \nabla U'(\rho) = 0 \qquad &\iff \\ \mathbf{u}\left(\partial_t \rho + \nabla \cdot (\rho \mathbf{u})\right) + \rho\left(\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla U'(\rho)\right) = 0, & \end{aligned}$$

the first parenthesis of this equation will be zero by the continuity equation, and leaving aside the trivial edge case of both ρ and $\mathbf{u}$ being everywhere zero, the final if and only if statement is,

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla U'(\rho) = 0.$$

□

6.1.2 The Euler equations in Lagrangian form

The Euler equations expressed in Equation (6.1) are in Eulerian coordinates, a coordinate system which is fixed to the domain of the problem. Sometimes a differential equation can be simplified by transforming the problem into a new frame of reference . Lagrangian coordinates are co-moving frames of reference for the particles of the fluid under motion. Let $\mathbb{R}_a$ denote the Eulerian coordinate system, and $\mathbb{R}_x$ denote a isomorphic space just referred to by the variable x. The function $X(x,t) : \mathbb{R}_x \times [0,\infty) \to \mathbb{R}_a$ maps the particle located at x at time $t = 0$ to the position in the Eulerian frame at which it is located at time t. In this case we consider the initial density $\rho(x,0)$ to be the initial positions of the particles, in other words $\rho(X(a,0)0) = \rho(x,0)$, and so $X(a,0) = x$.

Proposition 6.1.6. *If there exists $u : \mathbb{R}\times[0,\tau] \to \mathbb{R}$ such that $X \mapsto u(X,t)$ is Lipschitz, then for $x \in \mathbb{R}^n$ the differential equation*

$$\frac{\partial}{\partial t}X(a,t) = u(X(a,t),t), \tag{6.10}$$

where u is the Eulerian velocity, has unique solution. We will now define $V(a,t)$ to be the Lagrangian velocity which is just the composition $V(a,t) = (u \circ X)(a,t)$.

This proposition is fundamental to the concept of a Lagrangian velocity, it will be proven in conjunction with the rest of the Lagrangian frame in Section 6.2.

Material derivative

Consider the scalar and vector valued functions, $f : \mathbb{R}^3\times[0,\tau) \to \mathbb{R}$ and $F : \mathbb{R}^3\times[0,\tau) \to \mathbb{R}^3$. The 'derivative' of these functions along the trajectory specified by $X(a,t)$, recalling that a is just a label, is given by the multivariable chain rule

$$\frac{d}{dt}f(X(a,t),t) = \frac{\partial}{\partial t}f(X,t) + \frac{\partial X}{\partial t}\cdot\nabla f(x,t), \tag{6.11}$$

$$\frac{d}{dt}F(X(a,t),t) = \frac{\partial}{\partial t}F(X,t) + \left(\frac{\partial X}{\partial t}\cdot\nabla\right)F(x,t). \tag{6.12}$$

Then both equations can written in terms of the Lagrangian velocity $\partial_t X = V$.

Definition 6.1.7. The Euler equations in Lagrangian form are a pair of ordinary differential equations for $X(a,t) : \mathbb{R}^3 \times [0,\tau) \to \mathbb{R}^3$ and $V(a,t) : \mathbb{R}^3 \times [0,\tau) \to \mathbb{R}^3$ given by,

$$\frac{d}{dt}\begin{pmatrix} X \\ V \end{pmatrix} = \begin{pmatrix} V \\ -\nabla U'(\rho(X,t)) \end{pmatrix}, \tag{6.13}$$

where a, the labelling of the space at time zero is seen as an initial condition and thus fixed with respect to time.

A simplifying assumption is made in many of the examples discussed in that the density $\rho(X,t) = \rho(X)$ produces an autonomous differential equation for X, V. In other words, the density has no explicit time dependence.

Lemma 6.1.8. *Consider the pair (ρ, u), which is a solution to the Euler equations as outlined in Proposition 6.1.5. Then, the pair (X, V) solve the Lagrangian form of the Euler equations as given in Definition 6.1.7.*

Proof. The first row of Equation 6.16, $\partial_t X = V$ is simply Equation (6.13), the definition of the Lagrangian velocity. The second row is due to Equation (6.7). Consider the material derivative of a smooth vector function $F(X(a,t),t)$ with respect to time, as given in equation Equation (6.12). When this relation is applied to the function $\mathbf{u}(X(a,t),t)$ then the total derivative is,

$$\begin{aligned} &\partial_t \mathbf{u}(X(a,t),t) + (\mathbf{u}(X(a,t),t) \cdot \nabla)\, \mathbf{u}(X(a,t),t) + \nabla U'(\rho(X(a,t),t)) = 0, \\ &\frac{d}{dt}\mathbf{u}(X(a,t),t) = -\nabla U'(\rho(X(a,t),t)), \\ &\qquad \frac{d}{dt}V(a,t) = -\nabla U'(\rho(X(a,t),t)). \end{aligned}$$

□

Remark 6.1.9. The differential equations for X and V (Equation (6.13)) are a reframing of Equation (6.7), but have decoupled this equation from the continuity equation (Equation (6.6)). Thus it must be established if the density discussed in the following sections is a solution to the continuity equation as well. This subject will be discussed in Section 6.6

Remark 6.1.10. The differential equations (Equation (6.13)) do not describe how the the density changes over time. For the following section we make the assumption that the densities are invariant with respect to time.

6.2 Existence theory for solutions on $\mathbb{S}^2$

With suitable initial conditions, one can establish the existence and uniqueness of continuous solutions to the Euler equations via classical theorems based on Picard iteration in a closed and bounded subset of Euclidean space.

Theorem 6.2.1 (Birkhoff-Rota)**.** *[4, p. 113] If $g(y,t) : \mathbb{R}^3 \times [0,T] \to \mathbb{R}^3$ is a continuous function on a region $R \times [0,T]$ where $R = \{y \in \mathbb{R}^3 \mid \|y - y_0\| \leq r\}$ and is Lipschitz in the y variable, $\|g(y_2,t) - g(y_1,t)\| \leq K\|y_2 - y_1\|$ for all $t \in [0,\tau]$ and $y_1, y_2 \in R$. Then the differential equation,*

$$\frac{dy}{dt} = g(y,t) \tag{6.14}$$

with initial condition $y(0) = y_0$ has a unique solution on the interval $[0,\tau]$ where $\tau = \min\left(T, r/\sup_{R\times[0,T]} |g(y,t)|\right)$.

Proposition 6.2.2. *Let $\rho(x) \in C_b^2(\mathbb{R}^3, \mathbb{R}_+)$ be a density which is invariant over time and consider only $\rho(x)$ which have a minimum density $\rho_0 = \inf_{x\in\mathbb{R}^3}\{\rho(x)\} > 0$. Consider a region, Ω contained within a cuboid $R \subset \mathbb{R}^3$ such that $\int_\Omega \rho(x)dx = 1$. If $U(\rho) = \rho^\gamma$, then $\nabla U'(\rho(X))$ is continuous and Lipschitz on Ω. Thus the map,*

$$\begin{bmatrix} X \\ V \end{bmatrix} \mapsto \begin{bmatrix} V \\ \nabla U'(\rho(X)) \end{bmatrix} \tag{6.15}$$

is also Lipschitz. Therefore, the system of ODEs given in Lemma 6.1.8 has a solution on the interval $t \in [0,\tau]$ where $\tau = r/\sup_{x\in R} |\nabla U'(\rho(x))|$, and the solution is unique.

Proof. By Theorem 6.2.1 the ODE has a unique solution provided $-\nabla U'(\rho(X))$ Lipschitz continuous. First note $U'(\rho) = \gamma\rho^{\gamma-1}$, and thus $\nabla U'(\rho) = \gamma(\gamma-1)\rho^{\gamma-2}\nabla\rho$. Under the assumption that $\rho \in C_b^2$, then $\nabla\rho \in C_b^1$.

The function $f(x) = x^{\gamma-2}$ is differentiable on $\mathbb{R}_+$ provided the exponent is larger than 1. For $\gamma \in (1,2]$ however, the exponent will be smaller than 1, meaning $f(x) =$

$x^{\gamma-2}$ will not be differentiable at 0. By specifying a positive minimum density, the point of non-differentiability of $f(\rho) = \rho^{\gamma-2}$ is avoided. Hence the additional assumption of a minimum density $\rho_0 = \inf_{x\in\mathbb{R}^3}\{\rho(x)\} > 0$ implies that $f \circ \rho \in C_b^1$ and therefore $\nabla U'(\rho(x)) \in C_b^1$. Finally, a differentiable function is Lipschitz by the intermediate value theorem, with a Lipschitz constant equal to the maximum of the derivative, thus the ODE $\frac{dV}{dt} = \nabla U'(\rho(x))$ is Lipschitz and has a unique solution. With existence and uniqueness of V proven, its continuity on a bounded set imply the existence of a solution to $\frac{dX}{dt} = V$ by Peano's existence theorem [4]. □

6.3 On the sphere

Consider the ODE for $X(x,t) : \mathbb{R}^3 \times [0,\tau) \to \mathbb{R}^3$ and $V(x,t) : \mathbb{R}^3 \times [0,\tau) \to \mathbb{R}^3$ given in Definition 6.1.7. To adapt this differential equation so that solutions lie on the sphere one must introduce the lagrange multiplier λ.

$$\frac{\partial}{\partial t}\begin{pmatrix} X \\ V \end{pmatrix} = \begin{pmatrix} V \\ \lambda X - \nabla U'(\rho). \end{pmatrix} \tag{6.16}$$

Constraining the solution to the sphere implies that $\int \|x\|^2\rho(x)dx = 1$, hence by the method of Lagrange multipliers one can add a term to the Hamiltonian for the problem,

$$H(q,\rho) = \frac{1}{2}\int \|\nabla q\|^2\rho dx + \int U(\rho)dx + \lambda\left(\frac{1}{2}\int \|x\|^2\rho dx - 1\right). \tag{6.17}$$

The Euler equations on the sphere should be the canonical equations of this Hamiltonian.

$$\begin{aligned} \frac{\delta H}{\partial q} &= \frac{\partial H}{\partial q} - \nabla\frac{\partial H}{\partial \nabla q}, \\ &= -\nabla\cdot(\nabla q\rho), \\ \frac{\partial \rho}{\partial t} &= -\nabla\cdot(\nabla q\rho), \end{aligned}$$

Giving the continuity equation, and then the velocity update is

$$\begin{aligned}
\frac{\delta H}{\partial \rho} &= \|\nabla q\|^2 + U'(\rho) + \frac{\lambda}{2}\|x\|^2, \\
-\frac{\partial q}{\partial t} &= \frac{1}{2}\|u\|^2 + U'(\rho) + \frac{\lambda}{2}\|x\|^2, \\
\nabla\frac{\partial q}{\partial t} &= -\frac{1}{2}\nabla\|u\|^2 - \nabla U'(\rho) - \lambda x, \\
\frac{\partial u}{\partial t} &= -\frac{1}{2}\nabla\|u\|^2 - \nabla U'(\rho) - \lambda x.
\end{aligned}$$

Thus the Lagrangian velocity equation becomes

$$\begin{aligned}
\frac{d}{dt}V(a,t) &= \frac{\partial u}{\partial t} + \langle u, \nabla u\rangle, \\
\frac{\partial}{\partial t}V(a,t) &= -\frac{1}{2}\nabla\|u\|^2 - \nabla U'(\rho(X)) - \lambda X + \langle u, \nabla u\rangle, \\
&= -\nabla U'(\rho(X)) - \lambda X.
\end{aligned}$$

And so the coupled equations in the Lagrangian frame are as expressed in Equation (6.16), where the sign of λ is not important.

Proposition 6.3.1. *Consider the ODE with $\lambda = 1$ and initial conditions $\|X_0\| = 1$, $\langle X_0, V_0\rangle = 0$. If the density is constant (and thus the gradient of the internal energy is zero), then the problem reduces to a geodesic on the sphere.*

Proof. The ODE can be reframed as the equation

$$\frac{\partial}{\partial t}\begin{pmatrix} X \\ V \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} X \\ V \end{pmatrix},$$

for which the solution

$$\begin{pmatrix} X \\ V \end{pmatrix} = \exp\left(t\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right)\begin{pmatrix} X_0 \\ V_0 \end{pmatrix},$$

is evident. The skew symmetric nature of this constant matrix and its inclusion into $\mathfrak{su}(2)$ allow for its exponential to be calculated explicitly. As the second Pauli matrix,

let us represent the matrix by the symbol σ_2. Note that $\sigma_2^2 = -I$, and thus

$$\begin{aligned}
\exp\left(t\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\right) &= \sum_{n=-\infty}^{\infty} \frac{t^n \sigma_2^n}{n!}, \\
&= \sum_{n=-\infty}^{\infty} \frac{t^{2n}(-1)^n}{2n!} I + \sum_{n=-\infty}^{\infty} \frac{t^{2n+1}(-1)^n}{(2n+1)!}\sigma_2, \\
&= \cos(t) I + \sin(t)\sigma_2, \\
&= \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}.
\end{aligned}$$

This matrix is recognisable as a rotation of t radians around a fixed axis, it will clearly translate the frame $[X_0, V_0, X_0 \times V_0]$ along a geodesic on $\mathbb{S}^2$. □

Proposition 6.3.2. *Consider the ODE given in Equation* (6.16)*, with a density given by* $\rho(r,\theta) = C - A^2 \sin B \sin(\theta) r$ *and* $\gamma = 2$ *and* $\lambda = 0$*. The problem has solution*

$$X(t) = \hat{r}(B, At) = \begin{pmatrix} \sin(B)\cos(At) \\ \sin(B)\sin(At) \\ \cos(B) \end{pmatrix}.$$

and this solution corresponds to motion around the parallel at co-latitude $\theta = B$ *at constant angular velocity* A*. Provided that* $A^2 < C$*, the density* $\rho(r,\theta)$ *is positive out to a radius larger than* 1 *and the problem has a unique solution. For each valid* $A \in \mathbb{R}$ *there are a family of densities* $\{\rho_B(r,\theta) = C - A^2 \sin B \sin(\theta) r \mid B \in (0,\pi)\}$ *each with their own unique solution curve* $X_B(t)$*.*

Before proving the proposition, the intuition behind the choice of density can be explained. For $\gamma = 2$ the acceleration term $\nabla U'(\rho) = 2\nabla\rho$. In spherical polar coordinates the gradient operator is $\nabla = (\frac{\partial}{\partial r}, \frac{1}{r}\frac{\partial}{\partial\theta}, \frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi})$, and so the density $\rho(r,\theta) = A^2 \sin B \sin(\theta) r$ gives a gradient

$$\nabla\rho = -A^2 \sin B \sin\theta\, \hat{r} - A^2 \sin B \cos\theta\, \hat{\theta} + 0\, \hat{\varphi}. \tag{6.18}$$

Here we can already note that $-\sin\theta\, \hat{r} - \cos\theta\, \hat{\theta}$ is a unit vector pointing inwards along the radius of the parallel at θ (the word *radius* being used in light of the parallel

being a circle embedded in $\mathbb{R}^3$, not to be confused with the radial coordinate in spherical polar coordinates), calling this unit vector $\hat{R}_\theta$, $\nabla\rho = (A^2 \sin B)\, \hat{R}_\theta$. Circular motion is defined as motion in which the acceleration of the body is pointing inwards radially with magnitude $\omega^2 r$ where ω is the angular velocity of the body, and r the radius of the orbit. The parallel at $\theta = B$ has radius $\sin B$ and thus this ODE appears set up for circular motion along the parallel at $\theta = B$ with angular velocity A.

Proof. To prove the curve $X(t)$ is a solution to Equation (6.16) the second derivative is calculated.

$$\frac{d}{dt}X(t) = A \begin{pmatrix} -\sin(B)\sin(At) \\ \sin(B)\cos(At) \\ 0 \end{pmatrix}, \tag{6.19}$$

$$\frac{d^2}{dt^2}X(t) = -A^2 \begin{pmatrix} \sin(B)\cos(At) \\ \sin(B)\sin(At) \\ 0 \end{pmatrix}.$$

To compare this vector with the gradient of ρ which we expressed in spherical polar coordinates we must decompose it into its $(\hat{r}, \hat{\theta}, \hat{\varphi})$ components at the point (B, At)

$$\begin{aligned} -A^2 \begin{pmatrix} \sin(B)\cos(At) \\ \sin(B)\sin(At) \\ 0 \end{pmatrix} &= -A^2 \begin{pmatrix} (\sin^3(B) + \sin(B)\cos^2(B))\cos(At) \\ (\sin^3(B) + \sin(B)\cos^2(B))\sin(At) \\ \sin^2(B)\cos(B) - \sin^2(B)\cos(B) \end{pmatrix}, \\ &= -A^2 \sin^2(B) \begin{pmatrix} \sin(B)\cos(At) \\ \sin(B)\sin(At) \\ \cos(B) \end{pmatrix} \\ &\quad - A^2 \sin(B)\cos(B) \begin{pmatrix} \cos(B)\cos(At) \\ \cos(B)\sin(At) \\ -\sin(B) \end{pmatrix}, \\ &= -A^2 \sin^2(B)\hat{r}(B, At) - A^2 \sin(B)\cos(B)\hat{\theta}(B, At). \end{aligned} \tag{6.20}$$

Comparing this with $\nabla\rho$ as given in Equation (6.18) evaluated at coordinates (B, At)

shows that,

$$\frac{d^2}{dt^2}X(t) = \nabla\rho,$$

and $(X(t), V(t))$ where $V(t) := \frac{d}{dt}X(t)$ is the solution to the ODE. □

Remark 6.3.3. It is worth noting that plugging $(\theta(t), \varphi(t)) = (B, At)$ into Equation (2.6) gives Equation (6.20) directly, illustrating the usefulness of the frames based approach.

Remark 6.3.4. The above proposition is expressed with respect to a local coordinate system. The base of this coordinate system can be rotated through any element of $SO(3)$ and thus a curve, $X_B(t)$ tracing out any circle on the unit sphere will be a unique solution to the ODE (Equation (6.16)) with density $\rho_B(t)$ in some coordinate system.

6.4 General solutions on the sphere

First to caveat, the method used to evaluate the ODE here have been unable to specify a general set of simple conditions on ρ such that there exists a unique solution constrained to the sphere. The most general set of smooth solutions on the sphere are given by a unit speed parameterisation that satisfies $\nabla U'(\rho) = k_g\hat{r} \times \dot{\gamma}$, where k_g, the geodesic curvature of the curve, may vary.

The culmination of this avenue of thought is to express what properties the potential term must possess to constrain the solution to a general curve on the sphere. The question of whether a density could produce such a potential term is also discussed. The representation of a general curve on the sphere has been discussed in Section 1.2 and Section 2.1, depending on whether the Frenet-Serret frame is used or not.

6.4.1 Spherical geometric frame

Using the notation of Equation (2.2), let us consider a curve $\gamma(t)$ which satisfies the Euler equations in Lagrangian form, Equation (6.16). This curve thus satisfies the

differential equation,

$$\frac{d}{dt}\begin{bmatrix}\gamma(t)\\ \dot\gamma(t)\end{bmatrix} = \begin{bmatrix}\dot\gamma(t)\\ \lambda\hat r - \nabla U'\end{bmatrix}. \tag{6.21}$$

Ideally this pair of ODEs could be expanded into a frame, $[\gamma, \dot\gamma, \gamma\times\dot\gamma]^\top$. The best approach for a orthonormal frame of a general curve in $\mathbb{R}^3$ is to use the Serret-Frenet vectors $[\gamma', \gamma''/\|\gamma''\|, \gamma'\times\gamma''/\|\gamma'\times\gamma''\|]^\top$, but this requires another derivative of the curve to be calculated. Provided the curve is on the surface of a sphere, one can exploit the fact that the vector $\gamma(t)$ will always be orthogonal to the velocity $\dot\gamma$ and therefore the frame $[\gamma, \dot\gamma, \gamma\times\dot\gamma]^\top$ can be orthogonal. Consequently, the following arguments are slightly contrived — the only way the matrix differential equations actually define the evolution of a frame is if the curve lies on the sphere. Nevertheless, the first step in constructing a matrix version of the differential equations is through the construction of orthonormal vectors to express $\nabla U'$ with respect to, and this can be done by the Gram-Schmidt process.

Lemma 6.4.1. *The following vectors form an orthonormal basis for* $\mathbb{R}^3$*,*

$$e_1 = \frac{\gamma}{\|\gamma\|}, \qquad e_2 = \frac{\gamma\times\dot\gamma}{\|\gamma\times\dot\gamma\|},$$
$$e_3 = \left(\dot\gamma - \frac{\langle\gamma,\dot\gamma\rangle}{\|\gamma\|^2}\gamma\right)\|\dot\gamma - \frac{\langle\gamma,\dot\gamma\rangle}{\|\gamma\|^2}\gamma\|^{-1}.$$

Where the notation $\sigma = \|\dot\gamma - \frac{\langle\gamma,\dot\gamma\rangle}{\|\gamma\|^2}\gamma\|$ *will be used in future applications.*

Proof. Apply the Gram-Schmidt process to orthogonal γ and $\gamma\times\dot\gamma$. □

Let us expand the differential equations in to an equation for the evolution of the frame $[\gamma, \dot\gamma, \gamma\times\dot\gamma]^\top$.

Proposition 6.4.2. *The pair of ordinary differential equations in Equation* (6.21) *can be extended to the evolution of a non-orthogonal frame as*

$$\frac{d}{dt}\begin{bmatrix}\gamma\\ \dot\gamma\\ \gamma\times\dot\gamma\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ \Xi & \Upsilon & -\Lambda\\ -\Lambda\langle\gamma,\dot\gamma\rangle & \Lambda\|\gamma\|^2 & \Upsilon\end{bmatrix}\begin{bmatrix}\gamma\\ \dot\gamma\\ \gamma\times\dot\gamma\end{bmatrix} \tag{6.22}$$

Where

$$\Lambda = -\frac{\langle \nabla U', \gamma \times \dot{\gamma} \rangle}{\|\gamma \times \dot{\gamma}\|^2}$$
$$\Xi = \frac{\lambda}{\|\gamma\|} - \frac{\langle \nabla U', \gamma \rangle}{\|\gamma\|^2} + \frac{\langle \nabla U', \dot{\gamma} \rangle}{\sigma^2} \frac{\langle \dot{\gamma}, \gamma \rangle}{\|\gamma\|^2} - \frac{\langle \nabla U', \gamma \rangle}{\sigma^2} \frac{\langle \dot{\gamma}, \gamma \rangle}{\|\gamma\|^2}^2,$$
$$\Upsilon = -\frac{\langle \nabla U', \dot{\gamma} \rangle}{\sigma^2} + \frac{\langle \nabla U', \gamma \rangle}{\sigma^2} \frac{\langle \gamma, \dot{\gamma} \rangle}{\|\gamma\|^2}.$$

Proof. Express the vector field $\nabla U'$ with respect to the basis e_1, e_2, e_3 and then using Equation (6.21) and vector identities express the resulting mess of inner products in terms of $\gamma, \dot{\gamma}, \gamma \times \dot{\gamma}$. □

Now if $[\gamma, \dot{\gamma}, \gamma \times \dot{\gamma}]^\top$ was a frame, and the matrix in Proposition 6.4.2 was skew symmetric then the curve would stay on the surface of the sphere. Of course, neither seem likely. Nevertheless, making the matrix more skew symmetric might illuminate things, and one way to do that is to make $\Upsilon = 0$.

If our curve $\gamma(t)$ is regular – if $\|\dot{\gamma}(t)\| \neq 0$ for all t – then there exists a second curve $\tilde{\gamma}(s)$ which is parametrised by arc length (Definition 1.1.3). It agrees with $\gamma(t)$ at all points t, $\tilde{\gamma}(s(t)) = \gamma(t)$, but is unit speed [60, Cor.1.3.7]. The derivative of $\tilde{\gamma}(s)$ can be expressed in terms of $\dot{\gamma}(t)$,

$$\tilde{\gamma}'(s) = \frac{d}{ds}\tilde{\gamma}(s(t)) = \frac{1}{\frac{ds}{dt}} \frac{d}{dt}\gamma(t) = \frac{1}{\|\dot{\gamma}(t)\|}\dot{\gamma}(t), \tag{6.23}$$

And the second derivative can be calculated similarly.

Lemma 6.4.3. *The second derivative of $\tilde{\gamma}(s)$ with respect to s can be expressed in terms of the second derivative of $\gamma(t)$ with respect to t by,*

$$\tilde{\gamma}'' = \frac{\ddot{\gamma} - \langle \ddot{\gamma}, \tilde{\gamma}' \rangle \tilde{\gamma}'}{\|\dot{\gamma}\|^2}, \tag{6.24}$$

Where $'$ denotes differentiation with respect to s and $\dot{}$ denotes differentiation with respect to t.

Proof. Apply the quotient rule to the differentiation,

$$\begin{aligned}\frac{d}{ds}\tilde{\gamma}' &= \frac{1}{\frac{ds}{dt}}\frac{d}{dt}\frac{\dot{\gamma}(t)}{\|\dot{\gamma}(t)\|} \\ &= \frac{1}{\|\dot{\gamma}\|}\frac{\ddot{\gamma}\|\dot{\gamma}\| - \dot{\gamma}\frac{d}{dt}\|\dot{\gamma}\|}{\|\dot{\gamma}\|^2}.\end{aligned}$$

The derivative of the norm can be dealt with using the standard inner product on $\mathbb{R}^3$ and its compatibility with the product rule, along with the chain rule,

$$\begin{aligned}\frac{d}{dt}\|\dot{\gamma}\|^2 &= 2\langle\dot{\gamma},\ddot{\gamma}\rangle, \\ \frac{d}{dt}\|\dot{\gamma}\|^2 &= 2\|\dot{\gamma}\|\frac{d}{dt}\|\dot{\gamma}\|, \\ \frac{d}{dt}\|\dot{\gamma}\| &= \frac{\langle\dot{\gamma},\ddot{\gamma}\rangle}{\|\dot{\gamma}\|}.\end{aligned}$$

In conclusion, with use of the identity $\tilde{\gamma}' = \frac{\dot{\gamma}}{\|\dot{\gamma}\|}$, the second derivative can be expressed as in Equation (6.23). □

Proposition 6.4.4. *The frame in Proposition 6.4.2 can be replaced with another one along the same curve, only reparametrised to be unit speed. The new frame obeys the differential equation,*

$$\frac{d}{ds}\begin{bmatrix}\tilde{\gamma}\\ \tilde{\gamma}'\\ \tilde{\gamma}\times\tilde{\gamma}'\end{bmatrix} = \frac{1}{\|\dot{\gamma}\|^2}\begin{bmatrix}0 & \|\dot{\gamma}\|^2 & 0\\ \tilde{\Xi} & 0 & -\tilde{\Lambda}\\ -\tilde{\Lambda}\langle\tilde{\gamma},\tilde{\gamma}'\rangle & \tilde{\Lambda}\|\tilde{\gamma}\|^2 & 0\end{bmatrix}\begin{bmatrix}\tilde{\gamma}\\ \tilde{\gamma}'\\ \tilde{\gamma}\times\tilde{\gamma}'\end{bmatrix} \tag{6.25}$$

Where

$$\begin{aligned}\tilde{\Lambda} &= -\frac{\langle\nabla U',\tilde{\gamma}\times\tilde{\gamma}'\rangle}{\|\tilde{\gamma}\times\tilde{\gamma}'\|^2}, \\ \tilde{\Xi} &= \frac{\lambda}{\|\tilde{\gamma}\|} - \frac{\langle\nabla U',\tilde{\gamma}\rangle}{\|\tilde{\gamma}\|^2} + \frac{\langle\nabla U',\tilde{\gamma}'\rangle}{\sigma^2}\frac{\langle\tilde{\gamma}',\tilde{\gamma}\rangle}{\|\tilde{\gamma}\|^2} - \frac{\langle\nabla U',\tilde{\gamma}\rangle}{\sigma^2}\left(\frac{\langle\tilde{\gamma}',\tilde{\gamma}\rangle}{\|\tilde{\gamma}\|^2}\right)^2.\end{aligned}$$

Proposition 6.4.5. *The following are equivalent,*

(i) There exists a unique solution, $\gamma(t)$, to Euler's equations (Equation (6.16)) for a given $\rho(x)$ which lies on the unit sphere.

(ii) This curve $\gamma(t)$ has a unit speed reparametrisation, $\tilde{\gamma}(s)$, which evolves according to the matrix ODE,

$$\frac{d}{ds}\begin{bmatrix}\tilde{\gamma}\\ \tilde{\gamma}'\\ \tilde{\gamma}\times\tilde{\gamma}'\end{bmatrix} = \frac{1}{\|\dot{\gamma}\|^2}\begin{bmatrix}0 & \|\dot{\gamma}\|^2 & 0\\ \lambda+\langle\nabla U',\tilde{\gamma}\rangle & 0 & \langle\nabla U',\tilde{\gamma}\times\tilde{\gamma}'\rangle\\ 0 & -\langle\nabla U',\tilde{\gamma}\times\tilde{\gamma}'\rangle & 0\end{bmatrix}\begin{bmatrix}\tilde{\gamma}\\ \tilde{\gamma}'\\ \tilde{\gamma}\times\tilde{\gamma}'\end{bmatrix}, \tag{6.26}$$

and this matrix is skew symmetric.

Corollary 6.4.6. *The unit speed parametrisation above has*

$$\kappa_n = \frac{\lambda - \langle\nabla U',\gamma\rangle}{\|\dot{\gamma}\|^2}, \tag{6.27}$$

$$\kappa_g = \frac{\langle\nabla U',\gamma\times\gamma'\rangle}{\|\dot{\gamma}\|^2}. \tag{6.28}$$

In the case of a sphere, $\kappa_n = -1$.

For a curve parametrised by $(\theta(t),\phi(t))$ on the sphere, its geodesic curvature is given explicitly by Equation (2.7).

6.4.2 Using Frenet-Serret frame

If instead the general frame for a curve in $\mathbb{R}^3$ is used, then to calculate the curvature and torsion of this curve a further derivative is required. Under the assumption of this section that the density $\rho(x)$ is time invariant, the impulse of the curve is given by

$$\ddot{\gamma} = J(\nabla U'(\rho))\mathbf{t}, \tag{6.29}$$

where J represents the Jacobian matrix.

Proposition 6.4.7. *The curvature and torsion of the frame which flows along the solution to the Euler equations (Equation* (6.16)*) is given by*

$$\kappa = \|\lambda\hat{r} - \langle\nabla U',\boldsymbol{n}\rangle\boldsymbol{n} - \langle\nabla U',\boldsymbol{b}\rangle\boldsymbol{b}\|,$$
$$\tau = \langle J(\nabla U')\boldsymbol{t},\boldsymbol{b}\rangle.$$

6.4.3 Integral conditions on the solution

In this section, necessary conditions for curves which both solve the Euler equations (Equation (6.21)), and are constrained to the sphere are determined. These restrict the form that $\nabla U'$ can take. The hairy ball theorem says that there does not exist any smooth vector field lying in the tangent space of $\mathbb{S}^2$ which is non-zero at each point [13, Thm 2.2.2]. Thus for the following analysis, the surface S which is a region of the surface of the sphere is assumed not to contain any point for which $\nabla U'$ vanishes. In addition, a curve γ with velocity $\nabla U'$ will not be regular if it passes through a point at which $\nabla U'$ vanishes.

The Gauss'-Bonnet Theorem relates the integral of the Gaussian curvature of a surface to the geodesic curvature of its boundary.

Theorem 6.4.8 (Gauss'-Bonnet Theorem)**.** *If K denotes the Gaussian curvature of the surface S which has boundary equal to the curve γ with geodesic curvature κ_g then*

$$\int_S K \, dS + \int_\gamma \kappa_g \, d\gamma - 2\pi \, \chi(S) = 0, \tag{6.30}$$

where $\chi(S)$ denotes the Euler characteristic of the surface in question, which is calculated using $\chi = V - E + F$, where V, E and F are vertices, edges and faces respectively.

Example 6.4.9. For the class of solutions outlined in Proposition 6.3.2, in which solutions follow parallels, these values can be calculated explicitly. Equation (6.20) shows that $\langle \ddot{\gamma}, \dot{\gamma} \rangle = 0$, and therefore the unit speed parametrisation $\gamma'' = \ddot{\gamma}/\|\dot{\gamma}\|^2$. Note that $\|\dot{\gamma}\|^2 = A^2 \sin^2 B$ by Equation (6.19), then with reference to Equation (2.2) for the definition of κ_g and Equation (2.7) for the explicit value of $\ddot{\gamma}$,

$$\kappa_g = \frac{A^2 \sin B \cos B}{A^2 \sin^2 B} = \frac{\cos B}{\sin B}. \tag{6.31}$$

The integral of this quantity around the parallel at co-latitude $\theta = B$ is

$$\begin{aligned} \int_\gamma \kappa_g \, d\gamma &= \int_0^{2\pi} \frac{\cos B}{\sin B} \sin B \, d\varphi, \\ &= 2\pi \cos B. \end{aligned}$$

On the unit sphere, the Gaussian curvature is 1, making the integral of the Gaussian curvature simply equal to the area of the surface. The surface area of the polar cap with a boundary equal to the parallel at colatitude B is

$$\int_0^{2\pi} \int_0^B \sin\theta \, d\theta \, d\varphi = 2\pi(1 - \cos B).$$

Finally the Euler characteristic of a polar cap is equal to $\chi = 1$, so each term in the Gauss'-Bonnet equation has been calculated and

$$2\pi(1 - \cos B) + 2\pi \cos B - 2\pi = 0.$$

This theorem can be employed in the present context, in combination with Stokes' theorem. Embedding the unit sphere in Euclidean space $\mathbb{R}^3$ allows Stokes' Theorem to be stated using just the language of vector calculus.

Definition 6.4.10 (Stokes' Theorem)**.** The integral of the divergence of a smooth vector field over a surface is equal to the flux of that vector field through the boundary of the surface. Consider a smooth vector field F defined on a compact subset of $\mathbb{R}^3$ enclosing the simply connected surface S on the unit sphere, where the boundary of S is the path of a unit speed curve $\gamma(s)$,

$$\int_S \hat{r} \cdot (\nabla \times F) \, dS = \oint_\gamma F \cdot \mathbf{t} \, ds. \tag{6.32}$$

The vector $\hat{r}$ is everywhere normal to S and $\mathbf{t}$ denotes the tangent to the curve γ as employed in the Frenet-Serret frame.

Direct application of this theorem to $F = \nabla U'$ is not enlightening, as $\nabla \times \nabla\rho$ is zero for any density ρ. However, as it applies to any vector field F, one can consider $F = \hat{r} \times \nabla U'$, then $(\hat{r} \times \nabla U') \cdot \mathbf{t} = \nabla U' \cdot (\mathbf{t} \times \hat{r})$. As γ is on the sphere, $\gamma = \hat{r}$ and so

$\mathbf{t} \times \hat{r} = \gamma' \times \gamma$, thus by Equation (6.32),

$$\begin{aligned}\int_S \hat{r} \cdot (\nabla \times (\hat{r} \times \nabla U'))\, dS &= \oint_\gamma (\hat{r} \times \nabla U') \cdot \mathbf{t}\, ds, \\ &= \oint_\gamma \nabla U' \cdot (\gamma' \times \gamma)\, ds.\end{aligned}$$

The cross product $\nabla \times (\hat{r} \times \nabla U')$ is equal to

$$\nabla \times (\hat{r} \times \nabla U') = (\nabla \cdot \nabla U')\hat{r} - (\nabla \cdot \hat{r})\nabla U' + (\nabla U' \cdot \nabla)\hat{r} - (\hat{r} \cdot \nabla)\nabla U'.$$

Where the use of the inner product is a small abuse of notation, as neither of $\nabla \cdot \hat{r}$ or $\hat{r} \cdot \nabla$ are scalar quantities, they are instead a vector and a differential operator respectively. The inner product with $\hat{r}$ is

$$\begin{aligned}\hat{r} \cdot (\nabla \times (\hat{r} \times \nabla U')) = (\nabla \cdot \nabla U') - (\nabla \cdot \hat{r})(\hat{r} \cdot \nabla U') + \hat{r} \cdot (\nabla U' \cdot \nabla)\hat{r} \\ - \hat{r} \cdot (\hat{r} \cdot \nabla)\nabla U'.\end{aligned}$$

The first and last terms together are equal to the projection onto the tangent space of the gradient of the field, they remove any radial component of the gradient, and shall be denoted $\nabla_{\mathbb{S}^2} F = \nabla \cdot F - \left(\hat{r} \cdot \frac{\partial F}{\partial r}\right) \hat{r}$. If we denote the spherical components of $\nabla U'$ by $(U_r, U_\theta, U_\varphi)$, then $(\nabla U' \cdot \nabla)\hat{r} = \frac{U_\theta}{r}\hat{\theta} + \frac{U_\varphi}{r}\hat{\varphi}$ so it has no radial component. In addition, the divergence of the radial unit vector is $\nabla \cdot \hat{r} = 2/r$ and thus,

$$\hat{r} \cdot (\nabla \times (\hat{r} \times \nabla U')) = \nabla_{\mathbb{S}^2} \nabla U' - 2\frac{(\hat{r} \cdot \nabla U')}{r}.$$

Therefore Stokes theorem implies that,

$$\int_S \nabla_{\mathbb{S}^2} \nabla U' - 2\frac{(\hat{r} \cdot \nabla U')}{r}\, dS = \oint_\gamma \nabla U' \cdot (\gamma' \times \gamma)\, ds. \tag{6.33}$$

Proposition 6.4.11. *Consider the coupled differential equations given in Equation 6.21 for a unit speed γ, and the class Γ of curves which,*

(i) Lie on the surface of the sphere,

(ii) Form the boundary of a simply connected surface patch S.

Then if $\gamma \in \Gamma$ is a solution to Equation (6.21)*, the potential $\nabla U' - \lambda$ must satisfy,*

$$\int_S \nabla_{\mathbb{S}^2} \nabla U' - 2\lambda - 1 \, dS = 2\pi. \tag{6.34}$$

The region S is the patch enclosed by γ.

Proof. From Stokes theorem, and specifically Equation (6.33), a necessary condition on the existence of certain solutions to Equation (6.21) can be constructed. Consider a curve which solves Equation (6.21) and belongs to Γ. By Proposition 6.4.5 the curve has a unit speed reparametrisation, γ, which satisfies Equation (6.26). The integral along γ of its geodesic curvature, κ_g can be calculated by the Gauss' Bonnet theorem. But in addition, by Equation (6.28),

$$\oint_\gamma \kappa_g \, ds = \oint_\gamma \frac{\nabla U' \cdot (\gamma' \times \gamma)}{\|\dot{\gamma}\|^2} \, ds. \tag{6.35}$$

Assuming the solution was originally unit speed, making $\|\dot{\gamma}\|^2 = 1$, then

$$\int_S \nabla_{\mathbb{S}^2} \nabla U' - 2(\hat{r} \cdot \nabla U') \, dS = 2\pi - \int_S dS.$$

As a result, the vector field $\nabla U'$ will only lead to a solution of Equation (6.16) on the sphere (of unit speed) if

$$\int_S \nabla_{\mathbb{S}^2} \nabla U' - 2(\hat{r} \cdot \nabla U') + 1 \, dS = 2\pi.$$

Which can be further reduced by using $\kappa_n = 1$ and Equation (6.27), which imply $\hat{r} \cdot \nabla U' = 1 + \lambda$.

$$\int_S \nabla_{\mathbb{S}^2} \nabla U' - 2\lambda - 1 \, dS = 2\pi.$$

□

6.5 Solutions of a specific case

Having considered conditions on potential internal energies and initial densities which lead to solutions of the Euler equations which lie on the sphere, let us explore a specific problem in detail. Here we postulate a simple model of air pressure on earth, mostly uniform in density, with the cooler poles having denser air and the warmer equator less dense. Real weather data gives realistic values for atmospheric pressure of around 1010hPa (hectoPascals), with reasonable deviations for high and low pressure fronts of roughly ±20hPa [56]. The Legendre polynomials P_0^0 and P_2^0 can be used to produce a density of this form. Let the positive constant A be two orders of magnitude larger than the constant B, then the density may be well approximated by

$$\rho(\theta) = A + \frac{B}{2}(3\cos^2(\theta) - 1). \tag{6.36}$$

This density leads to a gradient of the internal energy of

$$-\nabla U'(\rho(\theta)) = -\gamma(\gamma-1)3B\cos(\theta)\sin(\theta)\left(A + \frac{B}{2}(3\cos^2(\theta)-1)\right)^{\gamma-2}\hat{\theta}. \tag{6.37}$$

With $\gamma = 2$, it reduces to $\nabla U'(\rho(\theta)) = 3B\sin(\theta)\cos(\theta)\hat{\theta}$. Thus the acceleration of the curve is given by the vector space $(-\lambda, -3B\sin(\theta)\cos(\theta), 0)$. This vector field is differentiable and thus Lipshitz on an open subset of $\mathbb{R}^3$ containing the unit sphere, and so the system has a unique solution. What is more, if $\lambda = \|V_0\|^2$ and $\langle V_0, \nabla U'\rangle = 0$ then the only solution to the system of equations which stays on the sphere is a constant speed curve with $\|V\|^2 = -\langle \hat{r}, \nabla U'\rangle = \lambda$.

Splitting the differential equation (Equation (6.16)) into orthogonal components $[\hat{r}, \hat{\theta}, \hat{\varphi}]$ using Equation (2.6),

$$-\lambda = -\left(\dot{\theta}^2 + \dot{\varphi}^2\sin^2(\theta)\right) \tag{6.38}$$

$$-3B\sin(\theta)\cos(\theta) = \ddot{\theta} - \dot{\varphi}^2\sin(\theta)\cos(\theta) \tag{6.39}$$

$$0 = 2\dot{\theta}\dot{\varphi}\cos\theta + \ddot{\varphi}\sin\theta \tag{6.40}$$

And solutions, $X(x,t) = \hat{r}(\theta(t), \varphi(t))$ in local coordinates must satisfy these three equations. This can be solved by direct integration in some cases, and in those cases it

is likely that elliptic integrals will arise.

Definition 6.5.1. An incomplete elliptic integral of the first kind [53, p.57] with complimentary modulus k is denoted,

$$\mathcal{F}(x|k) = \int_0^x \frac{1}{\sqrt{(1-y^2)(1-k^2y^2)}} dy.$$

The substitution of $x = \sin\varphi$ and $y = \sin\theta$ gives the variation in trigonometric form,

$$\mathcal{F}(\sin(\varphi)|k) = \int_0^{\sin\varphi} \frac{1}{\sqrt{1-k^2\sin^2(\theta)}} d\theta. \tag{6.41}$$

Definition 6.5.2. The Weierstrass $\wp$-function defined in the theory of elliptic curves [53, p.87], it inverts the incomplete elliptic integral,

$$\wp^{-1}(x) = \frac{1}{2}\int_\infty^x \frac{dy}{\sqrt{(y-e_1)(y-e_2)(y-e_3)}}. \tag{6.42}$$

Case 1

The simplest case is in which $\dot{\varphi}$ is the trivial solution to Equation (6.40), that is $\dot{\varphi} = 0$ and so $\varphi = a \in [0, 2\pi]$ a constant and the curve oscillates along one line of longitude. In this case Equation (6.39) reduces to a rescaled simple harmonic oscillator,

$$\ddot{\theta} = -\frac{3}{2}B\sin(2\theta),$$

where θ would represent the angle of the pendulum. The equation can be solved by an elliptic integral using separation of variables,

$$\begin{aligned}
\ddot{\theta}\dot{\theta} &= -3B\dot{\theta}\sin(\theta)\cos(\theta), \\
\frac{1}{2}\dot{\theta}^2 &= -\frac{3}{2}B\sin^2(\theta) + E, \\
\int \frac{d\theta}{\sqrt{2E - 3B\sin^2(\theta)}} &= \int dt, \\
\frac{1}{\sqrt{2E}}\mathcal{F}\left(\theta \mid \sqrt{\frac{3B}{2E}}\right) &= t + G.
\end{aligned}$$

Where $\mathcal{F}(x|\, k)$ denotes the incomplete elliptic integral of the first kind with modulus k.

Case 2

Equation (6.40) has a solution when $\varphi(t)$ is wholly dependent on $\theta(t)$, when $\varphi(t) = \frac{C^2}{\sin^2(\theta(t))}$ for some constant $C \in \mathbb{R}$. In this case, Equation (6.39) would reduce to the one variable problem,

$$\ddot{\theta} = \frac{C^2 \cos(\theta)}{\sin^3(\theta)} - 6B \sin(2\theta)\dot{\theta}. \tag{6.43}$$

Consider the case of $B = 0$ first. This corresponds to a uniform density, and so should give solutions which simply flow along geodesics. When $B = 0$ Equation (6.43) can be approached by separation of variables, multiply the equation by $\dot{\theta}$ and integrate to get

$$\frac{1}{2}\dot{\theta}^2 = -\frac{C^2}{2\sin^2(\theta)} + D, \tag{6.44}$$

where D is a constant of integration. Rearrange, square root and separate,

$$\begin{aligned}
\frac{1}{2}\dot{\theta}^2 \sin^2(\theta) &= -\frac{C^2}{2} + D\sin^2(\theta), \\
\frac{1}{2}\frac{\dot{\theta}^2 \sin^2(\theta)}{D\sin^2(\theta) - \frac{1}{2}C^2} &= 1, \\
\frac{1}{\sqrt{2}}\frac{\sin(\theta)}{\sqrt{D\sin^2(\theta) - \frac{1}{2}C^2}}\,\dot{\theta} &= 1, \\
\int \frac{-du}{\sqrt{Du^2 + (D - \frac{1}{2}C^2)}} &= \sqrt{2}\int dt, \\
\frac{1}{\sqrt{D - \frac{1}{2}C^2}} \int \frac{du}{\sqrt{\frac{Du^2}{D - \frac{1}{2}C^2} + 1}} &= -\sqrt{2}\int dt,
\end{aligned}$$

where the substitution $u = \cos(\theta)$ used. This is a rescaled standard integral, if $D >$

$C^2/2$ then the substitution is $u = \sqrt{\frac{D-\frac{1}{2}C^2}{D}}\sin(v)$,

$$\int \frac{1}{\sqrt{D}} \frac{\cos(v)dv}{\sqrt{\sin^2(v)+1}} = -\sqrt{2}\int dt,$$
$$\frac{1}{\sqrt{D}}v = -\sqrt{2}t + E,$$
$$\cos(\theta) = \sqrt{\frac{D-\frac{1}{2}C^2}{D}}\sin(-\sqrt{2D}t + E). \tag{6.45}$$

By solving the system of ODEs knowing the initial position and velocity, valid values of each of the constants can be derived from the initial condition. The constant C can be determined by an initial condition ($C = \dot{\varphi}(0)\sin^2(\theta(0))$) the other constants should then be expressed as functions of C. From Equation (6.44), and Equation (6.45),

$$D(C) = \frac{1}{2}\dot{\theta}(0)^2 + \frac{C^2}{2\sin^2(\theta(0))}, \tag{6.46}$$
$$E(D,C) = \sin^{-1}\left(\sqrt{\frac{D}{D-\frac{1}{2}C^2}}\cos(\theta(0))\right). \tag{6.47}$$

Case 3

The final case to consider is a non-zero B. If $\varphi(t) = \frac{C^2}{\sin^2(\theta(t))}$ then Equation (6.40) is satisfied and Equation (6.39) reduces to

$$\ddot{\theta} = \frac{C^2\cos(\theta)}{\sin^3(\theta)} - 3B\sin(\theta)\cos(\theta). \tag{6.48}$$

This equation can be integrated (though the method will depend on the values the

constants take) so that separation of variables gives

$$\begin{aligned}
\frac{1}{2}\dot{\theta}^2 &= -\frac{1}{2}\frac{C^2}{\sin^2(\theta)} - \frac{3B}{2}\sin^2(\theta) + F, \\
\int dt &= \int \frac{d\theta}{\sqrt{-\frac{C^2}{\sin^2(\theta)} - 3B\sin^2(\theta) + 2F}}, \\
\int dt &= \int \frac{\sin(\theta)d\theta}{\sqrt{-C^2 + 2D\sin^2(\theta) - 3B\sin^4(\theta)}}, \\
\int dt &= \int \frac{-dw}{\sqrt{2D - C^2 - 3B + (6B + 2D)w^2 - 3Bw^4}}.
\end{aligned} \tag{6.49}$$

The substitution $w = \cos(\theta)$ was taken. The denominator resembles a quadratic for w^2 and thus the quadratic formula will yield the roots. In the case in which the constants combine to make the denominator a perfect square (repeated roots require the discriminant to be zero) the constants satisfy $4D^2 - 12BC^2 = 0$, and the integral can be solved via a trigonometric substitution,

$$\begin{aligned}
\int dt &= \int \frac{-dw}{w^2 + \frac{6B+2D}{6B}}, \\
\int dt &= \int \frac{-dw}{\frac{6B+2D}{6B}\tan^2(v) + \frac{6B+2D}{6B}}, \\
\int dt &= \frac{6B}{6B+2D}\int \frac{-dw}{\sec^2(v)}, \\
\int dt &= \sqrt{\frac{6B}{6B+2D}}\int -dv, \\
t + G &= -\sqrt{\frac{6B}{6B+2D}}v,
\end{aligned}$$

where the substitution $\tan(v) = \sqrt{\frac{6B+2D}{6B}}w$ was used leaving

$$\tan\left(-(t+G)\sqrt{\frac{6B+2D}{6B}}\right) = \sqrt{\frac{6B+2D}{6B}}\cos(\theta). \tag{6.50}$$

This equation will only be valid for a small duration depending on the value of the

constants.

The general case of Equation (6.49) in which the constants do not imply a perfect quartic leads to a elliptic integral of the third kind. The denominator can be reduced in order using a birational substitution [64] of Legendre [53, §2.2].

Theorem 6.5.3. *Birational Substitution is a method of manipulating the degree of a rational polynomial integrand using the substitution $(x, y) \to (u, v)$, it converts a quartic polynomial in x into a cubic polynomial of u,*

$$\int \frac{1}{\sqrt{f(x)}} dx = -\int \frac{1}{\sqrt{f_1(u)}} du.$$

This is done via the substitutions $(x - \alpha) = \frac{1}{u}$ where α is a root of f, and $\frac{y}{(x-\alpha)^2} = v$ where $y^2 = f(x)$. The expression for v allows for a function purely of u and $v^2 = f_1(u)$ completes the substitution.

Proof. Let α denote a root of $f(x)$ and let $f(x) = g(x)(x - \alpha)$. The function $g(x)$ is a cubic polynomial and therefore has roots $\alpha_1, \alpha_2, \alpha_3$. Factorise $g(x)$ with respect to these roots and divide by $(x - \alpha)^3$, for $A \in \mathbb{R}$,

$$\begin{aligned}
\frac{g(x)}{(x-\alpha)^3} &= A\left(\frac{x-\alpha_1}{x-\alpha}\right)\left(\frac{x-\alpha_2}{x-\alpha}\right)\left(\frac{x-\alpha_3}{x-\alpha}\right) \\
&= A\left(\frac{x-\alpha+\alpha-\alpha_1}{x-\alpha}\right)\left(\frac{x-\alpha+\alpha-\alpha_2}{x-\alpha}\right)\left(\frac{x-\alpha+\alpha-\alpha_3}{x-\alpha}\right), \\
&= A(1+(\alpha-\alpha_1)u)(1+(\alpha-\alpha_2)u)(1+(\alpha-\alpha_3)u) \\
&= f_1(u).
\end{aligned}$$

This cubic polynomial in u is defined as $f_1(u)$ because $f_1(u) = v^2 = \frac{y^2}{(x-\alpha)^4} = \frac{g(x)}{(x-\alpha)^3} = \frac{f(x)}{(x-\alpha)^4}$. Lastly the change of variables is given by $dx = -\frac{1}{u^2}du$, and so the integral,

$$\begin{aligned}
\int \frac{1}{\sqrt{f(x)}} dx &= -\int \frac{1}{\sqrt{f_1(u)/u^4}} \frac{du}{u^2}, \\
&= -\int \frac{1}{\sqrt{f_1(u)}} du.
\end{aligned}$$

□

Returning to the problem at hand, the roots of the polynomial in Equation (6.49) can be found using the quadratic formula,

$$w^2 = \frac{6B - 2D}{6B} \pm \frac{\sqrt{4D^2 - 12BC^2}}{-6B} = \alpha_{\pm}.$$

This means $f(x) = -3B(x + \sqrt{\alpha_+})(x - \sqrt{\alpha_+})(x + \sqrt{\alpha_-})(x - \sqrt{\alpha_-})$, from which we can use the first root $\sqrt{\alpha_+}$ and the birational substitution to rewrite the integral as

$$\int \frac{du}{\sqrt{-3B(1 + 2\sqrt{\alpha_+}u)(1 + (\sqrt{\alpha_+} + \sqrt{\alpha_-})u)(1 - (\sqrt{\alpha_+} + \sqrt{\alpha_-})u)}}, \tag{6.51}$$

for $u = \frac{1}{\cos\theta - \sqrt{\alpha_+}}$. This integral takes the form of the inverted incomplete elliptic integral of Weierstrass' $\wp$ function with some scaling involved.

6.6 Consistent solutions on the sphere

From here on, the assumption is made that the density U satisfies the conditions necessary for the existence of solutions to the ODE, given in Proposition 6.2.2. Furthermore, U is also assumed to to be constrained to the sphere.

The discussion turns to whether the solution of the Lagrangian form of the Euler equations determines a pushforward for the initial density which satisfies the continuity equation. Then the solution is considered under a different light — for what duration of time can one be sure that this pushforward map produces Lebesgue measurable probability distributions. This question is explored in Section 6.6.2.

6.6.1 Consistency with the continuity equation

Chapter 4 introduces optimal transport as an approach to solve ODEs such as the Euler system without requiring as strict assumptions on the smoothness of the density as have been made in this chapter. The analysis in this section can be done without much measure theoretic analysis, aside from the definition of the pushforward of a measure.

Lemma 6.6.1. *Consider the divergence operator on the surface of the sphere. For $f \in C_b(\mathbb{S}^2)$, $\rho \in \mathcal{P}_2(\mathbb{S}^2)$ and V taken as the vector field solving Equation* (6.16) *in which*

the solution is constrained to the sphere,

$$\int_{\mathbb{S}^2} \nabla f \cdot (V\rho)\mu(dx) = -\int_{\mathbb{S}^2} f\nabla \cdot (V\rho)\mu(dx), \tag{6.52}$$

where μ is area measure on the sphere.

Proof. This is a specific statement of a wider dual relationship between the gradient and divergence operators. In this case, the statement follows from Stokes' theorem. First note the product rule for the divergence operator, $\nabla \cdot (fV\rho) = \nabla f \cdot (V\rho) + f\nabla \cdot (V\rho)$. Stokes' theorem states that if $d\eta$ is an exact 2-form, then on an orientable manifold Ω

$$\int_\Omega d\eta = \int_{\partial\Omega} \eta. \tag{6.53}$$

The sphere is a compact 2 dimensional manifold, area measure is given by $\mu(d\theta\, d\varphi) = \sin(\theta)d\theta d\varphi$ on the charts specified by longitude and latitude (θ, φ). Thus, if $\nabla \cdot (fV\rho)\sin(\theta)d\theta d\varphi$ is an exact 2-form, then it satisfies Stokes' theorem on the sphere, and the sphere has no boundary. Let $V = (V_\theta, V_\varphi)$ in $(\hat{\theta}, \hat{\varphi})$ coordinates and then the divergence operator is

$$\nabla_{\mathbb{S}^2} \cdot V = \left[\frac{\partial(V_\theta \sin\theta)}{\partial\theta} + \frac{\partial V_\varphi}{\partial\varphi}\right]\frac{1}{\sin\theta}. \tag{6.54}$$

If $\eta = fV_\varphi\rho d\theta - fV_\theta\rho d\varphi$ then it's differential $d\eta = \left[\frac{\partial(fV_\theta\rho\sin\theta)}{\partial\theta} + \frac{\partial(fV_\varphi\rho)}{\partial\varphi}\right] d\theta d\varphi$ is exact and so

$$\int_{\mathbb{S}^2} \nabla_{\mathbb{S}^2} \cdot (fV\rho)\mu(d\theta\, d\varphi) = \int_{\mathbb{S}^2} d\eta = 0. \tag{6.55}$$

Therefore, the expansion of the divergence operator using the product rule implies,

$$\int_{\mathbb{S}^2} \nabla_{\mathbb{S}^2} f \cdot (V\rho)\mu(dx) = -\int_{\mathbb{S}^2} f\nabla_{\mathbb{S}^2} \cdot (V\rho)\mu(dx). \tag{6.56}$$

□

Proposition 6.6.2. *Consider the triplet (X, V, ρ) being a valid solution to the Euler system in Lagrangian form (Equation (6.16)). Then the probability measure defined by the pushforward, $\sigma(x,t) = X(x,t)\#\rho(\cdot, 0)$, is the unique solution to the linear transport*

equation,

$$\frac{\partial}{\partial t}\sigma(X(x,t),t) + \nabla_{\mathbb{S}^2} \cdot (V(x,t)\sigma(x,t)) = 0 \tag{6.57}$$

with initial condition $\sigma(x,0) = \rho(x,0)$.

Proof. Let $\mu(dx)$ denote Lebesgue measure on the surface of the sphere $\mathbb{S}^2$. By Definition 4.1.2, of a pushforward measure,

$$\int_{\mathbb{S}^2} f(X(x,t))\rho(x,0)\mu(dx) = \int_{\mathbb{S}^2} f(y)\sigma(y,t)\mu(dy).$$

Taking the derivatve of this equation with respect to time, the derivative commutes with the integral due to the definition of ρ as a bounded Lebesgue measurable function.

$$\int \nabla_{\mathbb{S}^2} f(X(x,t)) \cdot V(x,t)\rho(x,0)\mu(dx) = \int f(y)\frac{\partial}{\partial t}\sigma(y,t)\mu(dy),$$

Recall the definition of the Lagrangian velocity $V = \mathbf{u} \circ X$ and thus one can take the pushforward of the left hand side of the equation,

$$\int \nabla_{\mathbb{S}^2} f(y,t) \cdot \mathbf{u}(y,t)\sigma(y,t)\mu(dy) = \int f(y)\frac{\partial}{\partial t}\sigma(y,t)\mu(dy).$$

Lemma 6.6.1 then implies that,

$$-\int f(y,t)\ \nabla_{\mathbb{S}^2} \cdot (\mathbf{u}(y,t)\sigma(y,t))\,\mu(dy) = \int f(y)\frac{\partial}{\partial t}\sigma(y,t)\mu(dy).$$

This holds for any $f \in C_b(\mathbb{R}^n)$ and therefore, the following ODE is satisfied weakly by $\sigma(y,t)$

$$-\frac{\partial}{\partial t}\sigma(y,t) + \nabla_{\mathbb{S}^2} \cdot (\mathbf{u}(y,t)\sigma(y,t)) = 0.$$

□

6.6.2 Estimates for the interval of validity

Theorem 6.2.1 establishes the existence of solutions solving the Lagrangian form of the ODE for time invariant $\rho(x,t)$, and so there exists a unique pair $X(x,t), V(x,t)$

for which $\sigma = X\#\rho(x,0)$ gives a weak solution to the continuity equation. The question remaining is, for how long is X invertible? When X is no longer invertible the assumption that the pushforward measure is in $\mathcal{P}_2(\mathbb{R}^n)$ is in jeopardy.

Lemma 6.6.3. *[38] A square matrix $A \in M_n(\mathbb{R}^n)$ defines an invertible linear map $A : \mathbb{R}^n \to \mathbb{R}^n$ if and only if Ker$(A) = \mathbf{0}$.*

Proof. For the contrapositive, consider $b \neq 0$ a solution to $Ax = 0$. Then the equation $Ax = c$ will not have a unique solution for each $c \in \mathbb{R}^n$, as if $Ad = c$ for some $d \in \mathbb{R}^n$, then $A(d + b) = c$ as well, and this implies the inverse map $A^{-1}(c)$ of c is not well defined, thus A isn't invertible.

An invertible map is injective and surjective. By the first Isomorphism theorem ($\dim\mathbb{R}^n = \mathrm{Ker}(A) + \mathrm{Im}(A)$), for the linear map to be surjective the kernel must have dimension 0, so it must be a singular point, and zero is never not inside the kernel of a linear map. □

Remark 6.6.4. Consider the sequence $\sum_{n=0}^{\infty}(A - I)x$ where $A \in M_n(\mathbb{R})$ and $x \in \mathbb{R}^n$. This sequence converges if $\|A - I\|_{op} < 1$. Furthermore the sequence is the geometric sum of $(I - (A - I))^{-1} = A^{-1}$, implying that where the series converges the inverse to A exists [37, Thm. 2.6.2]. Hence A is invertible if $\|A - I\|_{op} < 1$.

Lemma 6.6.5. *If the Jacobian $J_x(X)$ is invertible, and $\rho \in \mathcal{P}_2$ then $X\#\rho \in \mathcal{P}_2$.*

Proof. By the inverse function theorem, an invertible Jacobian implies that X is bijective in a neighbourhood. We can specify the measure using the definition of a pushforward, as if X is invertible then the preimage $X^{-1}(A) = \{X^{-1}(x) : x \in A\}$. Then as $\rho \in \mathcal{P}_2$ it can be represented by its density $\rho(A) = \int_A \rho(x)dx$ and so can $X\#\rho(A) = \int_A \rho(X^{-1}(x))dx$. □

Proposition 6.6.6. *An estimate for the length of time for which $X(x,t)$ is invertible is,*

$$t < \frac{1}{2L} \tag{6.58}$$

where L is the Lipschitz constant for $u(X,t)$.

Proof. By definition,

$$X(x,t) = \int_0^t V(x,s)ds + X(x,0),$$

where we take $X(x,0) = x$ without loss of generality. Consider now the Jacobian matrix of X, recalling that $x \in \mathbb{R}^3$.

$$\begin{aligned}\frac{\partial}{\partial x}X(x,t) &= \int_0^t \frac{\partial u(X(x,s),s)}{\partial X}\frac{\partial X(x,s)}{\partial x}ds + I \\ \frac{\partial}{\partial x}X(x,t) - I &= \int_0^t \frac{\partial u(X(x,s),s)}{\partial X}\left(\frac{\partial X(x,s)}{\partial x} - I\right)ds \\ &\quad + \int_0^t \frac{\partial u(X(x,s),s)}{\partial X}\frac{\partial X(x,s)}{\partial x}ds\end{aligned}$$

The Lipschitz constant for $u(X(x,t),t)$ is L,

$$\left\|\int_0^t \frac{\partial u(X(x,s),s)}{\partial X}\frac{\partial X(x,s)}{\partial x}ds\right\| \leq Lt$$

and if $f(t) = \sup_x \left\|\frac{\partial X(x,t)}{\partial x} - I\right\|$ then by Gronwall's inequality and the above estimate,

$$\begin{aligned}f(t) &\leq Lt + \int_0^t Lf(s)ds \\ f(t) &\leq Lt + \int_0^t L\exp\left(\int_0^t Lds\right)Lsds \\ &= Lt + \int_0^t L^2 se^{Ls}ds \\ &= Lt + \left[Lse^{Ls}\right]_0^t - \int Le^{Ls}ds \\ &= Lt + (Lt-1)e^{Lt} + 1.\end{aligned}$$

Then, the Jacobian $J_x(X) = \frac{\partial X(x,t)}{\partial x}$ is invertible if $f(t) = \|J_x(X) - I\|_{op} < 1$ by Remark 6.6.4. This condition is satisfied when,

$$\begin{aligned}Lt + (Lt-1)e^{Lt} + 1 &< 1 \\ \frac{Lt}{1-Lt} &< e^{Lt}\end{aligned}$$

which computationally comes out as $0 < Lt < 0.659$ to three significant figures, which is larger than one half. □

Chapter 7

The dam break problem

The dam break problem is a fluid mechanics problem in which the initial data is reminiscent of a dam on a river. The Saint-Venant equations model the leading edge of a reservoir of still water on a flat surface under the effect of gravity, when the barrier holding the water in is removed. Closed form solutions to this problem are known, making this a useful example to explore numerically.

7.1 One dimensional Euler equations

The Euler equations in one dimension are given by

$$\partial_t \rho + \partial_x(\rho u) = 0, \tag{7.1}$$

$$\partial_t(\rho u) + \partial_x(\rho u^2) + \partial_x P(\rho) = 0. \tag{7.2}$$

As derived in Section 6.1.1, the second Euler equation can be expressed in terms of internal energy by,

$$\partial_t u + u\partial_x u + \partial_x U'(\rho) = 0.$$

As discussed in n dimensions in Equation (6.7). In one dimension, the Euler equations form the canonical equations of motion for the Hamiltonian

$$H(\rho, q) = \frac{1}{2}\int \rho \left(\frac{\partial q}{\partial x}\right)^2 dx + \frac{\kappa}{\gamma(\gamma-1)}\int \rho^\gamma dx, \tag{7.3}$$

in which the potential energy has the form $U(\rho) = \frac{\kappa}{\gamma(\gamma-1)}\rho^{\gamma}$.

Lemma 7.1.1. *The Euler equations (Equations* (7.1) *and* (7.2)*) are the extremals of the Hamiltonian functional given in Equation* (7.3)*, also known as Hamilton's canonical equations of motion.*

Proof. The canonical equations of motion govern the dynamics of the Hamiltonian system, therefore recall the canonical equations of motion are:

$$\frac{\delta H}{\partial q} = \frac{\partial \rho}{\partial t}, \quad \text{and} \quad \frac{\delta H}{\partial \rho} = -\frac{\partial q}{\partial t}. \tag{7.4}$$

The integrand of a variation is the functional derivative, take the variation of H around q using a perturbation δq,

$$\begin{aligned}
\delta H &= \lim_{h\to 0} \frac{1}{h}\left(H(\rho, q + h\delta q) - H(\rho, q)\right) \\
&= \lim_{h\to 0} \frac{1}{2h} \int \rho \left(\frac{\partial}{\partial x}\left(q + h\delta q\right)\right)^2 - \left(\frac{\partial q}{\partial x}\right)^2 \Bigg) dx \\
&= \frac{1}{h}\int \rho h \frac{\partial q}{\partial x}\frac{\partial \delta q}{\partial x} dx \\
&= -\frac{1}{2}\int \frac{\partial}{\partial x}\left(\rho \frac{\partial q}{\partial x}\right)\delta q dx.
\end{aligned}$$

The pertubation f is assumed to be zero at the boundary conditions, hence the first term of the integration by parts is zero. Hence,

$$\frac{\delta H}{\partial q} = -\frac{1}{2}\frac{\partial}{\partial x}\left(\rho \frac{\partial q}{\partial x}\right).$$

By the canonical equation of motion, and $u = \partial_x q$ this establishes Equation (7.1). More simply than the first Equation, the functional derivative with respect to ρ is,

$$\frac{\delta H}{\partial \rho} = \frac{1}{2}\left(\frac{\partial q}{\partial x}\right)^2 + \frac{\kappa}{\gamma}\rho^{\gamma-1}.$$

Let $u = \partial_x q$ and take the partial derivative of the above with respect to x,

$$-\frac{\partial}{\partial x}\frac{\partial q}{\partial t} = \frac{\partial q}{\partial x}\frac{\partial^2 q}{\partial x^2} + \frac{\kappa}{\gamma}\frac{\partial}{\partial x}\rho^{\gamma-1},$$
$$-\partial_t u = u\partial_x u + \partial_x U'(\rho).$$

This establishes Equation (7.2) by its alternate form. □

7.2 The dam break problem

The dam break problem is a well known fluid mechanics problem in one dimension. The problem envisions a motionless lake of water behind a dam. The dam is assumed to have symmetry in the lateral direction to the original flow of the water, allowing that dimension to be ignored. The vertical height of the water is modelled not as a dimension but as a graph, of which the solution $(x, h(x,t))$ maps, the function $h(x,t)$ being the height of the water relative to $x = 0$ the position of the dam. The dynamics of the problem are as follows, at $t = 0$ the dam is removed, the water then flows under the effect of its internal gravitational potential energy. The motion of the fluid is given by the Saint-Venant equations [20].

Definition 7.2.1 (Dressler's form)**.** The Saint Venant equations as given by Dressler are,

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + 2c\frac{\partial c}{\partial x} = 0, \tag{7.5}$$
$$c\frac{\partial u}{\partial x} + 2\frac{\partial c}{\partial t} + 2u\frac{\partial c}{\partial x} = 0. \tag{7.6}$$

where $c = \sqrt{gh}$ and g is the gravitational constant, while h is the height of the water.

It is convenient to work with two forms of the Saint–Venant equations, one -taken from fluid mechanics- which illuminates the relationship with the Euler equations more clearly, and *Dressler's* form, which better produces the Riemann invariants.

Lemma 7.2.2. *The fluid mechanic's formulation of the Saint-Venant equations, given*

by Ref [15],

$$\partial_t h + \partial_x(hu) = 0, \tag{7.7}$$
$$\partial_t u + u\partial_x u + g\partial_x h = 0. \tag{7.8}$$

are equivalent to Dressler's form in Definition 7.2.1.

Proof. Using the change of variables $c^2 = hg$ one can verify,

$$\begin{aligned}
&\partial_t h + \partial_x(hu) = 0,\\
&\partial_t c^2 + \partial_x(c^2 u) = 0,\\
&2c\partial_t c + 2cu\partial_x c + c^2\partial_x u = 0,\\
&2\partial_t c + 2u\partial_x c + c\partial_x u = 0.
\end{aligned}$$

And

$$\begin{aligned}
&\partial_t u + u\partial_x u + g\partial_x h = 0,\\
&\partial_t u + u\partial_x u + 2c\partial_x c = 0.
\end{aligned}$$

□

The Euler equations in one dimension can describe the motion of the fluid in the dam break problem, with an internal energy of $U(\rho) = \frac{1}{2}g\rho^2$. In one dimension, the density of the fluid is the "mass of water within a unit interval of length". For a fixed density of fluid, the idea of the "mass of fluid within a unit interval" can be described by an equivalent variable, the height of the fluid in that interval. Hence the relation $\rho = h$ allows the 1 dimensional Euler equations to describe the motion of a 2 dimensional incompressible fluid in the same way that the Saint-Venant equations do.

Lemma 7.2.3. *The Euler equations are equivalent to the Saint-Venant equations when* $\gamma = 2$ *and the constant* $\kappa = g$.

Proof. As mentioned previously the change of variable $\rho = h$ is justified, with that change of variables the momentum conservation (Equation (7.1) and Equation (7.7))

are the same for Saint Venant and one dimensional Euler. Substituting $U(\rho) = \frac{1}{2}g\rho^2$ into the second Euler equation,

$$\begin{aligned}\partial_t u + u\partial_x u + \partial_x U'(\rho) = 0,\\ \partial_t u + u\partial_x u + g\partial_x \rho = 0,\end{aligned}$$

gives the second Saint-Venant equation (Equation (7.8)) again with $\rho = h$. □

The Euler equations are also equivalent to the alternate form of the Saint Venant equations due to Dressler via the substitution $c = \sqrt{g\rho}$.

Proof. To show equivalency between the two continuity equations substitute $c = \sqrt{g\rho}$ into Equation (7.1) and apply the product rule,

$$\begin{aligned}\frac{1}{g}\frac{\partial}{\partial t}c^2 + \frac{1}{g}\frac{\partial}{\partial x}(c^2 u) = 0,\\ \frac{2}{g}c\frac{\partial c}{\partial t} + \frac{1}{g}uc\frac{\partial c}{\partial x} + \frac{1}{g}c\frac{\partial}{\partial x}(uc) = 0,\\ 2\frac{\partial c}{\partial t} + 2u\frac{\partial c}{\partial x} + \frac{\partial u}{\partial x} = 0.\end{aligned}$$

Which is Equation (7.5).

To prove the equivalency of the momentum conservation equations the same approach is applied to Equation (7.2), in which Equation (7.5) is used to cancel some terms,

$$\begin{aligned}\frac{1}{g}\frac{\partial}{\partial t}(c^2 u) + \frac{1}{g}\frac{\partial}{\partial t}(c^2 u^2) + \frac{\partial}{\partial x}P(\rho) = 0,\\ 2cu\partial_t c + c^2\partial_t u + 2c^2 u\partial_x u + 2cu^2\partial_x c + g\partial_x P(\rho) = 0,\\ cu\left(u\partial_t c + c\partial_x u + 2u\partial_x c\right) + c^2\partial_t u + c^2 u\partial_x u + \frac{1}{2}\partial_x c^4 = 0,\\ \partial_t u + u\partial_x u + 2c\partial_x c = 0.\end{aligned}$$

with the assumption that we want solutions on the support of c. □

Euler equations for compressible gas	**Saint Venant equations for water in a reservoir**
ρ - pressure of the gas	h - height of the water
(ρ, u)	(h, u)
$\partial_t \rho + \partial_x(\rho u) = 0$ $\partial_t u + u\partial_x u + g\partial_x \rho = 0$	$\partial_t h + \partial_x(hu) = 0$ $\partial_t u + u\partial_x u + g\partial_x h = 0$

7.3 Characteristic curves, Riemann Invariants and the Ritter solution

Further examination of the Saint-Venant equations (and by extension the Euler equations) can be done by looking at the characteristic curves of the systems of PDEs, and using this formulation to find the Riemann Invariants. The Riemann Invariants are functions which are constant along the characteristic curves of the system. If $\gamma(x(t), t)$ is a characteristic curve (a parameterised curve which is a solution to the system) then it has derivative,

$$\frac{d}{dt}\gamma(x(t), t) = \frac{\partial}{\partial t}\gamma(x(t), t) + \frac{dx}{dt}\frac{\partial}{\partial x}\gamma(x(t), t). \tag{7.9}$$

Translating the Saint-Venant equations into this form gives

$$\frac{\partial}{\partial t}\begin{pmatrix} h(x(t), t) \\ u(x(t), t) \end{pmatrix} + \begin{pmatrix} u & h \\ g & u \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} h(x(t), t) \\ u(x(t), t) \end{pmatrix} = 0, \tag{7.10}$$

and thus for $\gamma(x(t), t)$ to be a characteristic curve of this system it must satisfy,

$$\partial_t \gamma(x(t), t) + \lambda \partial_x \gamma(x(t), t) = 0, \tag{7.11}$$

where λ is an eigenvalue of the matrix given in Equation (7.10). In addition to this constraint, γ must also be an eigenvector. The eigenvalues are $\lambda = u \pm \sqrt{gh} = u \pm c$. However the eigenvectors are

$$\begin{pmatrix} \pm\frac{c}{g} \\ 1 \end{pmatrix} \tag{7.12}$$

and due to the fact these are not constant, finding the Riemann invariants is more difficult. It may be possible to express the equations in a simpler form, namely in terms of exclusively c not h.

$$\frac{\partial}{\partial t}\begin{pmatrix} c(x(t),t)^2/g \\ u(x(t),t)\end{pmatrix} + \begin{pmatrix} u & h \\ g & u\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} c(x(t),t)^2/g \\ u(x(t),t)\end{pmatrix} = 0,$$
$$\begin{pmatrix} 2c/g & 0 \\ 0 & 1\end{pmatrix}\frac{\partial}{\partial t}\begin{pmatrix} c(x(t),t) \\ u(x(t),t)\end{pmatrix} + \begin{pmatrix} u & c^2/g \\ c^2/h & u\end{pmatrix}\begin{pmatrix} 2c/g & 0 \\ 0 & 1\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} c(x(t),t) \\ u(x(t),t)\end{pmatrix} = 0,$$
$$\frac{\partial}{\partial t}\begin{pmatrix} c(x(t),t) \\ u(x(t),t)\end{pmatrix} + \begin{pmatrix} u & 2c \\ c/2 & u\end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} c(x(t),t) \\ u(x(t),t)\end{pmatrix} = 0.$$

This new form of the equation has the same eigenvalues but the eigenvectors are now $[\pm 1, 2]^\top$. To find the Riemann invariants in this case, the search is for a function $R(x(t),t)$ such that $\partial_t R + \lambda \partial_x R = 0$. When $\lambda = u \pm c$ the functions $R(x(t),t) = u \pm 2c$ satisfy the equation, simply add Equation (7.5) and Equation (7.6) or subtract them depending on the eigenvalue. Thus the Riemann invariants are $(u \pm 2c)$, and

$$\frac{d}{dx}(u \pm 2c) = 0. \tag{7.13}$$

7.3.1 The Ritter solution

The Ritter solution to the Saint Venant equations is

$$u = \frac{2}{3}\left(\frac{x}{t} + c_0\right), \tag{7.14}$$
$$h = \frac{1}{9g}\left(2c_0 - \frac{x}{t}\right)^2. \tag{7.15}$$

The constant c_0 is defined to be $c_0 = \sqrt{gh_0}$ where h_0 is the initial height of the reservoir, and again g is the gravitational constant.

7.3.2 The characteristic curves

The system of PDEs known as the Saint–Venant equations has been reduced to the ODEs

$$\frac{d}{dx}(u \pm 2c) = 0, \qquad \frac{d}{dt}x(t) = u \pm c, \tag{7.16}$$

for the Riemann invariants and eigenvalues respectively.

Lemma 7.3.1. *The characteristic curves of the system are,*

$$\gamma_+ : x = mt, \qquad \gamma_- : x = 2t - 3\alpha t^{1/3}. \tag{7.17}$$

Proof. The first characteristic curve γ_+ is known from the form of the equations, it can also be seen from the Ritter solution. If $u = 2(x/t + 1)/3$ and $c = (2 - x/t)/3$ then the Riemann invariants are

$$u + 2c = 2, \qquad \text{and} \qquad u - 2c = \frac{4}{3}\frac{x}{t} - \frac{2}{3}. \tag{7.18}$$

the latter of which can only be constant if there exists a constant m such that $x/t = m$. The second characteristic curve, γ_-, is deduced from the first curve, γ_+, and the Riemann invariant condition in Equation (7.16).

$$\begin{aligned}
\frac{dx}{dt} &= u + c, \\
\frac{d}{dt}mt &= \frac{4}{3} + \frac{1}{3}\frac{x}{t}, \\
t\frac{dm}{dt} &= \frac{4}{3} - \frac{2m}{3}, \\
\int \frac{1}{(\frac{4}{3} - \frac{2}{3}m)}dm &= \int \frac{1}{t}dt, \\
-\frac{3}{2}\log\left(\frac{2}{3}m - \frac{4}{3}\right) &= \log(t) + \log(\alpha), \\
2m - 4 &= 3\alpha t^{\frac{2}{3}}, \\
x &= 2t + \frac{3\alpha}{2}t^{\frac{1}{3}}.
\end{aligned}$$

□

Chapter 8

Gibbs measure

In the following three chapters, the periodic nonlinear Schrödinger equation is discussed from the viewpoint of statistical mechanics. Instead discussing classical solutions for certain types of initial condition, the theory in this field is interested in weak solutions, or the *distribution* of *typical* solutions.

Why is this important? For nonlinear PDE, existence theorems for classical solutions to initial value problems may introduce hypothesis on the initial data that are very stringent or unrealistic. Further, even when the intial data is smooth, the solution to the intial value problem may not be smooth. Systems such as the Euler equations can form shocks (nondifferentiable, discontinuous solutions) in finite time [26]. To allow for functions which are discontinuous and nondifferentiable to solve the PDE one needs to deal with weak solutions.

One philosophy for understanding weak solutions is to instead conceptualise them as a random process. If a Cauchy problem is well posed on a set of initial data, then instead of working on classes of initial data and accompanying closed form solutions, one can discuss a *typical* solution. The set of all initial data forms a probability space, and at each time the random process documents how the distribution of the solutions (at said time) changes. This is especially useful in real systems where the intial conditions observed will not be exact. A density measurement of the atmosphere will return a range of values specified by the precision of the instrument. In a paper on the evolution of a measure on $L^2(\mathbb{T})$ under the dynamics of the NLSE, Lebowitz, Rose and Speer outline the philosophy. "Instead of trying to solve the initial value problem for a system

containing a very large (say 10^{23}) number of particles, which is clearly an impossible task even in principle, we obtain information about values of macroscopic observables by taking averages over Gibbs probability distributions containing only a few parameters (particle density, temperature, etc.). While the rigorous justification of the theory is still not fully understood, its success leaves no doubt about its utility. In fact, the results obtained from a suitable probability measure, which includes information about both typical behavior and fluctuations, are generally more relevant than the solution of a specific initial value problem for understanding the behavior of real systems."

– Lebowitz, Rose or Speer [48]

In the case of the the nonlinear Schrödinger equation, the Cauchy problem,

$$\begin{aligned} \frac{1}{i}\frac{\partial \psi}{\partial t} &= \frac{\partial^2 \psi}{\partial x^2} + \beta|\psi|^2\psi \\ \psi_0 &= \phi(x) \in H^1(\mathbb{T}, \mathbb{C}), \end{aligned}$$

is well posed [9, Thm. 1] for any initial data in H^1. Any solution is continuous in t and at each t is a function of x in $H^1(\mathbb{R}, \mathbb{C})$. To discuss typical solutions a measure is needed and there exists a measure describing this system which is invariant with respect to time. Within statistical mechanics the Hamiltonian describes the energy of a system. Louiville's theorem then says that on the phase space, $L^2(\mathbb{T}) \times L^2(\mathbb{T})$, there exists a measure which is invariant under the Hamiltonian flow. The Gibbs measure is this measure in the context of infinite dimensional phase space, and this chapter centres on defining the Gibbs measure.

8.1 Gaussian measure on the cylinder sets

Consider $L^2(\mathbb{T})$, and a finite dimensional linear subspace $C \subset L^2(\mathbb{T})$, where C is spanned by the collection $\{(\exp(in_j x) \ : n_j \in N \subset \mathbb{Z} \setminus \{0\}\}$ and N is ordered and contains m elements. The subspace C can be identified with the orthogonal projection onto $\mathbb{R}^m$ by the function ψ_C. If $(\exp(in_j x)_{j=1}^m$ is an orthonormal basis for C then $\psi_C : L^2(\mathbb{T}) \to \mathbb{R}^m; f \mapsto (\langle f, e_n \rangle_{L^2(\mathbb{T})})_{j=1}^m$ is a projection of f onto $\mathbb{R}^m$.

Definition 8.1.1. For ψ_C as defined above, let $A = \times_{j=1}^m A_j \subset \mathbb{R}^m$ be a Borel set in

$\mathbb{R}^m$, (so $A_j \in \mathcal{B}(\mathbb{R}), \forall j$). A cylinder set of $L^2(\mathbb{T})$ is the pre-image of A by ψ_C, [46],

$$\psi_C^{-1}(A) = \{f \in L^2(\mathbb{T}) \,:\, \psi_C(f) \in A\}.$$

Lemma 8.1.2. *The collection of cylinder sets form an algebra.*

Proof. **$L^2(\mathbb{T})$ is a cylinder set** For any C of dimension m, take $A = \mathbb{R}^m$. $\psi_C^{-1}(A) = L^2(\mathbb{T})$.

Compliments of cylinder sets are cylinder sets Consider the cylinder $\psi_C^{-1}(A)$, it has a compliment $(\psi_C^{-1}(A))^c = L^2(\mathbb{T}) \setminus \psi_C^{-1}(A)$.

$$\begin{aligned} L^2(\mathbb{T}) \setminus \psi_C^{-1}(A) &= \{f \in L^2(\mathbb{T}) : \psi_C(f) \notin A\}, \\ &= \{f \in L^2(\mathbb{T}) : \psi_C(f) \in A^c\}, \end{aligned}$$

And as $A \in \mathcal{B}(\mathbb{R})^m$, so is A^c and therefore $(\psi_C^{-1}(A))^c$ is a cylinder set.

Closed under finite intersections Take C_1, C_2 to be n and m dimensional subspaces with basis $(f_i)_{i=1}^n$ and $(l_i)_{i=1}^m$ respectively. And likewise A_1, A_2 are rectangles in $\mathbb{R}^n$ and $\mathbb{R}^m$.

$$\psi_{C_1}^{-1}(A_1) \cap \psi_{C_2}^{-1}(A_2) = \{f \in L^2(\mathbb{T}) : (\langle f, f_1\rangle, ..., \langle f, f_n\rangle, \langle f, l_1\rangle, ..., \langle f, l_m\rangle) \in A_1 \times A_2\}.$$

This looks like a cylinder, however, unless C_1 and C_2 have only intersect at $\mathbf{0}$, this formula can be reduced. Let $(e_i)_{i=1}^r$ be a basis of $C_1 \cap C_2$, $(g_i)_{i=1}^p$ of $C_1 \cap C_2^\perp$ and $(h_i)_{i=1}^q$ of $C_1^\perp \cap C_2$ and then $n + m = 2r + p + q$. Then $\text{span}\{f_i\} = \text{span}\{e_j, g_k : j = 1, ..., r.\ k = 1, ..., p.\}$ and $\text{span}\{l_i\} = \text{span}\{e_j, h_k : j = 1, ..., r.\ k = 1, ..., q.\}$, therefore A_1 and A_2 can be expressed in coordinates with respect to bases $\{e_j, g_k : j = 1, ..., r.\ k = 1, ..., p.\}$ and $\{e_j, h_k : j = 1, ..., r.\ k = 1, ..., q.\}$ respectively. One can consider them to be subspaces of $\mathbb{R}^{r+p+q}$ and let B be their intersection. The intersection of two rectangles is a rectangle and thus,

$$\begin{aligned} &\psi_{C_1}^{-1}(A_1) \cap \psi_{C_2}^{-1}(A_2) = \\ &\quad \{f \in L^2(\mathbb{T}) : (\langle f, e_1\rangle, ..., \langle f, e_r\rangle, \langle f, g_1\rangle, ..., \langle f, g_p\rangle, \langle f, h_1\rangle, ..., \langle f, h_q\rangle) \in B\}, \end{aligned}$$

is a cylinder set.

□

Before defining a cylindrical measure, the domain of the function ψ_C can be defined as the quotient space $L^2(\mathbb{T})/C_1$, where the equivalence class is given by $f \sim y \iff \langle f, e_i\rangle = \langle g, e_i\rangle \forall i$. Then, for $n = \dim(C_1) < \dim(C_2) = m$ the projection,

$$\pi_{C_2,C_1} : L^2/C_2 \to L^2/C_1, \tag{8.1}$$

$$: \sum_{i=1}^{m} \langle f, e_i\rangle e_i \mapsto \sum_{i=1}^{n} \langle f, e_i\rangle e_i. \tag{8.2}$$

Now note that, if A is again Borel and defined as $A = \times_{j=1}^{n} A_j \subset \mathbb{R}^n$, then $\pi^{-1}_{C_2,C_1}(\psi^{-1}_{C_1}(A)) = \psi^{-1}_{C_2}((\times A_i)_{i=1}^{n} \times \mathbb{R}^{m-n-1})$.

Definition 8.1.3 ([63] p.172)**.** A *cylindrical measure* is a finitely additive measure ν on the cylinders S of a Hilbert space H, formed by a family of measures μ_{C_i} on the subspaces spanned by finitely many basis vectors. Let S_i denote a cylinder $S_i = \psi^{-1}_{C_i}(A)$ for some $A \in \mathcal{B}(\mathbb{R}^m)$, assume $m = \dim(C_1) < \dim(C_2)$. The cylindrical measure is then defined $\nu(S_1) = \mu_{C_1}(\psi^{-1}_{C_1}(A))$, and the projection map π_{C_2,C_1} must compose with ψ_{C_1} so that $\mu_{C_2}(\pi^{-1}_{C_2,C_1}(\psi^{-1}_{C_1}(A))) = \mu_{C_1}(\psi^{-1}_{C_1}(A))$ implying that the measure of a cylinder $\nu(S_i) = \psi_{C_i}(A)$ does not depend on the choice of base A or generator subset C_i (as there are many base and generator set combinations which produce the same cylinder).

Definition 8.1.4. Gaussian measure on $\mathbb{R}^m$ is a well known probability measure on the Borel sets $A \in \mathcal{B}(\mathbb{R}^m)$ given by,

$$\mu^m(A) = \frac{1}{\sqrt{2\pi}^m} \int_{A_1} ... \int_{A_m} \exp\left(-\frac{1}{2}\sum_{j=1}^{m} x_j^2\right) \prod_{j=1}^{m} dx_j. \tag{8.3}$$

Proposition 8.1.5. *The cylindrical measure $\mathcal{V}$ on $L^2(\mathbb{T})$ is defined by the family of Gaussian measures μ_C on the cylinder sets. Let C be m dimensional and spanned by the basis $\{e^{in_1 x}, e^{in_2 x}, ..., e^{in_m x}\}$, a subset of $\{e^{inx} : n \in \mathbb{N}\}$ which is a basis of $L^2(\mathbb{T})$. The function $f = \sum_{n=-\infty}^{\infty} a_n e^{inx}$ is in the cylinder set $\psi_C^{-1}(A)$ if $a_{n_k} \in A_k$ for $k = 1, ..., m$. The pushforward of the measure of this cylinder set by the map ψ_C is simply Gaussian*

measure on $\mathbb{R}^m$,

$$(\mu_C \# \psi_C)(A) = \mu_C(\psi_C^{-1}(A)) = \mu^m(A) = \frac{1}{\sqrt{2\pi}^m} \int_{A_1} ... \int_{A_m} \exp\left(-\frac{1}{2}\sum_{j=1}^{m} x_j^2\right) \prod_{j=1}^{m} dx_j. \tag{8.4}$$

Proof. Let C_1 be a linear subspace of $L^2(\mathbb{T})$ spanned by a collection of Fourier modes $(e_k)_{k=1}^n$ where k is a subindex of $j \in \mathbb{Z} \setminus \{0\}$ and $e_k = e^{ijkx}$. One can expand the collection of basis vectors $(e_k)_{k=1}^n$ so that the new collection $(e_k)_{k=1}^m$ spans the second linear subspace C_2 of dimension $m > n$. Next consider a *base* space $A = \times_{j=1}^n A_j \subset \mathbb{R}^n$ where each $A_j \in \mathcal{B}(\mathbb{R})$. Now consider ψ_{C_i} as functions from the quotient space L^2/C_i and the projection map π_{C_1,C_2} as defined in Equation (8.1). To prove μ_C is cylindrical, it must be established that $\mu_{C_1}(\psi_{C_1}^{-1}(A)) = \mu_{C_2}(\pi_{C_2,C_1}^{-1}(\psi_{C_1}^{-1}(A)))$, to this end,

$$\begin{aligned}
\mu_{C_2}(\pi_{C_2,C_1}^{-1}(\psi_{C_1}^{-1}(A))) &= \mu_{C_2}\left(\psi_{C_2}^{-1}((\times A_i)_{i=1}^n \times \mathbb{R}^{m-n-1})\right) \\
&= \mu^m(\times A_i)_{i=1}^n \times \mathbb{R}^{m-n-1}) \\
&= \mu^n(A), \\
&= \mu_{C_1}(\psi_{C_1}^{-1}(A))
\end{aligned}$$

□

Remark 8.1.6. The cylindrical measure is only finitely additive.

Proof. The cylinder sets form an algebra but not a σ-algebra. Assume that the cylindrical measure is σ-additive and consider the set [46, p.55],

$$B_n := \{f \in L^2(\mathbb{T}) \,:\, |\langle f, e^{ijx}\rangle \leq n,\ j = 1, .., a_n\}. \tag{8.5}$$

The countable union $\cup_{n=1}^{\infty} B_n$ is equal to $L^2(\mathbb{T})$. The measure of $\mathcal{V}(L^2(\mathbb{T})) = \mu_C(\psi_C^{-1}(\mathbb{R})) = \mu^1(\mathbb{R}) = 1$ where C_1 is the span of any basis vector $\exp(ijx)$, $j \in \mathbb{Z} \setminus \{0\}$. However, the measure of B_n is

$$\begin{aligned}
\mathcal{V}(B_n) &= \frac{1}{\sqrt{2\pi}^{a_n}} \int_{-n}^{n} ... \int_{-n}^{n} \exp\left(-\frac{1}{2}\sum_{j=1}^{a_n} x_j^2\right) \prod_{j=1}^{a_n} dx_j, \\
&= \left(\frac{1}{\sqrt{2\pi}} \int_{-n}^{n} \exp\left(-\frac{x^2}{2}\right) dx\right)^{a_n}.
\end{aligned}$$

Each integral is smaller than 1 and so the value of a_n can be chosen so that the value of $\mathcal{V}(B_n) < \frac{1}{2^{n+1}}$. Choose this a_n to be the sequence in the definition of B_n for each n as $n \to \infty$. Then the sequence

$$\sum_{n=1}^{\infty} \mathcal{V}(B_n) < \frac{1}{2},$$

thus, on the non-disjoint sequence of cylinders B_n, if $\mathcal{V}$ was σ-additive then it would be subadditive. Instead, on this sequence the measure is superadditive in the sense that,

$$\sum_{n=1}^{\infty} \mathcal{V}(B_n) < \mathcal{V}\left(\cup_{n=1}^{\infty} B_n\right).$$

□

8.2 Radonification and Wiener loop measure

Having established the Gaussian measure as a finitely additive measure on the cylinder sets of $L^2(\mathbb{T})$, we can look to apply Radonification by the following theorem of Sazonov.

Definition 8.2.1 (Hilbert Schmidt operator)**.** The operator $T : H \to H$ is Hilbert-Schmidt if and only if

$$\sum_{n=1}^{\infty} \|T(e_n)\|^2 < \infty.$$

Theorem 8.2.1 (Sazonov)**.** *[63, p.215] Let H be a Hilbert space, $T : H \to H$ be a Hilbert-Schmidt operator, and μ be a cylindrical measure concentrated on the balls of H. Then the pushforward of μ by T is a Radon measure on H.*

Lemma 8.2.2. *The linear operator $u : L^2(\mathbb{T}) \to L^2(\mathbb{T}); e^{inx} \mapsto \frac{1}{n} e^{inx}$ is Hilbert-Schmidt.*

Proof. The norm,

$$\|u(e^{inx})\|^2 = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{n^2} |e^{inx}|^2 dx = \frac{1}{n^2}.$$

Then the series,

$$\sum_{n=1}^{\infty} \|u(e^{inx})\|^2 = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

□

Definition 8.2.3. A cylindrical measure μ_C on the Hilbert space H is scalarly concentrated up to δ on the subsets $B \subset H$ if, given a $\delta \in [0,1]$, for every one dimensional linear subspace C of H $\mu_C(\{\psi_C(f)|f \in B\}) \geq 1-\delta$.

Proposition 8.2.4. *The Cylindrical measure μ_C given in Equation* (8.4) *is scalarly concentrated on the balls of $L^2(\mathbb{T})$.*

Proof. The ball of radius r in $L^2(\mathbb{T})$ which is centred at zero is denoted $B_r(0)$. The preimage of the projection of this ball on to the 1 dimensional subspace spanned by e_i gives $\{f \in L^2(\mathbb{T}) : |\langle f, e_i\rangle| \leq r\}$. The measure of this ball is

$$\mu_C(\psi_C^{-1}(B_r(0)) = \frac{1}{\sqrt{2\pi}} \int_{-r}^{r} e^{-\frac{x^2}{2}} dx,$$

for any one dimensional subspace C. This is just one dimensional Gaussian measure. If we take the definition of $\text{erf}(r) := \frac{1}{\sqrt{2\pi}} \int_{-r}^{r} e^{-\frac{x^2}{2}} dx$ then note that the function is monotone increasing with codomain $[0,1]$. For any $\delta \in (0,1)$, there exists a $R \in \mathbb{R}$ such that $\text{erf}(R) = 1-\delta$ and then for all $r > R$, $\mu_C(\psi_C^{-1}(B_r(0)) \geq 1-\delta$ and so the measure is scalarly concentrated on the balls of $L^2(\mathbb{T})$. □

Proposition 8.2.5. *The pushforward of the cylindrical measure μ_C by the Hilbert-Schmidt operator u is the measure $\mathcal{W} = u_{\#}\mu_C$ on $L^2(\mathbb{T})$ is a Radon measure, which is denoted $\mathcal{W}$ for Wiener loop.*

Proof. This follows from an application of Theorem 8.2.1. The definition of the new measure is best explained by looking at the Fourier series of the relevant functions $f = \sum_{-\infty}^{\infty} a_n e^{inx}$. Let us consider the projection to a one dimensional linear subspace $C_n = \text{span}(e^{inx})$, the pushforward measure on the cylinder set $\psi_{C_n}^{-1}(A)$

$$\begin{aligned} u_{\#}\mu_{C_n}(\psi_{C_n}^{-1}(A)) &= \mu_{C_n}\left(u^{-1}\left(\{f \in L^2(\mathbb{T}) \,:\, \langle f, e^{inx}\rangle = a_n \in A\}\right)\right) \\ &= \mu_{C_n}\left(\{f \in L^2(\mathbb{T}) \,:\, \langle u(f), e^{inx}\rangle = \frac{a_n}{n} \in A\}\right). \end{aligned}$$

For the measure μ_C on the cylinder sets, a_n is distributed like a standard normal distribution $N(0,1)$, therefore when it is scaled by $\frac{1}{n}$ the new distibution will be $N(0,\frac{1}{n}^2)$. The measure is given by,

$$u_{\#}\mu_{C_n}(\psi_{C_n}^{-1}(A)) = \frac{1}{\sqrt{2\pi}}\int_A \exp\left(-\frac{n^2x^2}{2}\right) ndx.$$

□

This concludes the method of constructing the Gibbs measure developed by Schwartz. This example is somewhat of a canonical measure on a periodic function space, and as such was informative in developing the wider theory of translation invariant Radon measures on infinite dimensional Hilbert spaces (See the Cameron-Martin Theorem[14]).

The construction by Weiner

This specific example was discovered earlier thanks to Wiener, hence the name Wiener loop measure. It can be enlightening to follow his construction. His development of the measure made use of an abstract space Ω of infinitely many independent standard normal distributions $N(0,1)$.

Consider a measurable function $\phi : \Omega \to L^2(\mathbb{T})$, denote each independent standard normal distribution by $\gamma_n(\omega)$, then the measurable function is $\phi : \omega \mapsto \sum_{n=-\infty}^{\infty} \gamma_n(\omega)e^{inx}$.

The pushforward of the measure on Ω by ϕ gives the cylindrical measure μ_C from Proposition 8.1.5. For a cylinder set $\psi_C^{-1}(A)$ in $L^2(\mathbb{T})$ spanned by $\{e^{in_1x}, ..., e^{in_mx}\}$ the preimage of that set $\phi^{-1}(\psi_C^{-1}(A)) = \{\omega \ : \gamma_{n_k}(\omega) \in A_k,\ k = 1, ..., m\}$. The measure of this set is that of Gaussian measure of A in $\mathbb{R}^m$.

Radonification can be applied by composing ϕ with u to get $u\circ\phi : \omega \mapsto \sum_{n=-\infty}^{\infty} \frac{\gamma_n(\omega)}{n} e^{inx}$. The pushforward of the measure on Ω by $u \circ \phi$ gives the Wiener loop measure $\mathcal{W}$. For a cylinder set $\psi_C^{-1}(A)$ in $L^2(\mathbb{T})$ spanned by $\{e^{in_1x}, ..., e^{in_mx}\}$ the preimage of that set under u is

$$u^{-1}(\psi_C^{-1}(A)) = \left\{ \sum_{n=-\infty}^{\infty} a_n e^{inx} \ : \ \sum_{n=-\infty}^{\infty} \frac{a_n}{n} e^{inx} \in \psi_C^{-1}(A) \subset L^2(\mathbb{T}) \right\}.$$

The preimage of this set under ϕ is $\phi^{-1}(u^{-1}(\psi_C^{-1}(A))) = \{\omega : \frac{\gamma_{n_k}(\omega)}{n} \in A_k,\ k = 1, ..., m\}$.

Each random variable $\gamma_{n_k}(\omega)/n$ is distributed as $N(0, 1/n^2)$, and so the probability

$$\mathbb{P}(\{\omega \ : \frac{\gamma_{n_k}(\omega)}{n} \in A_k, \ k = 1, ..., m\}) = \frac{1}{\sqrt{2\pi}^m} \int_{A_1} ... \int_{A_m} \exp\left(-\frac{1}{2}\sum_{n=1}^{m} n^2 x_n^2\right) \prod_{n=1}^{m} n dx_n.$$

This is the pushforward of the cylindrical measure μ_C by u and is therefore the Wiener loop measure $\mathcal{W}$.

Finally consider this measure on the subspace $H^1 \subset L^2(\mathbb{T})$ as defined in Definition 3.2.7. So here $\tilde{\psi}_C^{-1}(A) = \{f \in H^1 \ : \psi_C(f) \in A\}$.

$$\int_{\tilde{\psi}_C^{-1}(A)} d\mathcal{W} = \frac{1}{\sqrt{2\pi}^m} \int_{A_1} ... \int_{A_m} \exp\left(-\frac{1}{2}\sum_{j=1}^{m} j^2 x_j^2\right) \prod_{j=1}^{m} j dx_j, \tag{8.6}$$

Now moving the expression back into the function space using the definition of the pushforward ψ, one can write down a expression for the measure on subsets of H^1. The expression only derives any meaning from its equality to the line above however.

$$= \frac{1}{\sqrt{2\pi}^m} \int_{\tilde{\psi}_C^{-1}(A)} \exp\left(-\frac{1}{2}\|\psi_C(f'(x))\|_{L^2}\right) \prod_{j=1}^{m} j dx_j, \tag{8.7}$$

$$= \frac{1}{\sqrt{2\pi}^m} \int_{\tilde{\psi}_C^{-1}(A)} \exp\left(-\frac{1}{2}\int_0^{2\pi} (\psi_C(f'(x)))^2 dx\right) \prod_{j=1}^{m} j dx_j. \tag{8.8}$$

Thanks to Sazonov's theorem, after radonification the measure now countably additive so Equation (8.6) can extend to limits as $\dim(C) \to \infty$. Note however that although the measure is built on the compact subsets of L^2 in H_1, the measure is zero on any of these subsets, (for example $\{f \in H^1 \ : \|f\|_{H^1} \leq K\}$).

8.3 Defining the Gibbs measure

Definition 8.3.1. If $P, Q \in L^2(\mathbb{T})$ then the modified canonical ensemble of the Gibbs measure on $L^2(\mathbb{T}) \times L^2(\mathbb{T})$ is defined as,

$$\nu_{\beta,K}(dP, dQ) = \frac{1}{Z_K} \mathbb{I}_{B_K} \exp\left(-H(P, Q)\right) \prod dP dQ, \tag{8.9}$$

where Z_K is a normalisation constant, $B_K = \{f \in L^2(\mathbb{T}) : \|f\|_{L^2(\mathbb{T})} < K\}$ and $H(P,Q)$ is the Hamiltonian for the NLS given in Equation (5.2).

To interpret Equation (8.9), note that the second and third terms of the Hamiltonian resemble the density terms for Wiener loop measure given in Equation (8.8). In addition, the preimage of the ball B_K under the map ϕ defined in the last section is the set $\Omega_K = \{\omega \in \Omega : \sum_{i=-\infty}^{\infty} \gamma_i(\omega)^2/i^2 \leq K\}$.

Thus the measure $\nu_{\beta,K}$ on $L^2(\mathbb{T}) \times L^2(\mathbb{T})$ can be defined with two copies of Wiener loop measure $\mathcal{W}$, a density $\exp(\frac{\beta}{4}\int (P^2+Q^2)^2 ds)$, and a cutoff on the set B_K by

$$\nu_{\beta,K}(dP,dQ) = \frac{1}{Z_K}\mathbb{I}_{B_K} \exp\left(\int_0^{2\pi} \frac{\beta}{4}(P^2+Q^2)^2 ds\right) \mathcal{W}(dP)\mathcal{W}(dQ). \tag{8.10}$$

However, $\exp\left(\frac{\beta}{4}\int (P^2+Q^2)^2 ds\right)$ may not be integrable with respect to $\mathcal{W}(dP)\mathcal{W}(dQ)$.

Lemma 8.3.2 (Leborowitz, Rose and Speer)**.** *[48, Thm 2.2] The modified canonical ensemble of the Gibbs measure is finite, and so normalizable, for any $\beta \in \mathbb{R}$ and $N > 0$. In other words,*

$$\int_{L^2(\mathbb{T})\times L^2(\mathbb{T})} \mathbb{I}_{B_K} \exp\left(\frac{\beta}{4}\int_0^{2\pi} (P^2+Q^2)^2 dx\right) \mathcal{W}(dP)\mathcal{W}(dQ) < \infty \tag{8.11}$$

for all $\beta > 0$. This integral is denoted Z_K.

Remark 8.3.3. The focussing case, in which $\beta < 0$, poses less of a problem. The exponential $\exp\left(\frac{\beta}{4}\int_0^{2\pi}(P^2+Q^2)^2 dx\right)$ will always be finite because the integral is positive.

8.3.1 Finite dimensional subspaces

The specific construction of the Gibbs measure using Fourier modes of functions in H^1 lead conveniently to a family of finite dimensional subspaces, $M^{(n)}$ and their accompanying Gibbs measures $\nu^{(n)}_{\beta,K}$. The following is taken from earlier work in Ref. [8].

Let D_n be the Dirichlet projection taking $\sum_{k=-\infty}^{\infty}(a_k+ib_k)e^{ik\theta}$ to $\sum_{k=-n}^{n}(a_k+ib_k)e^{ik\theta}$. Following [10], we truncate the random Fourier series of $u = P + iQ = \sum_{k=-\infty}^{\infty}(a_k +$

$ib_k)e^{ik\theta}$ to $u_n = P_n + iQ_n = \sum_{k=-n}^{n}(a_k + ib_k)e^{ik\theta}$ and correspondingly modify the Hamiltonian to

$$H_3^{(n)}((a_k),(b_k)) = \frac{1}{2}\sum_{k=-n}^{n} k^2(a_k^2+b_k^2) + \frac{\beta}{4}\int\Big|\sum_{k=-n}^{n}(a_k+ib_k)e^{ik\theta}\Big|^4\frac{d\theta}{2\pi} \tag{8.12}$$

for the real canonical variables $((a_k,b_k))_{k=-n}^{n}$. Then the canonical equations become a coupled system of ordinary differential differential equations in the Fourier coefficients. We introduce the polar decomposition $P_n + iQ_n = \kappa_n e^{i\sigma_n}$, and observe that in terms of these noncanonical variables, the Hamiltonians $H_1^{(n)} = \int_{\mathbb{T}}\kappa_n^2 d\theta$ and

$$H_3^{(n)} = \frac{1}{2}\int_{\mathbb{T}}\Big(\Big(\frac{\partial\kappa_n}{\partial\theta}\Big)^2 + \kappa_n^2\Big(\frac{\partial\sigma_n}{\partial\theta}\Big)^2\Big)\frac{d\theta}{2\pi} + \frac{\beta}{4}\int_{\mathbb{T}}\kappa_n^4\frac{d\theta}{2\pi} \tag{8.13}$$

are invariants under the flow.

The corresponding Gibbs measure is

$$d\nu_{\beta,K}^{(n)} = Z(K,\beta,n)^{-1}\mathbb{I}_{B_K}(u_n)\exp\Big(\frac{-\beta}{4}\int_{\mathbb{T}}|u_n(\theta)|^4\frac{d\theta}{2\pi}\Big)W(du_n) \tag{8.14}$$

in which $W(du_n)$ is the finite dimensional projection of Wiener loop measure and is defined in terms of the Fourier modes as

$$W(du_n) = \prod_{j=-n;j\neq 0}^{n}\exp\Big(-\frac{j^2}{2}(a_j^2+b_j^2)\Big)\frac{j^2 da_j db_j}{2\pi}. \tag{8.15}$$

Consider the map $u(x,t)\mapsto u(x+h,t)$ of translation in the space variable. This commutes with D_n, and the Gibbs measures $\nu_{\beta,K}^{(n)}$ are all invariant under this translation. In terms of Fourier components, we have $M_\infty = B_K$ and

$$M_n = \Big\{(a_j,b_j)_{j=-n}^{n} : a_j,b_j\in\mathbb{R} : \sum_{j=-n}^{n}(a_j^2+b_j^2)\leq K\Big\} \tag{8.16}$$

with the canonical inclusions of metric spaces $(M_1,\ell^2)\subset(M_2,\ell^2)\subset\cdots\subset(M_\infty,\ell^2)$ defined by adding zeros at the start and end of the sequences, which gives a sequence of isometric embeddings for the ℓ^2 metric on sequences. When we identify $(a_j,b_j)_{j=-n}^{n}$ with $\sum_{j=-n}^{n}(a_j+ib_j)e^{ij\theta}$, then we have a corresponding embedding for the L^2 metric.

Here $(M_n, L^2, \nu^{(n)}_{\beta,K})$ is a finite-dimensional manifold and a metric probability space. We now show that these spaces converge to $(M_\infty, L^2, \mu_{K,\beta})$ as $n \to \infty$.

Definition 8.3.4. (Convergence of metric measure spaces)

(i) For M a nonempty set, a pseudometric is a function $\delta : M \to [0, \infty]$ such that

$$\delta(x,y) = \delta(y,x), \quad \delta(x,x) = 0, \quad \delta(x,z) \le \delta(x,y) + \delta(y,z) \qquad (x,y,z \in M); \quad (8.17)$$

then (M, δ) is a pseudometric space.

(ii) Given pseudo metric spaces (M_1, δ_1) and (M, δ_2), a coupling is a pseudo metric $\delta : M \to [0, \infty]$ where $M = M_1 \sqcup M_2$ such that $\delta \mid M_1 \times M_1 = \delta_1$ and $\delta \mid M_2 \times M_2 = \delta_2$.

(iii) Suppose that $\hat{M}_1 = (M_1, \delta_1, \mu_1)$ and $\hat{M}_2 = (M_2, \delta_2, \mu_2)$ are complete separable metric spaces endowed with probability measures. Consider a coupling (M, δ) and a probability measure π on $M_1 \times M_2$ with marginals $\pi_1 = \mu_1$ and $\pi_2 = \mu_2$. Then the L^2 distance between $\hat{M}_1$ and $\hat{M}_2$ is

$$\mathfrak{D}_{L^2}(\hat{M}_1, \hat{M}_2) = \inf_{\delta,\pi} \Big(\int_{M \times M} \delta(x,y)^2 \pi(dxdy) \Big)^{1/2} \quad (8.18)$$

Lemma 8.3.5. *(i) Suppose that* $0 < -\beta K < 3/(14\pi^2)$. *Then* $\hat{M}_n = (M_n, L^2, \nu^{(n)}_{\beta,K})$ *has*

$$\mathfrak{D}_{L^2}(\hat{M}_n, \hat{M}_\infty) \to 0 \qquad (n \to \infty). \quad (8.19)$$

(ii) The measures $\nu^{(n)}_{\beta,K}$ *converge in total variation norm to* $\nu_{K,\beta}$ *as* $n \to \infty$.

Proof. (i) This is proved in Theorem 3.2 of Ref. [6]; see also Example 3.8 of Ref. [69]. Let $W_2(\nu^{(n)}, \nu)$ be the Wasserstein transportation distance between free Brownian loop measure μ and the pushforward of μ under the Dirichlet projection, $\nu^{(n)} = D_n \sharp \nu$, for the cost function $\|u - v\|^2_{L^2}$.

The key point is

$$\begin{aligned} W_2(\mu^{(n)}, \mu)^2 &\le \int \|D_n u - u\|^2_{L^2} \mu(du) \\ &= \mathbb{E} \sum_{k:|k|>n} \frac{|\gamma_k|^2}{k^2} = O\Big(\frac{1}{n}\Big) \qquad (n \to \infty). \end{aligned} \quad (8.20)$$

(ii) The measures $\nu_{\beta,K}^{(n)}$ converge in total variation norm to μ_K, by an observation of McKean[51] in his step 7. By M. Riesz's theorem, there exists $c_4 > 0$ such that $\int_{\mathbb{T}} |D_n u|^4 d\theta \le c_4 \int_{\mathbb{T}} |u|^4 d\theta$, and by [48] the integral

$$\int_{B_K} \exp\Big(\lambda c_4 \int_{\mathbb{T}} |u(\theta)|^4 d\theta\Big) W(du) \tag{8.21}$$

is finite, so we can use the integrand as a dominating function to show

$$\int_{B_K} \Big|\exp\Big(\lambda \int_{\mathbb{T}} |D_n u(\theta)|^4 d\theta\Big) - \exp\Big(\lambda \int_{\mathbb{T}} |u(\theta)|^4 d\theta\Big)\Big| W(du) \to 0 \qquad (n \to \infty). \tag{8.22}$$

□

Chapter 9

Weak convergence of solutions to the NLSE

This chapter is Section IV of [8] in collaboration with G. Blower. It discusses the Lax pair for the NLSE as derived in Chapter 5, and uses facts about Lie algebra's from Section 2.3.1. The main impetus of the chapter is to deduce under what conditions weak solutions to the Lax pair formulation exist. As discussed when introducing the previous chapter, weak solutions allow for discussion of the evolution of an ensemble of typical solutions given typical initial conditions — for example in Chapter 10 this takes the form of a stochastic process.

In the case of the the nonlinear Schrödinger equation, consider the Cauchy problem,

$$\begin{aligned} \frac{1}{i}\frac{\partial \psi}{\partial t} &= \frac{\partial^2 \psi}{\partial x^2} + \beta|\psi|^2\psi \\ \psi_0 &= \phi(x) \in H^1(\mathbb{T}, \mathbb{C}). \end{aligned}$$

Bourgain proves that this problem is well posed [9, Thm. 1], and Lebowitz, Rose and Speer [48] prove the Gibbs measure is invariant under the flow of the NLSE, as discussed in Chapter 8. This means the distribution of the random process at each time point is given by the Gibbs measure.

As discussed in Chapter 5, if $\psi = P + iQ = \kappa e^{i\sigma}$ according to the Hasimoto transform, then the partial differential equation for the NLS is transformed into the Lax pair

of equations (5.15) and (5.16),

$$\frac{\partial}{\partial x}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}, \tag{9.1}$$

$$\frac{\partial}{\partial t}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix} = \begin{bmatrix}0 & -\tau\kappa & \frac{\partial \kappa}{\partial x}\\ \tau\kappa & 0 & -\mu\\ -\frac{\partial \kappa}{\partial x} & \mu & 0\end{bmatrix}\begin{bmatrix}\mathbf{t}\\ \mathbf{n}\\ \mathbf{b}\end{bmatrix}. \tag{9.2}$$

and this chapter discusses whether weak solutions to this pair of ODEs exist given initial data as in the above Cauchy problem. Later a stochastic differential equation is constructed for the same pair of ODEs.

9.1 Gibbs measure transported to the frames

The compact Lie group $SO(3)$ of real orthogonal matrices with determinant one is a subset of $M_{3\times 3}(\mathbb{R})$, which has the scalar product $\langle X, Y\rangle = \text{trace}(XY^\top)$ and associated metric $d(X,Y) = \langle X-Y, X-Y\rangle^{1/2}$ such that $\langle XU, YU\rangle = \langle X, Y\rangle$ and $d(XU, YU) = d(X,Y)$ for all $U \in SO(3)$ and $X, Y \in M_{3\times 3}(\mathbb{R})$. The Lie group $SO(3)$ has tangent space at the identity element give by the skew symmetric matrices $so(3)$, so the tangent space $T_X SO(3)$ at $X \in SO(3)$ consists of $\{\Omega X : \Omega \in so(3)\}$, where $so(3)$ is a Lie algebra for $[x,y] = xy - yx$, $x, y \in so(3)$, and the exponential map is surjective $so(3) \to SO(3)$.

Consider the differential equation

$$\frac{dX}{dt} = \Omega(t)X; \quad X(0) = X_0 \tag{9.3}$$

where $t \in [0,1]$ is the evolving time, and $X \in SO(3)$. We consider a column vector $x \in \mathbb{R}^3$, satisfying $\frac{dx}{dt} = \Omega x$ which gives a velocity, and $\|x\| = 1$ because $\Omega \in so(3)$. Following Otto's interpretation[74] of optimal transport in the setting of partial differential equations, one constructs a weakly continuous family of probability measures, $\tilde{\nu}_t$ on $\mathbb{S}^2$ for $t \in [0,1]$, which satisfy the weak continuity equation,

$$\frac{\partial \tilde{\nu}_t}{\partial t} + \nabla \cdot \left(\Omega x \tilde{\nu}_t\right) = 0. \tag{9.4}$$

Likewise the differential equation (9.3) gives a weakly continuous family of probability measures, ν_t on $SO(3)$. If the integral

$$\int_0^1 \int_{SO(3)} \|\Omega X\|^2_{M_{3\times 3}(\mathbb{R})} \nu_t(dX)dt < \infty, \tag{9.5}$$

and ΩX is locally bounded, then ΩX is locally Lipschitz and ν_t is the unique solution to the weak continuity equation by Thm 5.34 of Ref.[74]. Recall that for the operator norm on $M_{3\times 3}(\mathbb{R})$, $\|A\| = \sup\{\|Ay\| : y \in \mathbb{R}^3\}$, where $\|X\| = 1$ for all $X \in SO(3)$ so $\|\Omega X\| \leq \|\Omega\|$.

The weak continuity equation is equivalent to

$$\int_{SO(3)} f(X)\nu_t(dX) = \int_{SO(3)} f(X_t(X_0))\nu_0(dX_0) \tag{9.6}$$

for all $f \in C(SO(3);\mathbb{R})$, where $X_0 \mapsto X_t(X_0)$ gives the dependence of the solution of (9.3) on the initial condition. The velocity field ΩX is associated with a transportation plan taking ν_{t_1} to ν_{t_2} which is possibly not optimal, but does give an upper bound on the Wasserstein distance for the cost $d(X,Y)^2$ on $SO(3)$ of

$$\frac{W_2(\nu_{t_2},\nu_{t_1})^2}{t_t - t_1} \leq \int_{t_1}^{t_2} \int_{SO(3)} \|\Omega\|^2_{M_{3\times 3}(\mathbb{R})} \nu_t(dX)dt \qquad (0 < t_1 < t_2 < 1). \tag{9.7}$$

Then by Theorem 23.9 of Ref. [73], the path (ν_t) of probability measures is absolutely continuous, so there exists $\ell \in L^1[0,1]$ such that $W_2(\nu_{t_2},\nu_{t_1}) \leq \int_{t_1}^{t_2} \ell(t)dt$ and 1/2-Hölder continuous, so there exists $C > 0$ such that $W_2(\nu_{t_2},\nu_{t_1}) \leq C|t_2 - t_1|^{1/2}$.

Example 9.1.1. (i) If $\Omega_t \in M_{3\times 3}(\mathbb{R})$ is skew, and X_t, Y_t give solutions of the differential equation

$$\frac{dX}{dt} = \Omega_t X, X(0) = X_0; \quad \frac{dY}{dt} = \Omega_t Y, Y(0) = Y_0 \tag{9.8}$$

then $d(X_t,Y_t) = d(X_0,Y_0)$. We deduce that if X_0 is distributed according to Haar measure on $SO(3)$, then X_t is also distributed according to Haar measure since the measure, the metric and solutions are all preserved via $X \mapsto XU$. Haar measure on $SO(3)$ was derived by Hurwitz [18, §3.2] and can be expressed explicitly in terms of Euler angles as $\mu = 2^{\frac{3}{2}}\sin(\phi)d\theta d\phi d\psi$ where $0 \leq \theta \leq \pi$, $0 \leq \phi, \psi < 2\pi$ are the Euler angles and the measure is invariant up to a multiplicative constant.

(ii) As an alternative, we can consider X_0 to have first column $[0;0;1]$ and observe the evolution of the first column T of X under the (9.3) where T evolves on $\mathbb{S}^2$.

We now consider the case in which Ω as in (9.2) is a $so(3)$-valued random variable over $(M_\infty, \mu_{K,\beta}, L^2)$.

Proposition 9.1.2. *Suppose that $\Omega = \Omega(u(\cdot,t))$ where $u(x,t)$ is a solution of NLS and that*

$$\int_{B_K} \|\Omega(u(\cdot,0))\|^2_{M_{3\times 3}(\mathbb{R})} \mu_{K,\beta}(du) \tag{9.9}$$

converges. Then for almost all u with respect to $\mu_{K,\beta}$, there exists a flow $(\nu_t(dX;u))$ of probability measures on $SO(3)$.

Proof. Each solution u of NLS determines Ω so that the associated ODE (9.3) transports the initial distribution of $X_0 \in SO(3)$ to a probability measure on $SO(3)$; then we average over the u with respect to $\mu_K(du)$. This Gibbs measure is invariant under the NLS flow, so by Fubini's theorem

$$\int_{B_K} \int_0^1 \int_{SO(3)} \|\Omega(u(\cdot,t))\|^2_{M_{3\times 3}(\mathbb{R})} \nu_t(dX) dt \mu_K(du) \tag{9.10}$$

converges. Hence the condition (9.5) is satisfied, for almost all u, and we can invoke Theorem 23.9 of Ref.[73]. □

For the finite-dimensional M_n of (8.16) and solutions $u_n = \kappa_n e^{i\sigma_n}$, the modified Hasimoto differential equations are

$$\frac{\partial}{\partial x} X^{(n)}(x,t) = \begin{bmatrix} 0 & \kappa_n & 0 \\ -\kappa_n & 0 & \tau_n \\ 0 & -\tau_n & 0 \end{bmatrix} X^{(n)}(x,t), \tag{9.11}$$

and

$$\frac{\partial}{\partial t} X^{(n)}(x,t) = \begin{bmatrix} 0 & -\tau_n\kappa_n & \frac{\partial \kappa_n}{\partial x} \\ \tau_n\kappa_n & 0 & \frac{\partial \sigma_n}{\partial t} + \beta\kappa_n^2 \\ -\frac{\partial \kappa_n}{\partial x} & -\frac{\partial \sigma_n}{\partial t} - \beta\kappa_n^2 & 0 \end{bmatrix} X^{(n)}(x,t) \tag{9.12}$$

involves $\tau_n = \frac{\partial \sigma_n}{\partial x}$ and $(\frac{\partial \kappa_n}{\partial x})^2 + \tau_n^2\kappa_n^2 = (\frac{\partial P_n}{\partial x})^2 + (\frac{\partial Q_n}{\partial x})^2$ which is continuous, so there exists a solution $X^{(n)}(x,t) \in SO(3)$. We can interpret the solutions as elements of a fibre bundle over $(M_n, \mu_K^{(n)}, L^2)$ with fibres that are isomorphic to $SO(3)$.

Let $P + iQ = \kappa e^{i\sigma}$ be a solution of NLS and let

$$\Omega_1 = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix}. \tag{9.13}$$

Proposition 9.1.3. *(i) Let $P + iQ = \kappa e^{i\sigma}$ be a solution of NLS with initial data in $P(x,0) + iQ(x,0) \in B_K \cap H^1$. Then Ω_1 in (9.13) gives an $so(3)$-valued vector field in $L^2(\kappa^2(x,t)dx)$.*

(ii) Let $P + iQ = \kappa e^{i\sigma}$ be a solution of NLS with initial data $P(x,0) + iQ(x,0) \in H^1 \cap B_K$, and let $P_n + iQ_n = \kappa_n e^{i\sigma_n}$ be the corresponding solution of the NLS truncated in Fourier space, giving matrix $\Omega_1^{(n)}$. Let $X_t^{(n)}(x)$ be a solution of (9.11) and suppose that $X^{(n)}$ converges weakly in L^2 to $X_t(x)$. Then X_t gives a weak solution of Equation (9.1).

Proof. (i) With $\omega = \sqrt{\kappa^2 + \tau^2}$, we have

$$\exp(h\Omega_1) = I + \frac{\sin h\omega}{\omega}\Omega_1 + \frac{1 - \cos h\omega}{\omega^2}\Omega_1^2$$

where the entries of Ω_1^2 are bounded by $\kappa^2 + \tau^2$, hence

$$\int_{\mathbb{T}} \|\Omega_1(x,t)\|^2_{M_{3\times 3}(\mathbb{R})} \kappa(x,t)^2 dx < \infty \tag{9.14}$$

for $u \in H^1$; however, there is no reason to suppose that τ itself is integrable with respect to dx.

(ii) By (5.18) and (5.19), we have $\kappa\Omega_1 \in L^2_x$ for all $u \in H^1$. Moreover, Bourgain [9] has shown that for initial data $P(x,0) + iQ(x,0) = \kappa(x,0)e^{i\sigma(x,0)}$ in $H^1 \cap B_K$, the map

$$\kappa(x,0)e^{i\sigma(x,0)} \mapsto \kappa(x,t)\Omega_1(x,t) \in L^2 \tag{9.15}$$

is Lipschitz continuous for $0 \le t \le t_0$ with Lipschitz constant depending upon $t_0, K > 0$.

We have

$$\frac{\|\kappa(x+h,t)X(x+h,t)-\kappa(x,t)X(x,t)\|^2}{h^2} \le 2\Big(\frac{1}{h}\int_x^{x+h}\big|\frac{\partial\kappa}{\partial y}(y,t)\big|dy\Big)^2 + 2\Big(\frac{1}{h}\int_x^{x+h}\kappa(y,t)\|\Omega_1(y,t)\|dy\Big)^2 \quad (9.16)$$

where the right-hand side is integrable with respect to x by the Hardy–Littlewood maximal inequality and (9.14). Suppose that $X^{(n)}$ is a solution of Equation (9.11). We take τ_n to be locally bounded. Then by applying Cauchy–Schwarz inequality to the integral

$$X^{(n)}(x+h,t)-X^{(n)}(x,t)=\int_0^h \Omega_1^{(n)}(x+s,t)X^{(n)}(x+s,t)ds,$$

we deduce that

$$\int_{[0,2\pi]}\|X^{(n)}(x+s,t)-X^{(n)}(x,t)\|^2_{M_{3\times3}(\mathbb{R})}\kappa_n(x,t)^2dx \le h\int_0^h\int_{[0,2\pi]}\|\Omega_1^{(n)}(x+s,t)\|^2_{M_{3\times3}(\mathbb{R})}\kappa_n(x,t)^2dxds \quad (9.17)$$

where the integral is finite by (9.14). Also

$$\sum_{j=1}^N \frac{\|X^{(n)}(x_j,t)-X^{(n)}(x_{j-1},t)\|^2_{M_{3\times3}(\mathbb{R})}}{x_j-x_{j-1}} \le \int_{x_0}^{x_N}\|\Omega_1^{(n)}(x,t)\|^2dx$$

for $0<x_1<x_2<\cdots<x_N<2\pi$. We have

$$\frac{\partial}{\partial x}\big(\kappa_n X^{(n)}\big)=\frac{\partial\kappa_n}{\partial x}X^{(n)}+\kappa^{(n)}\Omega_1^{(n)}X^{(n)} \quad (9.18)$$

so for $Z\in C^\infty([0,2\pi];M_{3\times3}(\mathbb{R}))$ and the inner product on $M_{3\times3}(\mathbb{R})$, we have

$$\begin{aligned}&\langle\kappa_n(2\pi)X^{(n)}(2\pi),Z(2\pi)\rangle-\langle\kappa_n(0)X^{(n)}(0),Z(0)\rangle-\int_0^{2\pi}\kappa_n(x)\langle X^{(n)}(x),Z(x)\rangle\,dx\\ &=\int_0^{2\pi}\frac{\partial\kappa_n}{\partial x}\langle X^{(n)}(x),Z(x)\rangle\,dx+\int_0^{2\pi}\langle X^{(n)},\kappa_n(x)\Omega_1^{(n)}(x)^\top Z(x)\rangle dx \quad (9.19)\end{aligned}$$

where $\kappa_n\to\kappa$ in H^1, so with norm convergence, we have $\frac{\partial\kappa_n}{\partial x}\to\frac{\partial\kappa}{\partial x}$ in L^2, and $\kappa_n\Omega^{(n)}\to$

$\kappa\Omega_1$ as $n \to \infty$, and with weak convergence in L^2, we have $X^{(n)} \to X$, so

$$\langle \kappa(2\pi)X(2\pi), Z(2\pi)\rangle - \langle \kappa(0)X(0), Z(0)\rangle - \int_0^{2\pi} \kappa(x)\langle X(x), Z(x)\rangle \, dx \tag{9.20}$$
$$= \int_0^{2\pi} \frac{\partial \kappa}{\partial x}\langle X(x), Z(x)\rangle \, dx + \int_0^{2\pi} \langle X, \kappa(x)\Omega_1(x)^\top Z(x)\rangle dx.$$

□

The simulation of this differential equation computes $X_x \in \mathbb{S}^2$ starting with $X_0 = [0; 0; 1]$ and produces a frame $\{X_x, \Omega_x X_x, X_x \times \Omega_x X_x\}$ of orthogonal vectors. Geodesics on $\mathbb{S}^2$ are the curves such that the principal normal is parallel to the position vector, namely the great circles. For a geodesic, $X_x \times \Omega_x X_x$ is perpendicular to the plane that contains the great circle.

Let $P + iQ = \kappa e^{i\sigma}$ be a solution of NLS and let

$$\Omega_2 = \begin{bmatrix} 0 & -\kappa\tau & \frac{\partial \kappa}{\partial x} \\ \kappa\tau & 0 & 0 \\ -\frac{\partial \kappa}{\partial x} & 0 & 0 \end{bmatrix}. \tag{9.21}$$

Proposition 9.1.4. *(i) Let $P + iQ = \kappa e^{i\sigma}$ be a solution of NLS with initial data $P(x, 0) + iQ(x, 0) \in B_K$. Then $x \mapsto \int_0^x \Omega_2(y, t)dy$ gives a so(3)-valued stochastic of finite quadratic variation on $[0, 2\pi]$ almost surely with respect to $\mu_K(dPdQ)$.*

(ii) Let $P + iQ = \kappa e^{i\sigma}$ be a solution of NLS with initial data $P(x, 0) + iQ(x, 0) \in H^1 \cap B_K$, and let $P_n + iQ_n = \kappa_n e^{i\sigma_n}$ be the corresponding solution of the NLS truncated in Fourier space, giving matrix $\Omega_2^{(n)}$. Let $X_t^{(n)}$ be a solution of (9.12). Then $X_t^{(n)}$ converges in L_x^2 norm to X_t as $n \to \infty$ where X_t gives a weak solution of (9.2).

Proof. (i) The essential estimate is

$$\begin{aligned}
&\int_{B_K} \sum_j |\kappa(x_{j+1},t)-\kappa(x_j,t)|^2 \mu_K(du) \\
&\qquad \le \sum_j \Big(\int_{B_K} |u(x_{j+1},t)-u(x_j,t)|^2 \mu_K(du) \Big) \\
&\qquad \le \sum_j \Big(\int_{B_K} |u(x_{j+1},t)-u(x_j,t)|^4 W_K(du) \Big)^{1/2} \Big(\int_{B_K} \Big(\frac{d\mu_K}{dW} \Big)^2 dW \Big)^{1/2} \\
&\qquad \le C \sum_j \Big(\int_{B_K} |u(x_{j+1},t)-u(x_j,t)|^2 W(du) \Big)^{1/2} \\
&\qquad \le C \sum_j (x_{j+1}-x_j) \le 2\pi C. \qquad (9.22)
\end{aligned}$$

The function σ is a progressively measurable stochastic process adapted with respect to a suitable filtration, and with differential satisfying an Ito integral equation[24]. Therefore, we can control the $\kappa\tau$ term via

$$\int_0^x (\kappa d\sigma - 2^{-1}\kappa^2 \langle d\sigma, d\sigma \rangle) = \int_0^x \kappa \nabla \sigma \cdot \begin{bmatrix} dP \\ dQ \end{bmatrix} = \int_0^x \frac{-Q dP + P dQ}{\sqrt{P^2+Q^2}} \qquad (9.23)$$

which is a bounded martingale transform of Wiener loop. As dP and dQ are martingale differences, the integral is a martingale transform [12].

(ii) By (5.18) and (5.19), we have $\Omega_2 \in L^2_x$ for all $u \in H^1$. Bourgain [9] has shown that for initial data $P(x,0)+iQ(x,0) = \kappa(x,0)e^{i\sigma(x,0)}$ in $H^1 \cap B_K$, the map

$$\kappa(x,0)e^{i\sigma(x,0)} \mapsto \Omega_2(x,t) \in L^2_x \qquad (9.24)$$

is Lipschitz continuous for $0 \le t \le t_0$ with Lipschitz constant depending upon $t_0, K > 0$. We have

$$\int_0^{2\pi} \|\Omega_2(x)\|^2 dx \le 2 \int_0^{2\pi} \Big(\Big(\frac{\partial \kappa}{\partial x} \Big)^2 + \kappa(x)^2 \tau(x)^2 + \kappa(x)^4 \Big) dx,$$

where the final integral is part of the Hamiltonian. With $Z \in C^\infty(\mathbb{T}; M_{3\times 3}(\mathbb{R}))$, we

have the integral equation for the pairing $\langle \cdot, \cdot \rangle$ on $L^2([0, 2\pi], M_{3\times 3}(\mathbb{R}))$

$$\langle X_t^{(n)}, Z \rangle = \langle X_0^{(n)}, Z \rangle + \int_0^t \langle X_s^{(n)}, (\Omega_s^{(n)})^\top Z \rangle \, ds. \tag{9.25}$$

Consider the variational differential equation in $L^2([0, 2\pi], M_{3\times 3}(\mathbb{R}))$

$$\begin{aligned} \frac{d}{dt}(X^{(m)}(x,t) - X^{(n)}(x,t)) &= \Omega_2^{(n)}(x,t)(X^{(m)}(x,t) - X^{(n)}(x,t)) \\ &\quad + (\Omega_2^{(m)}(x,t) - \Omega_2^{(n)}(x,t))X^{(m)}(x,t) \end{aligned} \tag{9.26}$$

where $\Omega_2^{(n)}(x,t)$ and $\Omega_2^{(m)}(x,t) - \Omega_2^{(n)}(x,t)$ are skew.

We introduce a family of matrices $U^{(n)}(x;t,s)$ such that $U^{(n)}(x;t,r)U^{(n)}(x;r,s) = U^{(n)}(x;t,s)$ for $t > r > s$ and $U^{(n)}(x;t,t) = I$ such that

$$\frac{\partial}{\partial t} U^{(n)}(x;t,s) = \Omega_2^{(n)}(x;t) U^{(n)}(x;t,s). \tag{9.27}$$

Then the variational equation has solution

$$\begin{aligned} X^{(m)}(x,t) - X^{(n)}(x,t) &= U^{(n)}(x;t,0)(X^{(m)}(x,0) - X^{(n)}(x,0)) \\ &\quad + \int_0^t U^{(n)}(x;t,r)(\Omega_2^{(m)}(x;r) - \Omega_2^{(n)}(x;r))X^{(m)}(x,r)dr. \end{aligned}$$

Then

$$\begin{aligned} &\frac{d}{dt}\big\langle X^{(m)}(t) - X^{(n)}(t), X^{(m)}(t) - X^{(n)}(t) \big\rangle_{L_x^2} \\ &\quad = 2\Re\big\langle (\Omega_2^{(m)}(t) - \Omega_2^{(n)}(t))X^{(m)}(t), X^{(m)}(t) - X^{(n)}(t) \big\rangle_{L_x^2} \\ &\quad \le \|\Omega_2^{(m)}(t) - \Omega_2^{(n)}(t)\|_{L_x^2}^2 \|X^{(m)}(t)\|_{L_x^2}^2 + \|X^{(m)}(t) - X^{(n)}(t)\|_{L_x^2}^2 \end{aligned} \tag{9.28}$$

so from this differential inequality we have

$$\|X^{(m)}(t) - X^{(n)}(t)\|_{L_x^2}^2 \le e^t \|X^{(m)}(0) - X^{(n)}(0)\|_{L_x^2}^2 + \int_0^t e^{t-s} \|\Omega_2^{(m)}(s) - \Omega_2^{(n)}(s)\|_{L_x^2}^2 ds. \tag{9.29}$$

Now $X^{(m)}(0) - X^{(n)}(0) \to 0$ and $\Omega_2^{(m)}(s) - \Omega_2^{(n)}(s) \to 0$ in L_x^2 norm as $n, m \to \infty$, so there exists $X(x,t) \in L_x^2$ such that $X(x,t) - X^{(n)}(x,t) \to 0$ in L_x^2 norm as $n \to \infty$.

We deduce that

$$\langle X(t), Z\rangle_{L^2_x} = \langle X_0, Z\rangle_{L^2_x} + \int_0^t \langle X_u, (\Omega_2(u))^\top Z\rangle_{L^2_x}\, du, \tag{9.30}$$

so we have a weak solution of the ODE. □

Chapter 10

Numerics of the Hasimoto frame equation

The objective of this chapter is to build on the analysis of the nonlinear Schrödinger equation carried out in Chapters 5 and 9. Simulating a random numerical approximation to the solution of the first differential equation of the Lax pair for the NLSE, Equation (5.15). This equation is the evolution of the Frenet-Serret frame with the torsion and curvature specified by the NLSE. The evolution of the frame is modelled as a stochastic process and the stochastic differential equation it satisfies will be discussed in the following section. Solutions to the NLSE are Gibbs measurable functions [9]. The differential equations discussed provide a way to push this measure forward onto the sphere. We consider the case where the parameter β in (5.1) is equal to 0. In this case, the Gibbs measure is reduced to Wiener loop measure and stochastic processes with the Wiener loop measure as their law are by definition Brownian loop.

After producing the SDE, a numerical method for solving it can be implemented. The empirical measure of many sample paths of the process is then compared with the theoretical distribution statistically.

There exists stumbling blocks in applying this methodology more broadly, for example in the second differential equation of the Lax pair (Equation (5.16)) the function $\mu(x,t)$ includes a term equal to the derivative $\frac{\partial \tau}{\partial t}$ and if this is interpreted as I have done in this chapter, then the function μ is too rough to construct a SDE.

10.1 Background on stochastic calculus

Before applying stochastic calculus to the problem, an explanation of the theory behind the stochastic integral is introduced, starting with the definition of a stochastic process.

Definition 10.1.1. [42] A stochastic process is a collection of random variables, each on the same probability space, indexed by time.

$$X_t(\omega) \coloneqq X(\omega, t) : (\Omega, \mathcal{B}(\Omega)) \times ([0, T), \mathcal{B}([0, T))) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)). \tag{10.1}$$

We take the state space to be $\mathbb{R}^n$ with its Borel sets and the sample space Ω has on it an unknown measure ν. For a fixed event ω_i in the probability space Ω the measurable function $X_t(\omega_i)$ is known as a *sample path*. For each given time t the random variable $X_{t_i}(\omega)$ has a *distribution* or *law*, which is the pushforward of the measure ν to the state space, $\mathcal{P}(X_{t_i} \in A) = \nu(X_{t_i}^{-1}(A))$.

Definition 10.1.2. A continuous stochastic process is a stochastic process in which the sample paths are continuous functions.

Left and right continuous processes are defined analogously, and the concept of cadlag processes (continuous from the left with limits from the right) capture the broadest class of discontinuous processes considered.

Definition 10.1.3. [42] A filtration, $\{\mathcal{F}_t \mid t \in [0, \infty)\}$, is a collection of σ-algebras which are increasing: if $t_1 < t_2$ then $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$. A stochastic process can generate a filtration, simply take $\mathcal{F}_t$ to be the smallest σ-algebra generated by the random variables $\{X_s \mid s \in [0, t]\}$.

Definition 10.1.4. [42] An adapted process $(X_t, \mathcal{F}_t)$ is a stochastic process and a filtration that the process is measurable with respect to.

A progressively measurable stochastic process is such that $(\omega, t) \mapsto X_t(\omega) : (\Omega \times [0, s], \mathcal{F}_s \otimes \mathcal{B}([0, s])) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is measurable for each $0 \leq s$ [42, Def 1.1.11]. This is equivalent to being right continuous and adapted to the filtration [42, Prop. 1.1.13].

Definition 10.1.5. [42, Def 2.1.1] Brownian motion can be defined as a continuous stochastic process. The continuous stochastic process and filtration to which it is adapted, $(W_t, \mathcal{F}_t)$, are Brownian motion if they satisfy the following conditions.

(i) $W_0 = 0$ with probability 1.

(ii) Let $0 \leq s < t$, then the increment $W_t - W_s$ is independent of $\mathcal{F}_s$ and distributed as $N(0, \sqrt{t-s})$.

Let $\mathbb{I}$ denote the set of partitions of an interval L, and let $(\eta_n)_{n\in\mathbb{N}} \in \mathbb{I}$ denote a sequence of partitions, which are refinements (η_2 has added some subintervals between the partition defined by η_1). The Riemann integral over L is defined by approximating a function over any sequence of partitions η_n such that $\sup_{x_i\in\eta_n} |x_{i+1} - x_i| \to 0$ as $n \to \infty$. The same notion is used to calculate the variation of a function. The variation of f is defined,

$$\sup_{\eta\in\mathbb{I}} \sum_{x_i\in\eta_n} |f(x_i) - f(x_{i-1})|.$$

The Riemann-Stieltjes integral is defined for integrators of bounded variation. If f is a real valued function on the interval L and g is of bounded variation, then the integral

$$\int_L f(x)dg(x) = \lim \sum_{x_i\in\eta_n} f(x_{i-1})|g(x_i) - g(x_{i-1})|,$$

is the Riemann-Stieltjes integral of f. The limit is taken over any sequence of partitions η_n which $\sup_{x_i\in\eta_n} |x_{i+1} - x_i| \to 0$ as $n \to \infty$. This integral is applied to the sample paths of stochastic processes which are of bounded variation. However, for stochastic processes which are not of bounded variation such as Brownian motion, this concept of an integral is not defined. Processes such as Brownian motion do not have bounded variation however they do have bounded quadratic variation:

$$\sup_{\eta\in\mathbb{I}} \sum_{x_i\in\eta_n} |f(x_i) - f(x_{i-1})|^2. \tag{10.2}$$

To develop Ito's concept of a stochastic integral the idea of a martingale and the Doob-Meyer decomposition need to be introduced. This is a technical subject complicated enough to warrant a long exposition. Here the basic concepts behind the stochastic integral will be explained, and details can be found in Shreve [42].

Definition 10.1.6. A martingale is a filtration and an adapted stochastic process $(M_t, \mathcal{F}_t)$ which satisfies $\mathbb{E}(M_t|\mathcal{F}_s) = M_s$ for any $0 < s < t$ in the range of definition of

M_t. A *submartingale* satisfies $\mathbb{E}(M_t|\mathcal{F}_s) \geq M_s$ instead, and a *supermartingale* satisfies $\mathbb{E}(M_t|\mathcal{F}_s) \leq M_s$.

The space of *continuous* martingales, M_t, with finite second moments (or *quadratic variation*) $\mathbb{E}M_t^2 < \infty$ is denoted $L^2_{\mathbb{E}}$.

Definition 10.1.7 (Doob-Meyer decomposition)**.** [42, Thm. 1.4.10] If $X_t, \mathcal{F}_t$ is a right continuous submartingale with left limits, $X_0 = 0$ almost everywhere and X_t is of class D (a definition regarding stopping times [42, Def. 1.4.8]). Then X_t permits a decomposition

$$X_t = A_t + M_t, \tag{10.3}$$

where A_t is a natural [42, Def. 1.4.5] increasing process and M_t is a martingale.

By construction, a natural increasing stochastic process A_t can act as an integrator for pathwise Riemann-Stieltjes integrals,

$$I(X) = \int X_t dA_t \tag{10.4}$$

for measurable X_t. If X_t is right continuous with left limits (càdlàg) and adapted to the filtration of A_t, then $I(X)$ is also right continuous with left limits and adapted — provided it is finite [42, Rem. 1.4.6]

Definition 10.1.8. The quadratic variation of a process $X_t \in L^2_{\mathbb{E}}$ is denoted $\langle X \rangle_t$. Take the Doob-Meyer decomposition of X^2 and then the natural increasing process A_t from the decomposition is the variation of X_t, $\langle X \rangle_t = A_t$.

The function A_t is called the quadratic variation because the quadratic variation (as given in Equation 10.2) of X_t will converge to A_t in probability [42, Thm. 5.8].

The stochastic integral of $X_t \in L^2_{\mathbb{E}}$ with respect to a continuous martingale $M_t \in L^2_{\mathbb{E}}$

$$I(X) = \int_0^T X_t(\omega) dM_t(\omega) \tag{10.5}$$

is not defined as a pathwise Riemann-Stieltjes integral because M_t does not have bounded variation. However, it can be defined for simple processes.

10.1.1 The stochastic integral

This subsection defines the stochastic integral. To start, the integrator is specified by a martingale $Y_t \in L^2_{\mathbb{E}}$ and its filtration $\mathcal{F}_t$. With respect to this integrator, a measure is introduced,

$$\mu_Y(A) = \mathbb{E} \int_0^T \mathbb{I}_A(t,\omega) d\langle Y \rangle_t \tag{10.6}$$

where $A \in \mathcal{B}([0,T]) \otimes \mathcal{B}(\Omega))$. This measure is used to construct an L^2 space of the appropriate X_t using the norm,

$$[X]_T = \int_0^T X_t^2 d\langle Y \rangle_t \tag{10.7}$$

and this norm gives a equivalence relation upon which we build a representative space of processes. The subtleties of using equivalence classes rather than processes themselves does not matter in this work.

Definition 10.1.9. The space of valid integrands is denoted $\mathcal{L}^*$ and includes all càdlàg (right continuous with left limits) $\mathcal{F}_t$-measurable adapted processes, X_t such that $[X]_T < \infty$.

Definition 10.1.10. A simple process is a stochastic process made up of a finite number of random variables, ζ_k, and indicator functions,

$$S_t(\omega) = \sum_{k=1}^{n} \zeta_k(\omega)\ \mathbb{I}_{(t_{k+1},t_k]}(t) \tag{10.8}$$

where $0 < t_1 < ... < T$ partition the inteval. The random variables ζ_k must satisfy the condition $\sup_k |\zeta_k(\omega)| < C$ for some constant C and almost every ω. The process is intentionally defined to *stick out into the future* [52, p.29].

The simple process can be integrated with respect to Y_t provided each ζ_k is $\mathcal{F}_{t_k}$-measurable. The resulting sum is a martingale transform of Y_t,

$$I_T(S) = \int_0^T S_t dY_t = \sum_{k=1}^{n} \zeta_k (Y_{t_{k+1}} - Y_{t_k}) \tag{10.9}$$

where for simplicity it is assumed $t_n = T$. The sum is a martingale transform, meaning

the integral itself is a martingale. This is straightforward to prove, for the term in the sum $k = n$, $\mathbb{E}(\zeta_k(Y_{t_{k+1}} - Y_{t_k})|\mathcal{F}_n) = \zeta_k\left(\mathbb{E}(Y_{t_{k+1}}|\mathcal{F}_n) - \mathbb{E}(Y_{t_k}|\mathcal{F}_n)\right) = \zeta_k(Y_{t_k} - Y_{t_k}) = 0$. For earlier terms, note that $\mathbb{E}(Y_{n-1}|\mathcal{F}_n) = Y_{n-1}$ because the conditioning is on a larger σ-algebra than $\sigma(Y_{n-1})$, hence the sum is a martingale.

The collection of simple processes in which each ζ_k is $\mathcal{F}_{t_k}$-measurable is dense in $\mathcal{L}$ with respect to the $[\ ]_T$ norm [42, Lem. 3.2.7]. Thus the stochastic integral of $X \in \mathcal{L}$ with respect to the martingale $Y_t \in L^2_{\mathbb{E}}$ is defined.

Definition 10.1.11. [42, Def. 3.2.9] The integral of $X_t \in \mathcal{L}^*$ with respect to the square integrable continuous martingale integrator $Y_t \in L^2_{\mathbb{E}}$ is the martingale $(I_t(X_t), \mathcal{F}_t)$ which satisfies $\lim_{i\to\infty} \|I(S^i_t) - I(X_t)\| = 0$ for any sequence $S^{(i)}$ of simple processes such that $\lim_{i\to\infty}[S^{(i)} - X_t]_T = 0$.

In this thesis the only martingale considered is Brownian motion, Itô's Lemma is defined for any continuous semi-martingale though I restrict the theorem to continuous submartingales to avoid further definitions. There exists decompositions for semi-martingales analogous to the Doob-Meyer decomposition [42, Def. 3.3.1].

Lemma 10.1.12 (Itô's Lemma)**.** *For any continuous submartingale X_t with Doob-Meyer decomposition $X_t = M_t + A_t$, and any $f \in C^2([0,T])$ with bounded second derivative, the fundamental theorem of calculus is replaced with Itô's Lemma:*

$$f(X_t) = f(X_0) + \int_0^t f'(X_s)dM_s + \int_0^t f'(X_s)dA_s + \frac{1}{2}\int_0^t f''(X_s)d\langle M\rangle_s.$$

When integrating with respect to a brownian motion W_t the quadratic variation of W_t is equal to t, thus $\langle W\rangle_t = t$. Therefore, in the case of Brownian motion $X_t = W_t$ and $f(x) = x^2$, Itô's Lemma implies that

$$W_t^2 = \int_0^t 2W_s dW_s + t. \tag{10.10}$$

10.2 A stochastic differential equation

The differential equation for the evolution of the Frenet-Serret frame discussed in Chapter 5 can be found in Equation (9.11).

Equation (9.11) is a PDE with respect to the space variable x, while the parameter of a stochastic process in a stochastic differential equation (SDE) is colloquially referred to as 'time'. To avoid confusion, in this section we refer to x as s; whereas the time variable t is suppressed. Thus the differential equation (Equation 5.15) is reinterpreted stochastically. The rate of change of X_s is governed by the stochastic processes P and Q. Recall the polar decomposition $P + iQ = \kappa e^{i\sigma}$ where, $\kappa = \sqrt{P^2 + Q^2}$ and σ is such that $\tau = \frac{\partial \sigma}{\partial s}$. This implies $\sigma = \tan^{-1}\left(\frac{Q}{P}\right)$ and the rate of change of τ is a function of Wiener loop and therefore will be represented as a drift term in the stochastic differential equation

$$dX_s = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} X_s ds + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} X_s \circ d\sigma. \tag{10.11}$$

The $\circ$ denotes the Stratonovich form of stochastic differential equation (SDE). When passing from ordinary differential equations to stochastic ones there is a choice of interpretation. In the literature [58, Remark 3] it is purported that the interpretation should be as a Stratonovich type SDE [42]. Numerical methods based of Euler-Maruyama require Itô type SDEs [50], so a conversion is necessary.

Let $\mathcal{F}_s$ denote the joint filtration of the independent Brownian processes P and Q. An existence theorem for solutions to the SDE will be stated later in the section, it requires all coefficients of the SDE to be adapted to the filtration $\mathcal{F}_s$, and then solution X_s will necessarily be adapted to $\mathcal{F}_s$. Lie group based SDE's are well posed under these assumptions [65].

Definition 10.2.1. The differential form of Itô's Lemma says that for independent Brownian processes P and Q, if f is bounded, twice differentiable and has continuous second derivatives:

$$df(P,Q) = \frac{\partial f}{\partial P} dP + \frac{\partial f}{\partial Q} dQ + \frac{1}{2}\left(\frac{\partial^2 f}{\partial P^2} + \frac{\partial^2 f}{\partial Q^2}\right) ds. \tag{10.12}$$

Implicit in the definition is the understanding that f is adapted to the joint filtration $\mathcal{F}_s$ of P_s and Q_s as it only depends on s through P and Q.

Lemma 10.2.2. *Define* $\sigma_\epsilon(P,Q) := \tan^{-1}(\frac{PQ}{P^2+\epsilon^2})$ *as the regularised Itô integral of* τ.

The Itô differential can be written as

$$d\sigma_\epsilon = f_1(P,Q)dP + f_2(P,Q)dQ + f_3(P,Q)ds, \tag{10.13}$$

where

$$\begin{aligned}
f_1(P,Q) &:= \frac{(\epsilon^2 - P^2)Q}{(\epsilon^2+P^2)^2 + P^2Q^2}, \\
f_2(P,Q) &:= \frac{P(\epsilon^2+P^2)}{(\epsilon^2+P^2)^2 + P^2Q^2}, \\
f_3(P,Q) &:= -\frac{2P^3Q(\epsilon^2+P^2)}{((\epsilon^2+P^2)^2+P^2Q^2)^2} - \frac{2PQ\left((\epsilon^2+P^2)^2+P^2Q^2\right)}{\left((\epsilon^2+P^2)^2+P^2Q^2\right)^2} \\
&\quad - \frac{(\epsilon^2-P^2)Q\left(2PQ^2+4P(\epsilon^2+P^2)\right)}{\left((\epsilon^2+P^2)^2+P^2Q^2\right)^2}
\end{aligned}$$

The stochastic differential equation for X_s given in Equation (10.12) can be updated using Lemma 10.2.2,

$$dX_s = \mathbf{D}X_s ds + \mathbf{B}X_s \circ dP + \mathbf{C}X_s \circ dQ. \tag{10.14}$$

Where,

$$\begin{aligned}
\mathbf{D} &= \begin{bmatrix} 0 & \sqrt{P^2+Q^2} & 0 \\ -\sqrt{P^2+Q^2} & 0 & f_3(P,Q) \\ 0 & -f_3(P,Q) & 0 \end{bmatrix}, \\
\mathbf{B} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & f_1(P,Q) \\ 0 & -f_1(P,Q) & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & f_2(P,Q) \\ 0 & -f_2(P,Q) & 0 \end{bmatrix}.
\end{aligned} \tag{10.15}$$

This differential equation is in the form of a Stratonovich SDE and not an Itô SDE. To convert from a Stratonovich SDE to a Itô SDE a correction term is introduced.

Theorem 10.2.1. *[43, p.159] If X_s is a strong solution to the Stratonovich SDE,*

$$dX_s = f(X_s)ds + g(X_s) \circ dW_s$$

where g and f are linear in X, and adapted to the filtration of W_s. Then X_s is also the solution to the equivalent Itô SDE,

$$dX_s = \left(f(X) + \frac{1}{2} \frac{\partial g(X)}{\partial X} g(X) \right) ds + g(X) dW.$$

The requirements for a stochastic differential equation to have a strong solution are outlined in Theorem 10.3.1. See Ref. [42, Prop. 2.21] for the precise meaning of equivalent in this case. More generally the functions f and g can be taken to be Lipschitz and twice bounded differentiable respectively. Applying the theorem one can convert Equation (10.12) from a Stranovich SDE into an Itô SDE as follows:

$$dX_s = \mathbf{A} X_s ds + \mathbf{B} X_s dP + \mathbf{C} X_s dQ \tag{10.16}$$

where,

$$\mathbf{A} = \begin{bmatrix} 0 & \sqrt{P^2+Q^2} & 0 \\ -\sqrt{P^2+Q^2} & \frac{1}{2} f_1^2(P,Q) + \frac{1}{2} f_2^2(P,Q) & f_3(P,Q) \\ 0 & -f_3(P,Q) & \frac{1}{2} f_1^2(P,Q) + \frac{1}{2} f_2^2(P,Q) \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & f_1(P,Q) \\ 0 & -f_1(P,Q) & 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & f_2(P,Q) \\ 0 & -f_2(P,Q) & 0 \end{bmatrix}. \tag{10.17}$$

As justified in the discussion about the Gibbs measure, when $\beta = 0$ the stochastic processes P and Q are each a Brownian bridge with period $T = 2\pi$, thus they can be expressed in terms of Brownian motions W_1 and W_2; that is,

$$P(s) = W_1(s) - \frac{s\, W_1(2\pi)}{2\pi}. \quad Q(s) = W_2(s) - \frac{s\, W_2(2\pi)}{2\pi} \tag{10.18}$$

The differentials can then be expressed as $dP = dW_1 - W_1(2\pi)/2\pi\, ds$ and equivalently for Q. Equation (10.16) is now written as a standard Itô SDE,

$$dX_s = \left(\mathbf{A} + \frac{W_1(2\pi)}{2\pi} \mathbf{B} + \frac{W_2(2\pi)}{2\pi} \mathbf{C} \right) X_s ds + \mathbf{B} X_s dW_1 + \mathbf{C} X_s dW_2 \tag{10.19}$$

where **A**, **B** and **C** are defined as in Equation (10.17). This is the SDE of interest in this chapter.

10.3 The Magnus expansion for SDEs

The genesis of the SDE (Equation (10.19)) was the Frenet-Serret frame, a matrix in the Lie algebra $\mathfrak{so}(3)$, and thus if $X_0 \in SO(3)$, then so did X_s for all s. This should be true of the SDE too. In Section 2.3.3 we discuss dealing with multivariable differential equations which have a solution in terms of a Magnus expansion. Section 2.3.2 shows the difficulty in calculating the derivative of the exponential map between the Lie algebra and its Lie group. Lemma 2.3.23 gives an expression for conversion of the matrix differential equation for X_s into one for Ω_s, where

$$X_s = \exp(\Omega(s))X_0,$$

for some initial condition X_0. The added complexity is product of working with a SDE rather than an ODE. That said, the problem is significantly more straight forward than for solutions within a general Lie group, as the exponential map from $\mathfrak{so}(3) \to SO(3)$ is surjective. This means the consistency problem between the evolution of Brownian processes on different coordinate charts does not need to be addressed.

The intricacies of Itô calculus are beyond the breadth of this thesis, but a simplified rule for working with differentials of Brownian motions W_i, W_j is,

$$\begin{aligned} dW_i dW_j &= \delta_{ij} dt, \\ dW_i dt = 0, \quad dt dt &= 0. \end{aligned}$$

Using these rules, along with Itôs lemma, we can establish the differential equation Ω would have to follow if X satisfied a general SDE in just one dimension.

A standard existence and uniqueness theorem for solutions to a stochastic differential equation of the form,

$$dX_s = a(s, X_s)ds + b(s, X_s)dW_1, \tag{10.20}$$

with initial value X_0 can be commonly found [43, 42]. A weak solution to an SDE is a process which satisfies the ODE for a specific choice of Wiener process W_1, a

second solution for a different Wiener process would have the same finite dimensional probability distributions (they are equivalent) but different sample paths. A strong solution to an SDE, X, is a process which for an equivalent Y,

$$\mathbb{P}(\sup_{0\leq t\leq T}|X_t - Y_t| > 0) = 0.$$

that is solutions are pathwise unique given a suitable initial condition. The conditions under which a weak solution exists for Equation (10.20) can be found in books on SDEs [43, Thm. 4.7.1].

Theorem 10.3.1. *[43, Thm 4.5.3] A strong solution to Equation* (10.20) *exists if,*

- *The functions $a(s,x)$ and $b(s,x)$ are jointly measurable as functions $([0,T],\mathcal{B}([0,T]))\times(\mathbb{R},\mathcal{B}(\mathbb{R}))\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$.*
- *The functions $a(s,x)$ and $b(s,x)$ are both Lipschitz in x.*
- *The functions $a(s,x)$ and $b(s,x)$ satisfy a growth inequality, for some $K>0$*

$$\|a(s,x)\|^2 \leq K^2(1+\|x\|^2)$$
$$\|b(s,x)\|^2 \leq K^2(1+\|x\|^2)$$

  for all $(s,x)\in[0,T]\times\mathbb{R}$.

- *The initial value, $X_0 = Z$, is a random variable which is $\mathcal{F}_0$ measurable and finite second moment.*

The existence and uniqueness of higher dimensional problems, when $x\in\mathbb{R}^n$ or a Riemannian manifold hold when the relevant adaptations are made to the norms.

In addition, an SDE of the form,

$$dX_s = a(s,X,W_s^1,W_s^2,\ldots,W_s^m)ds + \sum_{i=1}^{m} b_i(s,X,W_s^1,W_s^2,\ldots,W_s^m)dW_s^i \qquad (10.21)$$

can be reduced to a vector valued SDE of the form of Equation (10.20), simply let $\tilde{X}_s = [X_s, W_s^1, \ldots, W_s^m]$ and then

$$d\tilde{X}_s = \begin{bmatrix} a(s,\tilde{X}) \\ 0 \\ \vdots \\ 0 \end{bmatrix} ds + \sum_{i=1}^{m} \begin{bmatrix} b_i(s,\tilde{X}) \\ \mathbf{e}_i \end{bmatrix} dW_s^i, \tag{10.22}$$

where $\mathbf{e}_i$ is a m dimensional unit vector with 1 in the i^{th} position. Redefining new $\tilde{a}$ and $\tilde{b}$ will give an equation of the form of Equation (10.20).

Lemma 10.3.1. *If the one dimensional function $X(s) = \exp(\Omega(s))X_0$ is a solution to*

$$dX_s = g(s)X_s ds + h(s)X_s dW_1 + k(s)X_s dW_2$$

for independent Brownian's W_1 and W_2, where $g(s), h(s), k(s)$ are adapted to their joint filtration. Then Ω satisfies the SDE,

$$d\Omega_s = \left(g(s) - \frac{1}{2}(h(s)^2 + k(s)^2) \right) ds + h(s)dW_1 + k(s)dW_2 \tag{10.23}$$

Proof. Consider a function of Ω, $f(\Omega) = \exp(\Omega(s))$, so $X_s = f(\Omega_s)X_0$. Then assume X_s satisfies the assumption of the lemma, and let $d\Omega = l(s)ds + h(s)dW_1 + k(s)dW_2$ for some yet undefined functions $l(s), h(s), k(s)$. The differential of f in the sense of Itô is,

$$\begin{aligned} df(\Omega) &= f'(\Omega(s))d\Omega + \frac{1}{2}f''(\Omega(s))(d\Omega)^2, \\ &= f' \left(l(s)ds + h(s)dW_1 + k(s)dW_2 \right) + \\ &\qquad + \frac{1}{2}f'' \left(h(s)^2 ds + k(s)^2 ds \right), \end{aligned}$$

The derivatives with respect to Ω do not change the function $f(\Omega)$,

$$= \left(l(s) + \frac{1}{2}(h(s)^2 + k(s)^2) \right) f ds + h(s) f dW_1 + k(s) f dW_2.$$

Finally, matching up coefficients for $dX_s = df(\Omega_s)$ one shows that the function $l(s)$ must be $l(s) = g(s) + \frac{1}{2}(h(s)^2 + k(s)^2)$. □

For matrix valued coefficients the problem is more involved. It is advised to view the matrix as a multivariable problem. Let X_s denote a real matrix valued stochastic process, which is measurable and adapted to the joint filtration $\mathcal{F}_t$ of Brownian loops P and Q. In addition let X_s be the solution to

$$dX_s = \mathbf{A}X_s ds + \mathbf{B}X_s dW_1 + \mathbf{C}X_s dW_2$$

for independent Brownian's W_1 and W_2. The matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are given in Equation (10.17). They have had their dependencies suppressed for visual clarity, but depend on both P and Q, which vary over time.

Lemma 10.3.2. *If X_s satisfies the assumptions of the previous paragraph, the Magnus expansion $X_s = \exp(\Omega_s)X_0$ implies the exponent Ω_s satisfies the SDE,*

$$d\Omega_s = \mathbf{L}ds + \mathbf{M}dW_1 + \mathbf{N}dW_2 \tag{10.24}$$

where the matrices $\mathbf{L}, \mathbf{M}, \mathbf{N}$ again have had their dependencies suppressed, but are given by

$$\begin{aligned} \mathbf{M} &= \mathcal{E}_{-\Omega}(\mathbf{B}) \qquad \mathbf{N} = \mathcal{E}_{-\Omega}(\mathbf{C}) \\ \mathbf{L} &= \mathcal{E}_{-\Omega}\left(\mathbf{A} - \frac{1}{2}(\mathbf{B}^2 + \mathbf{C}^2 + F(\mathbf{M}) + F(\mathbf{N}))\right). \end{aligned}$$

The inverse of the partial derivative of the exponential map is defined explicitly in terms of its power series in Lemma (2.3.22). Although $\mathcal{E}_{-\Omega}$ is only well defined as an inverse if the argument is in $\mathfrak{so}(3)$, one can still interpret it as a formal power series for other matrices. The function $F(\mathbf{M})$ is defined within the proof.

Proof. Consider the function $f(\Omega) = \exp(\Omega(s))X_0$ and assume that the SDE for Ω_s takes the form $d\Omega_s = \mathbf{L}ds + \mathbf{M}dW_1 + \mathbf{N}dW_2$ for some matrices $\mathbf{L}, \mathbf{M}, \mathbf{N}$. Consider the differential of f in the sense of Itô by taking the Taylor expansion about each component.

$$df(\Omega) = \sum_{ij} \frac{\partial f}{\partial \Omega_{ij}} d\Omega_{ij} + \frac{1}{2}\sum_{ij}\sum_{kl} \frac{\partial}{\partial \Omega_{kl}}\frac{\partial f}{\partial \Omega_{ij}} d\Omega_{kl} d\Omega_{ij}. \tag{10.25}$$

The derivative of the exponential of a matrix is expressed in Equation (2.26). Denote

the matrix of zeros with a unit i, j component by E_{ij} and note that

$$\frac{\partial \exp(\Omega)}{\partial \Omega_{ij}} = \mathcal{D}_{-\Omega}\left(\frac{\partial \Omega}{\partial \Omega_{ij}}\right)\exp(\Omega) = \mathcal{D}_{-\Omega}(E_{ij})\exp(\Omega). \tag{10.26}$$

By the assumption on the form of $d\Omega$, the components of the SDE are $d\Omega_{ij} = \mathbf{L}_{ij}ds + \mathbf{M}_{ij}dW_1 + \mathbf{N}_{ij}dW_2$. The directional derivative of the exponential $\mathcal{D}_{-\Omega}(E_{ij})$, discussed explicitly in Section 2.3.2, is linear in its second argument so for any matrix $\mathbf{M}$, $\sum_{ij}\mathbf{M}_{ij}\mathcal{D}_{-\Omega}(E_{ij}) = \mathcal{D}_{-\Omega}(\mathbf{M})$. Thus the first term of Equation (10.25) is,

$$\begin{aligned}\sum_{ij}\frac{\partial f}{\partial \Omega_{ij}}d\Omega_{ij} &= \sum_{ij}\mathcal{D}_{-\Omega}(E_{ij})\left(\mathbf{L}_{ij}ds + \mathbf{M}_{ij}dW_1 + \mathbf{N}_{ij}dW_2\right)\exp(\Omega)\\ &= \mathcal{D}_{-\Omega}(\mathbf{L})f(\Omega)ds + \mathcal{D}_{-\Omega}(\mathbf{M})f(\Omega)dW_1 + \mathcal{D}_{-\Omega}(\mathbf{N})f(\Omega)dW_2.\end{aligned}$$

Taking into consideration the rules of Itô calculus, the product $d\Omega_{ij}d\Omega_{kl} = (B_{ij}B_{kl} + C_{ij}C_{kl})\,ds$. The second order partial derivatives of the exponential of the matrix Ω are,

$$\frac{\partial}{\partial \Omega_{kl}}\frac{\partial f}{\partial \Omega_{ij}} = \mathcal{D}_{-\Omega}(E_{kl})\mathcal{D}_{-\Omega}(E_{ij})\exp(\Omega) + \frac{\partial}{\partial \Omega_{kl}}\left(\mathcal{D}_{-\Omega}(E_{ij})\right)\exp(\Omega). \tag{10.27}$$

Details of the derivative of $\mathcal{D}$ can be found in the Appendix of [50], and uses the power series expansion given in Equation (2.26). The second term of Equation (10.25) is given by,

$$\sum_{ij}\sum_{kl}\frac{\partial}{\partial \Omega_{kl}}\frac{\partial f}{\partial \Omega_{ij}}d\Omega_{kl}d\Omega_{ij} = \left[\left(\mathcal{D}_{-\Omega}(\mathbf{M})\right)^2 + \left(\mathcal{E}_{-\Omega}(\mathbf{N})\right)^2 + F(\mathbf{M}) + F(\mathbf{N})\right]f(\Omega)ds.$$

Where the function F is given by,

$$F(\mathbf{M}) = \sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\frac{1}{p+q+2}\frac{(-1)^p}{(p!(q+1)!}\mathrm{ad}_{\Omega}^{p}(\mathrm{ad}_{\mathbf{M}}(\mathrm{ad}_{\Omega}^{q}(\mathbf{M}))).$$

The stochastic differential equation for $f(\Omega)$ can now be stated as,

$$df(\Omega) = \left[\mathcal{D}_{-\Omega}(\mathbf{L}) + \frac{1}{2}\left((\mathcal{D}_{-\Omega}(\mathbf{M}))^2 + (\mathcal{D}_{-\Omega}(\mathbf{N}))^2 + F(\mathbf{M}) + F(\mathbf{N})\right)\right] f(\Omega)ds$$
$$+ \mathcal{D}_{-\Omega}(\mathbf{M})f(\Omega)dW_1 + \mathcal{D}_{-\Omega}(\mathbf{N})f(\Omega)dW_2.$$

Equating $f(\Omega) = X$ the coefficients of $dX_s = \mathbf{A}X_s ds + \mathbf{B}X_s dW_1 + \mathbf{C}X_s dW_2$ can be compared with that of $df(\Omega)$. The expression $\mathcal{D}_\Omega$ has a well defined inverse $\mathcal{E}_\Omega$, and so,

$$\begin{aligned}
\mathbf{M} &= \mathcal{E}_{-\Omega}(\mathbf{B}), \\
\mathbf{N} &= \mathcal{E}_{-\Omega}(\mathbf{C}), \\
\mathbf{L} &= \mathcal{E}_{-\Omega}\left(\mathbf{A} - \frac{1}{2}(\mathbf{B}^2 + \mathbf{C}^2 + F(\mathbf{M}) + F(\mathbf{N}))\right).
\end{aligned}$$

□

10.4 The proposed numerical algorithm

The numerical method of choice for solving Equation (10.19) was developed by Pigott, Solo et al. [58], [50], and will be motivated and justified below. The method is a one step Euler Maruyama approximation done in the Lie algebra. It uses the derivative of the exponential map discussed in Section 2.3.2, though the power series for $\mathcal{E}_X$ (Lemma 2.3.22) is truncated heavily.

The stochastic differential equation for the first Frenet-Serret matrix with $\beta = 0$ is given in Equation (10.11). This is converted into an SDE in the Lie algebra by Lemma 10.3.2. As a result, the differential equation simulated in the Lie algebra is,

$$d\Omega_s = \mathbf{L}ds + \mathcal{E}_{-\Omega}(\mathbf{B})dP + \mathcal{E}_{-\Omega}(\mathbf{C})dQ, \tag{10.28}$$

Where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are the matrices as defined in Equation (10.17), and $\mathbf{L}$ is defined in terms of those matrices as

$$\mathbf{L} = \mathcal{E}_{-\Omega}\left(\mathbf{A} - \frac{1}{2}(\mathbf{B}^2 + \mathbf{C}^2 + F(\mathcal{E}_{-\Omega}(\mathbf{B})) + F(\mathcal{E}_{-\Omega}(\mathbf{C})))\right). \tag{10.29}$$

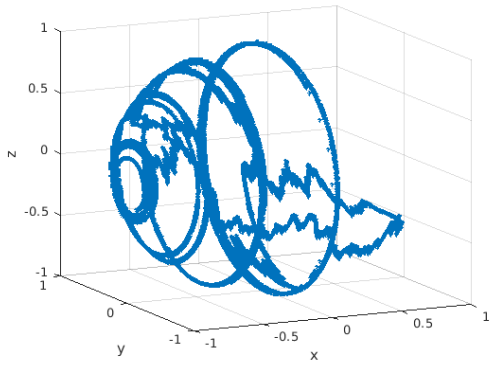

Figure 10.1: The figure demonstrates a sample path of the stochastic process X_s, which is a solution to Equation (10.19). As $X_s \in SO(3)$ the path is visualised as the action of X_s applied to a unit vector in $\mathbb{R}^3$. The numerical solution shown is for $s \in [0, 10]$, and has a step size of $h = 10^{-5}$.

The one step method takes $\Omega_s^{(0)} = \mathbf{0}$ the zero matrix at each timestep s, and calculates $\Omega_s^{(1)}$ by

$$\Omega_s^{(1)} = \Omega_s^{(0)} + h\mathbf{L} + \sqrt{h}\delta\mathcal{E}_{-\Omega_0}(\mathbf{B}) + \sqrt{h}\eta\mathcal{E}_{-\Omega_0}(\mathbf{C}), \tag{10.30}$$

where ϵ and η are distributed $N(0, 1)$. Therefore $P_{t+h} - P_t \sim \sqrt{h}\delta$ and equivalently for Q. For any matrix $\mathbf{M}$, $\mathcal{D}_{\mathbf{0}}(\mathbf{M}) = \mathbf{M}$ by definition. The same holds for $F(\mathbf{M}) = \frac{1}{2}[\mathbf{M}, \mathbf{M}] = 0$ for the first step as $\Omega_0 := \mathbf{0}$. Thus the one step algorithm further reduces to

$$\Omega_s^{(1)} = h\mathbf{A} - \frac{1}{2}(\mathbf{B}^2 + \mathbf{C}^2) + \sqrt{h}\delta\mathbf{B} + \sqrt{h}\eta\mathbf{C}, \tag{10.31}$$

$$= h\mathbf{D} + \sqrt{h}\delta\mathbf{B} + \sqrt{h}\eta\mathbf{C}, \tag{10.32}$$

where the matrix $\mathbf{D}$ is given in Equation (10.14) and each matrix $\mathbf{D}, \mathbf{B}, \mathbf{C}$ belong to $\mathfrak{so}(3)$ and depend on s. The final stage of the algorithm is to use the update in the Lie algebra to update the original function X. That is, $X_{n+1} = \exp(\Omega_1)X_n$.

Note that in Algorithm 10.1 it is assumed that $2\pi < T < 4\pi$ for the construction of periodic Brownian motion.

The resulting stochastic process $X_s \in SO(3)$ is then used to rotate the unit vector $y_0 = [0, 0, 1]^\top$ on $\mathbb{S}^2$ to $y_s = X_s y_0$, the third column of X_s. The sample paths of this process can be described by construction of a frame $\{y_s, y_s', y_s \times y_s'\}$. Figure 10.1

Algorithm 10.1 Algorithm for one sample path

Define h a constant timestep increment, and T the period to solve for
`no_timesteps` $\leftarrow T/h$
Define ϵ, a small regularisation constant.
$X_0 \leftarrow I$ initial condition.
Sample $\gamma_i \sim N(0,1)$ and $\tilde{\gamma}_i \sim N(0,1)$ for $i = 1, ..., $ `no_timesteps`.
$\eta \leftarrow \left[\gamma_1, \gamma_2, ..., \gamma_{T/h}\right]$
$\delta \leftarrow \left[\tilde{\gamma}_1, \tilde{\gamma}_2, ..., \tilde{\gamma}_{T/h}\right]$
$W_1 \leftarrow [W_1(h), W_1(2h), ..., W_1(T)]$ where $W_1(nh) = \sqrt{h}\sum_{j=1}^{n}\gamma_j$.
$W_2 \leftarrow [W_2(h), W_2(2h), ..., W_1(T)]$ where $W_2(nh) = \sqrt{h}\sum_{j=1}^{n}\tilde{\gamma}_j$.
$P_j \leftarrow \left[P_j(h), P_j(2h), ..., P_j(\frac{2\pi}{h})\right]$ where for $j = 1, 2$

$$P_j(nh) = W_1(nh + 2\pi(j-1)) - \frac{W_1(2\pi j)\, nh}{2\pi}$$

$Q_j \leftarrow \left[Q_j(h), Q_j(2h), ..., Q_j(\frac{2\pi}{h})\right]$ where for $j = 1, 2$

$$Q_j(nh) = W_2(nh + 2\pi(j-1)) - \frac{W_2(2\pi j)\, nh}{2\pi}$$

Concatenate P_1 with P_2 so the array has T/h elements. Repeat for Q.
The following operations are vectorised, each object is an array of T/h components.
$\kappa \leftarrow \sqrt{P^2 + Q^2}$ ▷ κ will be a $1 \times T/h$ array
$f_i \leftarrow f_i(P, Q)$ for $i = 1, 2, 3$ and f_i given in Eq. (10.13)
$\mathbf{D}, \mathbf{B}, \mathbf{C}$ defined as in Equation (10.14) using f_i above ▷ They are $3 \times 3 \times T/h$ arrays
$\Omega \leftarrow h\mathbf{D} + \sqrt{h}\delta\mathbf{B} + \sqrt{h}\eta\mathbf{C}$
$\|\omega\| \leftarrow \sqrt{\Omega_{21}^2 + \Omega_{31}^2 + \Omega_{32}^2}$
$\exp(\Omega) \leftarrow I + \frac{\sin(h\|\omega\|)}{\|\omega\|}\Omega + \frac{1-\cos(h\|\omega\|)}{\|\omega\|^2}\Omega^2$
$X_i \leftarrow \prod_{j=1}^{i}\exp(\Omega)_j X_0$ ▷ $\exp(\Omega)_j$ represents the j'th element of the array.
The $\prod$ operator here acts as:

$$\prod_{j=1}^{i}\exp(\Omega)_j = \exp(\Omega)_i ... \exp(\Omega)_2 \exp(\Omega)_1$$

demonstrates a sample-path of y_s generated via Algorithm 10.1. The code used to simulate a sample path is available [44].

The sample paths start off on the great circle perpendicular to the y-axis, and so have constant binormal $y_s \times y_s'$. As a sample path extends past the great circle, the binormal vector at each point deviates slowly; as a consequence a sample path can be thought of as a precessing orbit.

10.5 Computational complexity

The purpose of this simulation is to observe the distribution of the stochastic process which solves the nonlinear Schrödinger equation. Many sample paths are required to approximate the distribution, hence the choice of a first order numerical scheme is needed to limit the computational complexity. Even the scheme outlined in Section 10.4 requires an order of $\mathcal{O}(T/h)$ matrix multiplications. For further insight into the constant on the T/h term, each matrix multiplication of $M \in M_3(\mathbb{R})$ involves $3^2(3+2) = 45$ individual operations, not to mention the additional operations involved in calculating each matrix using Rodriguez' formula. The algorithm scales with number of sample paths simulated, N, as $\mathcal{O}(N)$. As such, the limits of the computing power on hand meant, for $T = 10$, a timestep of $h = 10^{-5}$ and $N = 10^6$ sample paths.

The algorithm can be parallelised easily —using a machine equipped with an 8-core Intel Xeon Gold 6248R CPU with a clock speed of 2993 Mhz; with $h = 10^{-5}$ and $N = 10^6$ the run time of the simulation was 1 week.

10.6 Error estimation

The following is taken from Piggott [59], and establishes the *almost* linear convergence of error in expectation of the L^2 norm. The numerical method produces a step function

$$\hat{y}_h(s) = \prod_{j=1}^{j=\max_n\{nh\leq s\}} \exp(\Omega_j) X_0 y_0 \tag{10.33}$$

The step function $\hat{y}_h(s)$ converges to the solution $y(s)$ in the L^2 space of Itô processes

as the step size $h \to 0$,

$$\mathbb{E}\left[\sup_{0\le s<T} \|y_s - \hat{y}_{s,h}\|^2_{\mathbb{R}^3}\right] = \mathcal{O}(h^{1-\varepsilon}), \tag{10.34}$$

for some small $\varepsilon > 0$ (See Piggott [59]). The derivation of this estimate is given in [59, Thm 3.1], and here the scale of the constants involved are investigated to determine if they can be controlled for feasible step sizes h. Let $\varepsilon = \frac{1}{r}$, the dominating term in the $\mathcal{O}(h^{1-\frac{1}{r}})$ expansion is given in [59, Equation (25)] by $(12 + 4T)TD_r e^{4T(2+T)}$. Naturally, the size of the interval, T plays a large role in the estimate. The constant $D_r = 1 + \frac{3^r T}{r-1} C_r$ and the constant C_r has the most effect controlling the $\mathcal{O}(h^{1-\varepsilon})$ term, it is of the form

$$\mathcal{C}_r = \frac{2^r \Gamma(r + \frac{1}{2}) h^{1-\frac{1}{r}}}{\sqrt{\pi}}. \tag{10.35}$$

Take the logarithm of the function and apply Stirling's approximation [71, p.151] to approximate the logarithm of the gamma function,

$$\begin{aligned}
\log\left(\frac{2^r \Gamma(r + \frac{1}{2}) h^{1-\frac{1}{r}}}{\sqrt{\pi}}\right) &= r\log 2 - \frac{1}{2}\log\pi + \left(1 - \frac{1}{r}\right)\log h + \log\Gamma\left(r + \frac{1}{2}\right) \\
&= r\log 2 - \frac{1}{2}\log\pi + \left(1 - \frac{1}{r}\right)\log h \\
&\quad + r\log r - r + \frac{1}{2}\log 2\pi + C,
\end{aligned}$$

for some constant C. The derivative of this expression will give the location of a minima,

$$\frac{\partial}{\partial r}\log\left(\frac{2^r \Gamma(r - \frac{1}{2}) h^{1-\frac{1}{r}}}{\sqrt{\pi}}\right) = \log 2 + \frac{1}{r^2}\log h + \log r.$$

Due to computational constraints, a value of $h = 10^{-5}$ is the smallest step size feasible to take (See Section 10.5). Therefore, this expression changes sign between 4 and 5. Further the second derivative is $-\frac{1}{r^3}\log h + \frac{1}{r} > 0$ for all positive r implying the stationary point is a minima. Taking $r = 4$ gives an upper bound on the constant of $\mathcal{C}_4 < 0.05$ showing it is possible to control the constants on the interval under consideration.

10.7 Hypothesis testing

The algorithm laid out in Algorithm 10.1 was run $N = 10^6$ times. The resulting stochastic process $X_s \in SO(3)$ was then used to rotate the unit vector $y_0 = [0, 0, 1]^\top$ on $\mathbb{S}^2$ to $y_s = X_s y_0$, the third column of X_s. For each sample path of the new stochastic process $y_s(\omega_i) \in S^2$ the position at times $s = 0.3, 0.6, \ldots, 6.0$ were recorded by their spherical coordinates of longitude $\theta_s \in [-\pi, \pi)$ and colatitude $\phi_s \in [0, \pi]$. The data was collected for the purpose of verifying that the joint distribution of (θ_s, ϕ_s) becomes independent and uniform over the sphere, a plausible suggestion from looking at the histogram in Figure 10.2. To this end the following hypotheses are made, for $s = 0.3, 0.6, \ldots, 6.0$.

H_0^s: The distribution of $\theta_s(\omega_i)$ and $\phi_s(\omega_i)$ are statistically independent at timestep s.

H_1^s: The distribution of $\theta_s(\omega_i)$ and $\phi_s(\omega_i)$ are dependent at time s.

The symbol H_0 denotes the null hypothesis, and H_1 is the alternative hypothesis. We perform a nonparametric independence test based on the Hilbert Schmidt Independence criterion (HSIC) measure of dependence [31]. In addition, under the assumption that the distribution of the marginals are independent at the later time $s = 10$, we make the following hypotheses,

H_0^θ: The longitude θ_s is distributed as a uniform distribution, against H_1^θ that θ_s follows a different distribution.

H_0^ϕ: The latitude ϕ_s follows a sine distribution (See Equation (10.40)), against the alternative H_1^ϕ that ϕ_s follows a different distribution.

The significance level of a statistical test is denoted α and represents a bound for the acceptable probability of a type I error. A type I error is if the statistical test was to reject the null hypothesis out of hand (incorrectly). One can quantify this probability using simple set theory and an index function,

$$I(H_0^s) = \begin{cases} 1 & \text{if } H_0 \text{ is rejected incorrectly,} \\ 0 & \text{if } H_0 \text{ is rejected correctly.} \end{cases}$$

The probability of a type I error can then be defined as $\mathbb{P}(I(H_0^s) = 1)$. For multiple tests the concept is extended to the probability of a nonzero family error rate. In other words, the chance that for at least one test a type I error occurred. A family consists of "obviously related group of observations collected from the same experiment whose statistical analysis falls into a singe mathematical framework" [54, p.34]. As such the hypotheses under consideration fall into two families split by the type of null hypothesis.

To evaluate the family-wide probability of a type I error consider the index set for our problem $\mathfrak{N} = \{0.3, 0.6, \ldots, 6\}$ consisting of 20 elements. The probability of a nonzero family error rate is $\mathbb{P}\left(\bigcup_{s\in\mathfrak{N}}(I(H_0^s) > 0)\right)$ and a bound on this probability,

$$\mathbb{P}\left(\bigcup_{s\in\mathfrak{N}}(I(H_0^s) = 1)\right) \leq \sum_{s\in\mathfrak{N}} \mathbb{P}(I(H_0^s) = 1), \tag{10.36}$$

follows from Boole's inequality (subadditivity). This inequality is known as the *Bonferroni inequality* [54, p.8] and provides a useful bound to ensure a specific acceptable α. Let α_s denote the significance level for each test H^s for $s \in \mathfrak{N}$, $\mathbb{P}(I(H_0^s) = 1) = \alpha_s$. Then Equation (10.36) implies that for N tests, setting $\alpha_s = \alpha/N$ will produce an overall significance level of α.

To be 99% sure that the statistical tests carried out do not give a false positive — rejecting the null hypotheses out of hand — the overall significance level is set to $\alpha = 0.01$ for each family of tests. To achieve this it is required that each of the 20 tests for $\{H_0^s \ : \ s \in \mathfrak{N}\}$ have a significance level of $\alpha_s = 0.01/20$, and so each test has a far stricter significance level than $\alpha_s = 0.01$. And for the second family $\{H_0^\theta, H_0^\phi\}$ the α values are each 0.005.

10.7.1 Independence of latitude and longitude marginals.

The statistical test chosen to study the interdependence of the two distributions θ_s and ϕ_s over time is based on the Hilbert-Schmidt Independence Criterion [31], a well known nonparametric independence test. The implementation used is thanks to Jitkittum [40]. The results are displayed in the following chart,

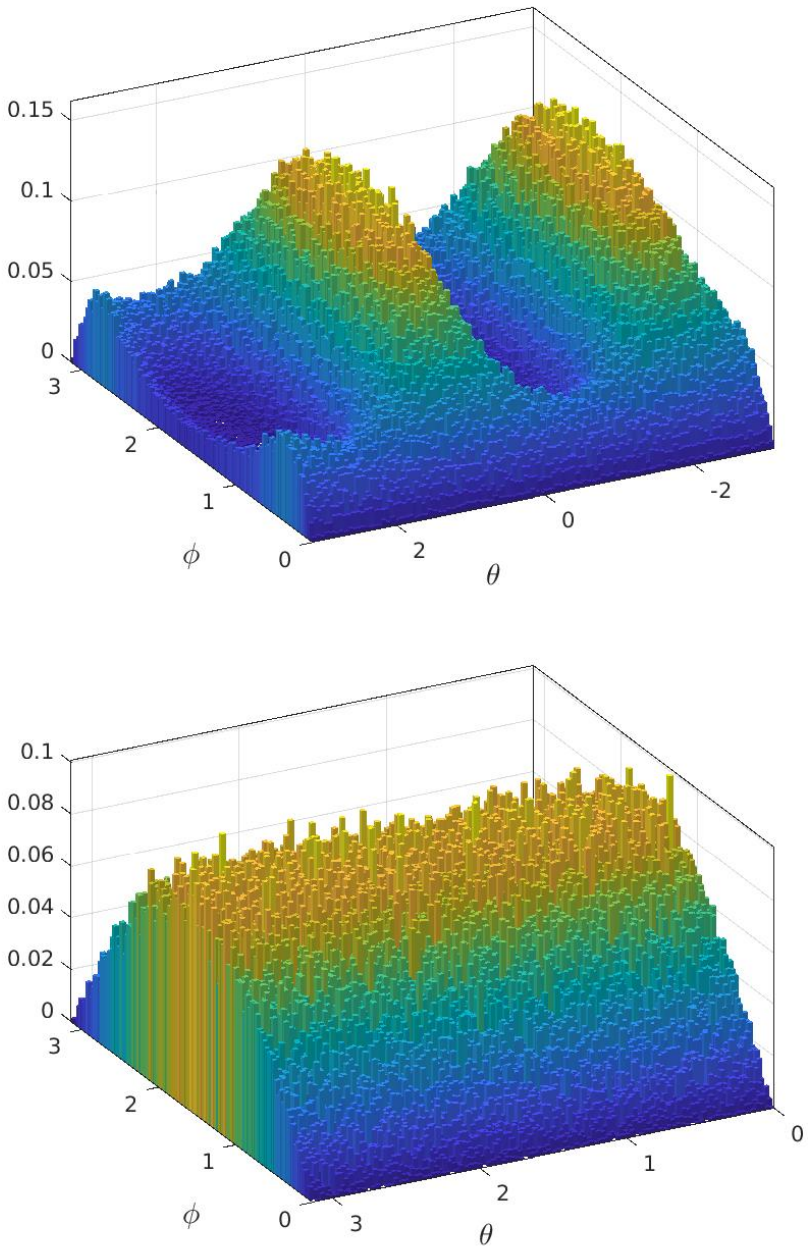

Figure 10.2: Histograms of the joint distribution of the third column of X_s at two timesteps, $s = 1$ (**left**) and $s = 10$ (**right**). The distribution lies on the sphere $\mathbb{S}^2$ and thus the axes are chosen as the longitude θ_s and colatitude ϕ_s. With respect to these axes the marginals of the distribution become more independent over time.

timestep (s)	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4
rejects H_0^s	T	T	T	T	T	T	T	T
timestep (s)	2.7	3.0	3.3	3.6	3.9	4.2	4.5	4.8
rejects H_0^s	F	F	F	F	F	F	F	F
timestep (s)	5.1	5.4	5.7	6.0				
rejects H_0^s	F	F	F	F				

With significance level 0.01 it is observed that from $s = 2.7$ the test is able to reject the null hypothesis H_0^s and support the claim that θ_s and ϕ_s evolve to be independent.

Although the significance levels for the test were defined stringently apriori, it is worth noting the relevance of this value after running the experiment. The chance of incorrectly rejecting the null hypothesis is in other words the chance that the test identified dependence between the distributions where there is in fact none. Looking at the result of the test this gives a method of quantifying the chance that before $s = 2.7$ the distributions were already independent, we can be 99% certain that they were not. It doesnt give any indication of the chance that no type II error occured in the interval $2.7-6$ (that the test failed to identify that the distributions were dependent here). The HSIC test as implemented has a maximum chance of type II error, β depending on the choice of α made. The probability that each timestep between 2.7 and 6 suffered from a type II error is β^{11}, the HSIC authors estimate [40] a value of $\beta = 0.2$ which would make β^{11} around 4×10^{-9}. This supports the claim that the θ_s and ϕ_s distributions start off dependent and become independent within the interval 2.7 to 6.

10.7.2 Wasserstein distance between measures

By reference to previous work [8], the Wasserstein distance between two measures $\nu_1, \nu_2 \in \mathcal{P}_2(\mathbb{S}^2)$ can be bound by terms involving their cumulative distribution functions.

We start by calculating the Wasserstein distance $W_1(\nu_1, \nu_2)$ between probability measures ν_1 and ν_2 on $\mathbb{S}^2$, which are absolutely continuous with respect to area and have disintegrations

$$d\nu_j = f_j(\theta) g_j(\phi \mid \theta) \sin\phi \, d\phi d\theta \qquad (\theta \in [-\pi, \pi], \phi \in [0, \pi], j = 1, 2)$$

where f_j $(j = 1, 2)$ are probability density functions on $[-\pi, \pi]$ that give the marginal distributions of ν_j in the longitude θ variable, and g_j in the colatitude variable. Let F_j be the cumulative distribution function of $f_j(\theta)d\theta$ and G_j be the cumulative distribution function of $g_j(\phi)\sin\phi\, d\phi$. We measure $W_1(\nu_1, \nu_2)$ in terms of one-dimensional distributions. Given distributions on $\mathbb{R}$ with cumulative distribution functions F_1 and F_2, we write $W_1(F_1, F_2)$ for the Wasserstein distance between the distributions for cost function $|x - y|$. Let $\psi : [-\pi, \pi] \to [-\pi, \pi]$ be an increasing function that induces $f_2(\theta)d\theta$ from $f_1(\theta)d\theta$; then

$$W_1(\nu_1, \nu_2) \le W_1\big(F_1, F_2\big) + \int_{-\pi}^{\pi} W_1\big(G_2(\cdot \mid \psi(\theta)), G_1(\cdot \mid \theta)\big) f_1(\theta)d\theta.$$

In particular, for $f_1(\theta) = 1/(2\pi)$ and $g_1(\phi) = 1/2$, we have a product measure $\nu_1(d\theta d\phi) = (4\pi)^{-1}\sin\phi\ d\phi d\theta$ giving normalized surface area on the sphere. Then $F_1(\theta) = (\theta + \pi)/(2\pi)$ and $F_2(\psi(\theta)) = (\theta + \pi)/(2\pi)$, so $\psi(2\pi(\tau - 1/2))$ for $\tau \in [0, 1]$ gives the inverse function of F_2. We deduce that

$$W_1(F_1, F_2) = \int_{-\pi}^{\pi} \Big| \frac{\theta + \pi}{2\pi} - F_2(\theta) \Big| d\theta \tag{10.37}$$

and

$$W_1\big(G_2(\cdot \mid \psi(\theta)), G_1(\cdot \mid \theta)\big) = \int_0^{\pi} \Big| \int_0^{\phi} \big(g_2(\phi' \mid \psi(\theta)) - (1/2)\big) \sin\phi' d\phi' \Big| d\phi \tag{10.38}$$

Hence the Wasserstein distance can be bounded in terms of the cumulative distribution functions by

$$\begin{aligned} W_1(\nu_1, \nu_2) &\le W_1(F_2, F_1) + W_1(G_2, G_1) + \int_{-\pi}^{\pi} W_1(G_2(\cdot \mid \theta), G_2) dF_1(\theta) \\ &= \int_{-\pi}^{\pi} \Big| \frac{\theta + \pi}{2\pi} - F_2(\theta) \Big| d\theta + \int_0^{\pi} \Big| G_2(\phi) - \frac{1 - \cos\phi}{2} \Big| d\phi \\ &\quad + \int_{-\pi}^{\pi} \int_0^{\pi} \big| G_2(\phi \mid \theta) - G_2(\phi) \big| dF_1(\theta) d\phi. \end{aligned} \tag{10.39}$$

The triangle inequality has been used to obtain a more symmetrical expression involving the Wasserstein distances for the marginal distributions and the G conditional

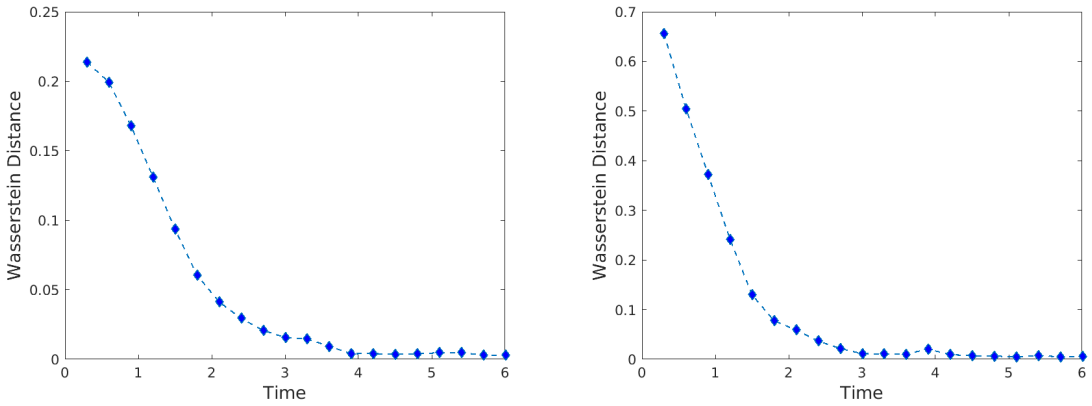

Figure 10.3: The plots involve the difference between the CDFs of two marginals. For θ_s, the predicted CDF $F_1(\theta) = (\theta + \pi)/(2\pi)$ is compared with the empirical CDF, $F_N^{\theta_s}$. The Wasserstein distance between $F_1(\theta)$ and $F_N^{\theta_s}$ is displayed on the **left**. For ϕ_s, the predicted CDF $G_1(\phi) = (1 - \cos(\phi))/2$ is compared with the empirical CDF $G_N^{\phi_s}$. The Wasserstein distance between $G_1(\phi)$ and $G_N^{\phi_s}$ is displayed on the **right**. The empirical measures considered are created using $N = 10^5$ samples and evaluated at each of the datapoints indicated on the graphs ($s = 0.3, 0.6, ..., 6.0$).

distributions, namely the dependence of the colatitude distribution on longitude.

Uniform distribution

As discussed in the previous section, the uniform measure on the sphere $\mathbb{S}^2$ is given by $\nu(A) = \int_A \sin\phi \; d\theta d\phi$, this gives marginal densities $\rho(\theta) = 1/2\pi$ and $\rho(\phi) = \sin(\phi)/2$. For each density there is a cumulative distribution function,

$$F_1(\theta) = \frac{\theta + \pi}{2\pi}, \quad \text{and} \quad G_1(\phi) = \frac{1 - \cos(\phi)}{2}, \tag{10.40}$$

as shown in Equation (10.39). The empirical cumulative distribution functions for $N \in \mathbb{N}$ samples are given by

$$F_N^{\theta_s}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{[0,\theta_s(\omega_i)]}(\theta), \quad \text{and} \quad G_N^{\phi_s}(\phi) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{[0,\phi_s(\omega_i)]}(\phi). \tag{10.41}$$

For $N = 10^6$ these empirical marginals are generated using the previously mentioned data at timesteps $s = 0.3, 0.6, 0.9, \ldots, 6.0$. Using these distributions the Wasserstein

distances $W_1(F_1, F_N^{\theta_s})$ and $W_1(G_1, G_N^{\phi_s})$ can be estimated.

For $s = 10$ the final two hypotheses were tested using a Kolmogorov-Smirnov test in MATLAB [39, kstest]. The hypothesis of H_0^θ was that at $s = 10$ the θ_s coordinate is distributed with CDF F_1, and with a significance of 0.005 the Kolmogorov-Smirnov test was unable to reject the null hypothesis. The final hypothesis test needed was H_0^θ, that at $s = 10$ the ϕ_s coordinate is distributed with CDF G_1, and with a significance of 0.005 the Kolmogorov-Smirnov test was also unable to reject this null hypothesis.

10.7.3 Concentration inequality

Estimates for the convergence of the empirical measure to its true measure in Wasserstein distance are given by Blower [8, Thm IV.3]. As a result of our hypothesis testing, it is significantly likely (99%) that the distribution of (θ_s, ϕ_s) at $s = 6$ is the uniform measure on the sphere. With the empirical data of $N = 10^6$ points used in the hypothesis testing, the expectation, $\mathbb{E}W_1(F_N^{\theta_s}, F_1)$ and $\mathbb{E}W_1(G_N^{\phi_s}, G_1)$ are close to zero by the estimates ,

$$\mathbb{E}W_1(F_N^{\theta_s}, F_1) \leq \frac{1}{\sqrt{N}} \int_{-\infty}^{\infty} \sqrt{F(t)(1 - F(t))}\, dt, \tag{10.42}$$

given in Proposition IV.4 of [8] and similarly for G. The Wasserstein distance can be calculated directly for the $N = 10^6$ samples, giving $W_1(F_N^{\theta_s}, F_1) \leq 0.010$ and $W_1(W_1(G_N^{\phi_s}, G_1) \leq 0.011$. The choice of $\varepsilon = 0.02$ and $N = 10^6$ and reasonable α in the estimate [8, Thm IV.3],

$$\mathbb{P}(\big(W_1((F_N^{\theta_s}, F_1)) > \varepsilon\big) < 2\exp(-N\alpha\varepsilon^2/2) < 0.01, \tag{10.43}$$

shows that with at least 0.99 probability it holds that $W_1((F_N^{\theta_s}, F_1)) \leq 0.02$. This inequality and the implications equally holds for G. The calculated Wasserstein distances are well within the predicted range with a significance level of $\alpha = 0.01$.

Chapter 11

Numerics of the NLSE by Fourier series

An alternate construction of the periodic nonlinear Schrödinger equation involves employing the Fourier coefficients of functions in $L^2(\mathbb{T})$. The Hamiltonian for the NLSE can be recast in terms of Fourier coefficients and then a new ODE can be constructed on the space truncated at the N^{th} Fourier coefficient. The invariants of the Hamiltonian system ensure the Fourier coefficients are constrained to a high dimensional sphere, and the ODE turns out to be of the appropriate form to consider evolution in the Lie algebra again. A numerical method is constructed to model the evolution of this system and its complexity is discussed. The large matrices hampered the ability to run significant numbers of simulations and perform statistical tests on the empirical distribution. However, this avenue of thought might be possible with access to substantial computational power.

11.1 The Schrödinger equation in Fourier space

Once again the focus is to solve the periodic NLSE with initial condition ϕ within the support of the Gibbs measure. Let $u \in H^1(\mathbb{T}, \mathbb{C})$ denote the solution. Truncate the Fourier series of $u = P + iQ \in H^1(\mathbb{T}, \mathbb{C})$ to the N^{th} term giving:

$$u_N(\theta) = \sum_{k=-N}^{N} (a_k + ib_k)e^{ik\theta}.$$

The finite dimensional Hamiltonian for the NLS can be expressed in terms of canonical variables $(a_k, b_k)_{k=-N}^N$ by

$$H_3^{(N)}((a_k),(b_k)) = \frac{1}{2}\sum_{k=-N}^{N} k^2(a_k^2 + b_k^2) + \frac{\beta}{4}\int \Big| \sum_{k=-N}^{N} (a_k + ib_k)e^{ik\theta}\Big|^4 \frac{d\theta}{2\pi}. \tag{11.1}$$

This truncated Hamiltonian incurs an approximation error between u, a solution to the Hamiltonian H in Definition 5.2 and u_N a solution to the Hamiltonian system $H_3^{(N)}$. If K represents the radius of the ball on which the Gibbs measure is normalised, and N the number of Fourier modes considered, then the head of the series is bounded by estimates such as [5, Eq. 2.19],

$$\left| \int_{\mathbb{T}} (u(\theta)^3 - (u(\theta) - u_N(\theta))^3) \frac{d\theta}{\sqrt{2\pi}} \right| \leq 7\alpha(2N+1)^{\frac{3}{2}} K^{\frac{3}{2}},$$

where α is a constant which can be derived from the Hamiltonian. This is related to the choice of initial condition, ϕ, such that $\|\phi\|_{L^2} \leq K$, and the invarance of the L^2 norm of the solution to the NLSE integrable system [9, Eq. 1.8].

The tail of the Fourier truncation can be bound in supremum norm using the fact that $u, u_N \in H^1$. This follows the same argument as Lemma 3.2.10 and can be found in the discussion in [6, Prop 3.1]. Let $\{|n| \geq N\}$ denote the set of $n \in \mathbb{Z} \setminus (-N, N)$, and invoke Cauchy-Schwartz,

$$\begin{aligned}\|u-u_N\|_\infty &\le \sum_{\{|n|\ge N\}} |a_n| = \left\langle \sum_{n=-\infty}^{\infty} e^{in\theta}, \sum_{\{|n|\ge N\}} |a_n| e^{in\theta} \right\rangle \\ &= \left\langle \sum_{n=-\infty}^{\infty} \frac{e^{in\theta}}{\sqrt{n^2+1}}, \sum_{\{|n|\ge N\}} |a_n|\sqrt{n^2+1}\, e^{in\theta} \right\rangle \\ &\le \left(\sum_{n=-\infty}^{\infty} \frac{1}{n^2+1} \right)^{\frac{1}{2}} \left(\sum_{\{|n|\ge N\}} (n^2+1)\, |a_n|^2 \right)^{\frac{1}{2}}.\end{aligned}$$

The series $\sum_{n=-\infty}^{\infty} \frac{1}{n^2+1}$ converges, and the second summation tends to zero as $N \to \infty$, therefore $\|u-u_N\|_\infty$ tends to zero as $N \to \infty$. In addition the second summation is equal to $\|u-u_N\|_{H^1}$.

The distance between H (Definition 5.2) and H_N can be bound as will be shown. The first term in the definition is known as the kinetic energy,

$$\sum_{n=-\infty}^{i} nftyk^2(a_k^2+b_k^2) - \sum_{|n|\ge N} k^2(a_k^2+b_k^2) = \|u'-u_N'\|_{L^2}^2. \tag{11.2}$$

which is bounded by the H^1 norm. The second term is called the potential energy, $U(u) = \int |u|^4 d\theta/2\pi = \|u\|_{L^4}^4$. By M. Riesz' theorem there exists a constant $C_4 > 0$ such that $\|u_N\|_{L^4}^4 \le C_4 \|u\|_{L^4}^4$, which implies that $U(u_N) \le C_4 U(u)$. Note that the potential energy functional is always positive and by virtue of its place within a Hamiltonian of an integrable system there exists K such that $\|u\|_{L^2}^2 \le K$ and by extension $\|u_N\|_{L^2}^2 \le K$. The functional U is convex, and Blower [6, Eq. 3.10] gives a bound on the distance between U and its chord between u and u_N for $0 \le t \le 1$,

$$0 \le tU(u) + (1-t)U(u_N) - U(tu+(1-t)u_N) \le 28Nt(1-t)\|u'-u_N'\|_{L^2}^2. \tag{11.3}$$

Combining this analysis, one can evaluate the full Hamiltonian,

$$\begin{aligned}\left|H(u) - H_3^N(u_N)\right| \leq & \frac{1}{2}\left|\sum_{k=-\infty}^{\infty} k^2(a_k^2+b_k^2) - \sum_{k=-N}^{N} k^2(a_k^2+b_k^2)\right| \\ & + \frac{\beta}{4}\left|\int |u|^4 \frac{d\theta}{2\pi} - \int |u_N|^4 \frac{d\theta}{2\pi}\right| \to 0,\end{aligned}$$

as $N \to \infty$.

11.1.1 The differential equations for Fourier modes

Returning to the derivation of the numerical method, the canonical equations of motion of this Hamiltonian are

$$\frac{\delta H_3^{(N)}}{\delta a_j} = \frac{db_j}{dt}, \quad \frac{\delta H_3^{(N)}}{\delta b_j} = -\frac{da_j}{dt}.$$

Recall that the functions $e^{ik\theta}$ are orthogonal polynomials in $L^2(\mathbb{T})$, and $|z|^2 = \bar{z}z$. Therefore the canonical equations of motion can be expressed as:

$$\begin{aligned}\frac{da_j}{dt} = -j^2 b_j & + \frac{i\beta}{2} \sum_{-j+k-l+m=0} (a_k + ib_k)\overline{(a_l + ib_l)}(a_m + ib_m) \\ & - \frac{i\beta}{2} \sum_{-k+j-l+m=0} \overline{(a_k + ib_k)(a_l + ib_l)}(a_m + ib_m), \end{aligned} \tag{11.4}$$

and

$$\begin{aligned}\frac{db_j}{dt} = j^2 a_j & + \frac{\beta}{2} \sum_{-j+k-l+m=0} (a_k + ib_k)\overline{(a_l + ib_l)}(a_m + ib_m) \\ & + \frac{i\beta}{2} \sum_{-k+j-l+m=0} \overline{(a_k + ib_k)(a_l + ib_l)}(a_m + ib_m). \end{aligned} \tag{11.5}$$

If these equations can be expressed as a vector differential equation of the form $dX/dt = MX$ where M belongs to $\mathfrak{so}(4N+2)$ then $SO(4N+2)$ based methods for solving ODEs can be employed. Considering just the linear terms, the canonical equations of motion

can be written in compact form,

$$\frac{d}{dt}\begin{bmatrix} a_j \\ b_j \end{bmatrix} = \begin{pmatrix} 0 & -j^2 \\ j^2 & 0 \end{pmatrix} \begin{bmatrix} a_j \\ b_j \end{bmatrix}. \tag{11.6}$$

The whole $2N$ of these coupled equations can be expressed as

$$\frac{d}{dt}\begin{bmatrix} a_N \\ \vdots \\ a_{-N} \\ b_N \\ \vdots \\ b_{-N} \end{bmatrix} = \begin{pmatrix} & & & -N^2 & 0 & 0 \\ & & & 0 & \ddots & 0 \\ & & & 0 & 0 & -(-N)^2 \\ N^2 & 0 & 0 & & & \\ 0 & \ddots & 0 & & & \\ 0 & 0 & (-N)^2 & & & \end{pmatrix} \begin{bmatrix} a_N \\ \vdots \\ a_{-N} \\ b_N \\ \vdots \\ b_{-N} \end{bmatrix}. \tag{11.7}$$

So the linear term is definitely skew symmetric, now consider the nonlinear terms in Equations (11.4) and (11.5). Define the $N \times N$ matrix A by

$$A_{jk} = \sum_{l,m,\, m-l=j-k} \overline{(a_l + ib_l)}(a_m + ib_m). \tag{11.8}$$

From this definition it is evident that $\overline{A_{jk}} = A_{kj}$, and so $\overline{A} = A^T$ making A Hermitian. In addition A is a Toeplitz matrix. Then the nonlinear term for Equation (11.7) can be written as

$$\left(\begin{array}{c|c} iA - i\overline{A} & -A - \overline{A} \\ \\ \hline A + \overline{A} & iA - i\overline{A} \\ \\ \end{array}\right) \begin{bmatrix} a_N \\ \vdots \\ a_{-N} \\ b_N \\ \vdots \\ b_{-N} \end{bmatrix} \tag{11.9}$$

and this matrix is skew symmetric. It is skew symmetric not skew-Hermitian — one can check that it equals its complex conjugate, and hence must be a real matrix. Let $X := (a_N, ..., a_{-N}, b_N, ..., b_{-N})^T$, and denote the linear matrix of Equation (11.7) by L and the nonlinear matrix in Equation (11.9) by $\mathcal{N}(X)$. Both matrices are skew

symmetric and so is their sum, therefore

$$\frac{dX}{dt} = (L + \mathcal{N}(X))X \tag{11.10}$$

evolves in $SO(4N + 2)$.

11.2 Algorithm employing Trotters product formula

As a result of X evolving in $SO(N)$ which is implied by Equation (11.11), $\sum_{i=-N}^{N}(a_i^2 + b_i^2) = k$ where k is constant. As this holds for the interval $t \in [0, \tau]$ for which the differential equation is being solved, one can restrict the initial conditions so that $k = 1$ and one is left with a problem on the unit sphere $\mathbb{S}^{n-1}$. This means for a numerical method to be viable it should conform to the geometry of the sphere.

In addition to this constraint, Equation (11.11) has an exploitable linear term. The structure of the matrix L as given in Equation (11.7) means its exponential can be calculated in terms of trigonometric functions exactly analogous to the derivation of Rodriguez' formula. For further motivation to explore approaches which separate the exponential, the process used by MATLAB to calculate exponentials of general matrices is by the scaling and squaring method. The matrix is scaled so that for some l, $\|L\|/l \leq 1$. Details are covered in the next section but it suffices to say that a matrix with a large norm (such as L where $\|L\|_\infty = 10^6$) is particularly ill suited to this method. Yet, the nonlinear term, which cannot be dealt with using closed form expressions, has norm $\|\mathcal{N}(X)\|_\infty = \max_{j,k\in[-N,N]}\{\overline{2(a_k + ib_k)}(a_j + ib_j)\} \leq 1$, making it ideal for the scaling and squaring method.

Splitting methods [32, p.47] are commonly used to split one step of a method into two 'parts' such as motion due to a linear and nonlinear term. When dealing with non-commuting matrix differential equations, splitting methods will introduce additional errors. The Baker-Campbell-Hausdorff formula [68] shows that $e^{A+B} \neq e^A e^B$ unless $[A, B] = 0$. A work around is to use Trotters product formula, and this is the method used.

Considering the linear and non-linear parts of the differential equation separately,

$$\frac{dX}{dt} = (L + \mathcal{N}(X))X. \tag{11.11}$$

Employing a first order Euler method within the Lie algebra [32, Eq.8.10], results in an update $X_{n+1} = \exp(h(L + \mathcal{N}(X_n))X_n$. Then the Trotters formula [28, Thm 8.12] is used to split this exponential into a linear and nonlinear part:

$$\exp(hL + h\mathcal{N}) \approx \left(\exp(L\frac{h}{k})\exp(\mathcal{N}\frac{h}{k})\right)^k. \tag{11.12}$$

The exponential of the linear part can be done algebraically with trigonometric expressions. The algorithm follows.

Algorithm 11.1 Trotters product formula algorithm

$N \leftarrow$ number of Fourier modes,
$h \leftarrow$ timestep length,
`timesteps` $\leftarrow$ number of timesteps,
$k \leftarrow$ approximation constant for trotters algorithm
sample γ_i and $\tilde{\gamma}_i$ from a $N(0,1)$ distribution.
$a_i \leftarrow \gamma_i / i$
$b_i \leftarrow \tilde{\gamma}_i / i$
$X \leftarrow [a_N, \ldots a_{-N}, b_N \ldots b_{-N}]$ note both $a_0 = 0$ and $b_0 = 0$.
$X \leftarrow X/\|X\|$,
$A \leftarrow diag([-N : N].^2)$

$$L \leftarrow \begin{pmatrix} \cos(hA/k) & -\sin(hA/k) \\ \sin(hA/k) & \cos(hA/k) \end{pmatrix}$$

for $i = 1 :$ `timesteps` **do**
    $X \leftarrow (L * \exp(\mathcal{N}(X)h/k))^k X$ ▷ $\mathcal{N}(X)$ is given in Algorithm 11.2
end for

The error incurred in this algorithm will depend on k and h. In terms of h the symplectic Euler algorithm is a first order numerical scheme so the error will be linear with respect to h. The trotters formula incurs an error of the order $\frac{h}{\sqrt{k}}$. So in total the method has linear order with respect to step size h and is also proportional to $1/\sqrt{k}$.

11.2.1 Computational complexity

The complexity of the algorithm is linearly dependent on the number of steps required to simulate the solution over the interval required; for a solution on $[0, 1]$ it requires

Algorithm 11.2 Function $\mathcal{N}(X)$

function $\mathcal{N}$(X)

$Y \leftarrow (a_j + ib_j)_{j=-N}^{N}$ where each a_j, b_j are taken from X.

Let $T_i(Y)$ denote the right shift operator i positions, for example

$$T_1([a, b, c]) = [0, a, b]\,.$$

for $i = 1 : (2N + 1)$ **do**

$B_i \leftarrow YT_i(\overline{Y})^\top$

end for

$A \in M_{2N+1}(\mathbb{C})$ is given by $A_{jk} \leftarrow B_{k-j}$. Note $B_{-i} = \overline{B}_i$. ▷ See Eq. (11.8)

$$A = \begin{bmatrix} B_0 & B_1 & \dots \\ B_{-1} & B_0 & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

Define $\mathcal{N}$ from combining A with itself, ▷ See Eq. (11.9)

$$\mathcal{N} \leftarrow \left(\begin{array}{c|c} iA - i\overline{A} & -A - \overline{A} \\ \hline A + \overline{A} & iA - i\overline{A} \end{array}\right)$$

return $\mathcal{N}$

end function

$1/h$ timesteps. It is also depends additively on the number of sample paths required to build up an empirical measure.

Casting focus to one step of the algorithm, the number of Fourier modes considered is the most expensive parameter of the algorithm as it determines the size of the matrices. A matrix multiplication of two square N dimensional matrices requires $N^2(N + (N - 1)) \approx N^3$ operations, which is N multiplications and $N - 1$ additions for each entry of the N by N matrix. That said, there is a routine that MATLAB uses called BLAS [19] which minimizes the time required to calculate vector/matrix multiplications, so that for large matrices it performs faster than the naive N^3 approach. In addition the calculation of a matrix to the power of k can be done in far fewer operations than k multiplications, for instance if $k = 2^j$ then the matrix multiplication can be done with j squarings.

Nevertheless, the number of matrix multiplications is the largest factor in the complexity of the algorithm. The number needed per step depends on the method used to calculate the exponential of a matrix. The exponential map used in the Algorithm 0 is the MATLAB function `expm`; this function uses the scaling and squaring method. For an arbitrary matrix M it aims to use $\exp(M/l)^l = \exp(M)$ so that $\|M/l\| \leq 1$ and then use Padé approximants to find $\exp(M/l)$ [36]. The intricacies of the Padé approximation is discussed in Section 11.2.2. In terms of computational complexity, it is attractive to choose $l = 2^s$ as then the l matrix multiplications can instead be carried out by s squarings. Then the approximation of $\exp(M/l)$ is done by the Padé approximation $R_{mm}(M/l)$ which is a rational function with both numerator and denominator having degree m, implying $a(a + 1)$ multiplications. That said, estimates of just m multiplications are found in the literature [29, p.573]. The choice for s and m depend on the norm of M. Higham [36] provides values of θ_m so that if $\|2^{-s}M\| \leq \theta_m$ then the error incurred by `expm` is smaller than 2^{-53}. The dominant term in the computational complexity is thus

$$N^3 \left(s + m(m + 1)\right).$$

As an example, if one was to set a 8 term truncation, $\theta_8 = 1.5$, which implies acceptable error rates if $\|M\| \leq 2^{s+1}$, thus the best complexity in this case would be

$$N^3 \left(\frac{\log \|\mathcal{N}(X)\|}{\log 2} + \log 2 + 72 \right).$$

11.2.2 Constrained to the sphere

In Section 11.2 we mentioned that the method was first order dependent on a suitable choice of k. In this section we discuss how the algorithm preserves the invariance of the L^2 norm of the solution u, which is equivalent to the euclidean norm of X. And this invariance is preserved despite using an approximation for the exponential function.

The one-step ahead for the first order method proposed is

$$X_{n+1} = \left(\exp(L\frac{h}{k})\exp(\mathcal{N}(X_n)\frac{h}{k})\right)^k X_n.$$

Recall the initial data u is sampled from the space of Gibbs measurable functions with L^2 norm bounded by K, that is $u \in H^1 \cap B_K$. Now truncate its Fourier series up to the N^{th} mode

$$u_N(\theta) = \sum_{k=-N}^{N} (a_k + ib_k)e^{ik\theta}$$

and denote $X_0 = (a_N, ..., a_{-N}, b_N, ..., b_{-N})^T \in \mathbb{R}^{4N+2}$. Now $\|X_0\|^2 \leq K$ and the L^2 norm of each sample path/solution is invariant because of the skew symmetric form of the differential equation, hence $(X_n)_{n\in\mathbb{N}}$ all lie on the same $4N+2$ dimensional sphere in $\mathbb{R}^{4N+2}$. Thus each curve can be thought of as the action of the group $SO(4N+2)$ on the point X_0 and for convenience we will simply consider X_0 to be the first column of a matrix in $SO(4N+2)$.

A detailed discussion of the spaces involved mapping the space $SO(4N+2)$ to itself by the numerical algorithm needs to take place.

The map $\mathcal{N}(X) : SO(4N+2) \to G_1 \subset \mathfrak{so}(4N+2)$ has a bound,

$$\|\mathcal{N}\|_\infty = \frac{|\beta|}{4}\int |u|^4 \leq \frac{|\beta|}{4}N\int |u|^2 \leq \frac{|\beta|}{4}NK.$$

So $G_1 \coloneqq \mathfrak{so}(4N+2) \cap \{\|M\|_\infty \leq \frac{|\beta|}{4}NK\}$.

The exponential map used in the algorithm is the MATLAB function `expm`, this function uses the scaling and squaring method. For an arbitrary matrix M it aims to use $\exp(M/k)^k = \exp(M)$ so that $\|M/k\| \leq 1$ and then to use Padé approximants to find $\exp(M/k)$ [36]. For computational efficiency, it is attractive to choose $k = 2^s$ as

then the k matrix multiplications can instead be carried out by s squarings.

The Padé approximants are rational functions with integer coefficients and are given by $R_{ab}(x) = P_{ab}(x)/Q_{ab}(x)$ where $P_{ab}(x) = Q_{ba}(-x)$. The diagonal Padé approximants are the most efficient to calculate as the algorithm requires $s + \max(a, b)$ matrix multiplications [29], in other words setting $a = b$ gives the highest order approximation for the complexity. Considering this, the diagonal Padé approximants are given by $R_{aa}(x) = P_{aa}(x)/P_{aa}(-x)$ where

$$P_{aa}(x) = \sum_{j=0}^{a} \frac{(2a-j)!a!}{(2a)!j!(a-j)!} x^j. \tag{11.13}$$

Proposition 11.2.1. *The Padé approximant of a skew symmetric matrix is unitary. That is, if $M \in \mathfrak{so}(N)$, then $R_{aa}(M) \in SO(N)$ for any $a \in \mathbb{N}$.*

Proof. Ehle and Van Rossun have shown [62, Thm 1.1] that the diagonal Padé approximants have all of their zeros in the set $\{z \in \mathbb{C} \mid \Re(z) > 0\}$ and therefore their poles in the set $\{z \in \mathbb{C} \mid \Re(z) < 0\}$. A skew symmetric matrix has only purely imaginary eigenvalues and thus $R_{aa}(M)$ is a function defined on the spectrum of M.

Thus we can discuss the eigenvalues of $R_{aa}(M)$, which, if $i\alpha_j$ represent the eigenvalues of M, are given by $R_{aa}(i\alpha_j)$[47, Thm 5.3.4]. In addition to this, the eigenvalues $R_{aa}(i\alpha_j)$ lie on the unit circle, as

$$|R_{aa}(i\alpha_j)|^2 = R_{aa}(i\alpha_j)R_{aa}(-i\alpha_j) = 1, \tag{11.14}$$

due to the form $R_{aa}(x)$ takes. Then the spectral theorem says that a normal matrix is unitary iff all of its eigenvalues are on the unit circle. So, provided $R_{aa}(M)$ is normal, it is unitary, and a real unitary matrix is orthogonal. $R_{aa}(M)$ is normal because M is skew, and therefore normal and

$$R_{aa}(M)R_{aa}(M)^T = R_{aa}(M)R_{aa}(-M) = \frac{P_{aa}(M)}{P_{aa}(-M)} \frac{P_{aa}(-M)}{P_{aa}(M)}, \tag{11.15}$$

and polynomials in M commute because M commutes with itself. □

11.2.3 Implementation

I was unable to implement the algorithm in a form that permitted the number of samples required for statistical tests to be carried out on the empirical distribution. This is a result of how the complexity of the algorithm scales with the dimension of the problem. The challenge may be surmountable with more computing power, and if so I think there may be scope to compare the empirical distribution produced by the algorithm with a distribution created using a Markov Chain Monte Carlo (MCMC) method for the Gibbs [25]. Statistical methods such as the Kolmogorov-Smirnov test could then be applied in this setting to deduce if the two collections of samples are drawn from the same distribution.

Chapter 12

Numerics of the Euler equations via Transport

As discussed in Section 3.4, measures can be weak solutions to PDEs. In this chapter the intent is to consider the Euler equations in a manner in which measure valued solutions can be found. Hamiltonian systems provide a template for this. A smooth solution to a PDE that forms the canonical equations of motion for a Hamiltonian is an extremal of said Hamiltonian — constructed by calculus of variations. Interpreting the Hamiltonian of the system instead by an action functional applied to a measure, an analogous method is developed for measure valued or weak solutions to the system.

The Hamiltonian is a measure of the change in energy when moving between points in phase space. In the language of optimal transport, it is similar to a *cost function*. The evolution of the system between states — probability densities — in a way that minimises the change in energy — minimises the cost function — is an *optimal transport map*. In this case however, the dynamics of a solution to the isentropic Euler equations is also governed by an internal energy $\mathcal{U}$, which restricts the mass from moving optimally.

This chapter is adapted from the work of Gangbo et al [27]. In an effort to understand and explain the theory behind his numerical methods we have expanded on his paper with a discussion of the convexity of the operator under discussion and the existence and uniqueness of a minimiser of this operator on the appropriate space. In the next chapter his method is applied to the dam break problem, giving a comparison within a dynamic system with an exact solution.

12.1 Setting up the problem

The density function ρ represents the distribution of the mass of the fluid. Normalising so that the entire mass is 1, the density can instead be treated as a probability measure in $\mathcal{P}_2(\mathbb{R}^n)$. To give some evidence as to why, note that the function ρ represents the location of the mass, as no fluid will concentrate mass into areas of zero volume ρ should be absolutely continuous with respect to Lebesgue measure. Thus the probability measure ρ will have a Radon Nikodym derivative - its density - and this density will be $\rho(x)$. Using identical notation for the probability measure and its density could cause confusion, but in this case reference to the measure or the density should be clear from context.

Consider the Hamiltonian for the isentropic Euler equations,

$$H(\rho, u) = \frac{1}{2}\int \|u\|^2 \rho(x)dx + \int U(\rho)dx, \tag{12.1}$$

where one can denote the integral $\int U(\rho)dx = \mathcal{U}(\rho)$. The solution to the Euler equations will be the pair (ρ, u) which minimise the change in this Hamiltonian over a short timeframe τ. Instead of taking the variation of the Hamiltonian to be left with a set of PDEs we leave the Hamiltonian as an energy functional and implement ideas from the theory of gradient flows [2]. Following the work of Gangbo [27] the problem is equivalent to searching for the measure, ρ which, for $\rho_1 \in \mathcal{P}_2(\mathbb{R}^n)$ minimises

$$\frac{1}{2\tau} W_2(\rho_1, \rho)^2 + \mathcal{U}(\rho). \tag{12.2}$$

In addition, to control the internal energy, $\mathcal{U}(\rho) < \infty$ one must impose that $\rho \in \mathcal{P}_2(\mathbb{R}^n) \cap L^\gamma$. Finally, note that this construction offers no method to calculate the update in velocity of the fluid, and so the problem on the manifold $\mathcal{P}_2(\mathbb{R}^n)$ could be extended to its tangent space.

12.2 The tangent space

Understanding how the velocity will update between steps requires a measure on the tangent bundle, $\mathbb{TR}^n = \mathbb{R}^n \times T\mathbb{R}^n$.

$$\mathbb{TR}^n := \{(x, u) \mid x \in \mathbb{R}^n,\ u \in T_x\mathbb{R}^n\}. \tag{12.3}$$

Curves on the manifold can be lifted to a graph on the tangent bundle with use of the function $g : x \mapsto (x, u(x)); \mathbb{R}^n \to \mathbb{TR}^n$. This function can thus also act as a pushforward for measures from $\mathcal{P}(\mathbb{R}^n)$ to the tangent bundle. The set of measures which lie on the tangent bundle and have a marginal with respect to the base manifold (in this case $\mathbb{R}^n$) which is in $\mathcal{P}(\mathbb{R}^n)$ will be denoted by $\mathcal{P}(\mathbb{TR}^n)$.

Definition 12.2.1. The set $\mathcal{P}(\mathbb{TR}^n)$ is defined by the statement, $\mu \in \mathcal{P}(\mathbb{TR}^n)$ if and only if there exists $\rho \in \mathcal{P}(\mathbb{R}^n)$ and $u \in L^1(\mathbb{R}^n; \mathbb{R}^n\rho)$, such that the measure $g\#\rho = \mu$. Equivalently, for all test functions $f \in L^1(\mathbb{TR}^n, \mu)$

$$\int_{\mathbb{TR}^n} f(y)\mu(dy) = \int_{\mathbb{R}^n} f(g(x))\rho(x)dx. \tag{12.4}$$

The measure $\mu \in \mathcal{P}(\mathbb{TR}^n)$ has disintegration $\mu = \sigma\rho$, in other words, for any continuous and bounded test function f,

$$\int_{\mathbb{TR}^n} f(x, u)\mu(dx, du) = \int_{\mathbb{R}^n} \left(\int_{T\mathbb{R}^n} f(x, u)\sigma_x(du) \right) \rho(x)dx. \tag{12.5}$$

The simplest example would be when $\sigma_x(du)$ allows $\cup_x T_x\mathbb{R}^n$ to be treated as one vector field on $\mathbb{R}^n$, this is when $\sigma_x(du) = \delta_x(u)$, where δ denotes the Dirac measure.

12.2.1 Acceleration cost

For the Riemannian manifold on which $\mathcal{P}(\mathbb{R}^n)$ live there is a clear metric, the Wasserstein distance. For the measures on the tangent bundle we would also like a metric. One measure of distance between points in phase space would be the minimum acceleration required to move from point to point, this is known as the *average acceleration cost* and is the initial cost function proposed by Gangbo and Westdickenberg. The average acceleration cost can be calculated from a calculus of variation argument.

Proposition 12.2.2. *Let γ, parametrised by time, represent any curve with $\gamma(0) = x_1$, $\gamma(\tau) = x_2$, $\dot{\gamma}(0) = u_1$ and $\dot{\gamma}(\tau) = u_2$. The curve with the minimum average acceleration has an acceleration of*

$$\frac{1}{\tau}\int_0^\tau |\ddot{\gamma}(s)|^2 ds = \frac{12}{\tau^2}\left|\frac{x_2 - x_1}{\tau} - \frac{u_2 + u_1}{2}\right|^2 + |\frac{u_2 - u_1}{\tau}|^2. \tag{12.6}$$

Proof. The proposition is proven by applying calculus of variations to the action

$$S[\gamma] = \frac{1}{\tau}\int_0^\tau |\ddot{\gamma}(s)|^2 ds,$$

in order to find its minima. Consider a variation of the acceleration along the curve γ. That is a perturbation of the curve γ, by a second curve, denoted f, which has zero initial and final positions and velocities, such that $\gamma + f$ satisfies the same initial and final conditions as γ. I will use superscript numbers in parenthesis to denote derivatives higher than one, so $\gamma^{(2)}$ denotes the acceleration and $\gamma^{(3)}$ the impulse of the curve.

Then the limit gives the variation of the action,

$$\begin{aligned}
\delta S &= \lim_{\varepsilon\to 0}\frac{S[\gamma + \varepsilon f] - S[\gamma]}{\varepsilon}, \\
&= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon\tau}\int_0^\tau |\gamma^{(2)}(s) + \epsilon f^{(2)}(s)|^2 + |\gamma^{(2)}(s)|^2 ds, \\
&= \lim_{\varepsilon\to 0}\frac{1}{\tau}\int_0^\tau 2\gamma^{(2)}(s) f^{(2)}(s) + \varepsilon |f^{(2)}(s)|^2 ds, \\
&= \frac{1}{\tau}\int_0^\tau 2\gamma^{(2)}(s) f^{(2)}(s) ds.
\end{aligned} \tag{12.7}$$

Then by definition, $f(0) = 0$, $f(\tau) = 0$, $\dot{f}(0) = 0$ and $\dot{f}(\tau) = 0$. Thus, applying integration by parts to Equation (12.7) so that the $h^{(2)}(s)$ term is differentiated twice the functional is simplified:

$$\frac{1}{\tau}\int_0^\tau 2\gamma^{(2)}(s) f^{(2)}(s) ds = \frac{2}{\tau}\int_0^\tau \gamma^{(3)}(s)\dot{f}(s) ds - \frac{2}{\tau}\left[\gamma^{(2)}(s)\dot{f}(s)\right]_0^\tau, \tag{12.8}$$

$$= \frac{2}{\tau}\int_0^\tau \gamma^{(4)}(s) f(s) ds - \frac{2}{\tau}\left[\gamma^{(3)}(s) f(s)\right]_0^\tau. \tag{12.9}$$

And thus the Euler-Lagrange equation to find the extremal of this functional is

$$\gamma^{(4)}(s) = \frac{d^4\gamma(s)}{ds^4} = 0, \tag{12.10}$$

implying that the minimising curve is a cubic polynomial. Next let $\gamma(s) = As^3 + Bs^2 + Cs + D$, and calculate the coefficients which minimise $S[\gamma]$ subject to the initial and final conditions. □

Definition 12.2.3. The expression for average acceleration leads to a definition of a cost function A_τ between points in the tangent bundle at time $t = 0$ and at time $t = \tau$,

$$\begin{aligned} A_\tau \ &: \mathbb{TR}^n_0 \times \mathbb{TR}^n_\tau \to [0, \infty) \\ &: ((x_1, u_1), (x_2, u_2)) \mapsto 3 \left| \frac{x_2 - x_1}{\tau} - \frac{u_2 + u_1}{2} \right|^2 + \frac{1}{4} |u_2 - u_1|^2 , \end{aligned}$$

which is equal to the minimum acceleration cost up to a scale factor of $1/4\tau^2$.

This cost function A_τ can be extended to act on the space of measures $\mathcal{P}(\mathbb{TR}^n)$, treating A_τ as the cost function for an optimal transport condition between two measures:

$$\mathbf{A}_\tau(\mu_1, \mu_2) := \inf\{ \int \int_{\mathbb{TR}^n \times \mathbb{TR}^n} A_\tau(\mathbf{x}, \mathbf{y}) \pi(d\mathbf{x}, d\mathbf{y}) \ : \ \pi \in \Pi(\mu_1, \mu_2)\}. \tag{12.11}$$

where $\Pi(\mu_1, \mu_2)$ is the set of all product measures with marginals μ_1 and μ_2. Then the optimal transport map, $\hat{\pi}$, minimises the acceleration cost for the entire tangent bundle.

12.2.2 The functional W_τ

The issue with Equation (12.11) is that A_τ is not a well defined metric. This distance function, A_τ, fits with our intuition about free motion, if a particle is at position x_1 with velocity u_1 and ends up at position $x_2 = x_1 + \tau u_1$ with the same velocity $u_1 = u_2$, then the acceleration cost should be zero - the particle hasn't accelerated. This is true, $A((x_1, u_1), (x_1 + \tau u_1, u_1)) = 0$. However, as a metric, this makes A_τ non-positive definite. In addition to this, it is also not symmetric. A metric must be positive definite

and symmetric and so A_τ is not suitable to consider optimal transport problems with respect to.

The distance function A_τ can be manipulated so that it measures the distance between the unaffected trajectory of a point from $t = 0$ to $t = \tau$ and its perturbed location under a minimal acceleration. This is achieved by exchanging the space $\mathbb{TR}_0^n$ with the tangent bundle at time $t = \tau$ if there was no acceleration after time $t = 0$, denoted $\tilde{\mathbb{TR}}_\tau^n$. This space is mapped to from $\mathbb{TR}_0^n$ via the map

$$X_\tau : \mathbb{TR}_0^n \to \tilde{\mathbb{TR}}_\tau^n \tag{12.12}$$

$$(x, u) \mapsto (x + \tau u, u). \tag{12.13}$$

Using this map we can define the new distance function, the function W_τ is defined in terms of A_τ by $W_\tau(X_\tau(\mathbf{x}), \mathbf{y}) = A_\tau(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x} \in \mathbb{TR}_0^n, \mathbf{y} \in \mathbb{TR}_\tau^n$. Both arguments of W_τ are on tangent bundles at time $t = \tau$ and it is given explicitly by

$$W_\tau((x_1, u_1), (x_2, u_2)) \mapsto \frac{3}{\tau^2} \left| \frac{x_2 - x_1}{\tau} - \frac{u_2 - u_1}{2} \right|^2 + \frac{1}{4} \left| \frac{u_2 - u_1}{\tau} \right|^2. \tag{12.14}$$

Mirroring $\mathbf{A}_\tau$, this metric can be extended to a distance between measures on $\mathcal{P}(\mathbb{TR}^n)$ by an optimal transport condition,

$$\mathbf{W}_\tau(\mu_1, \mu_2) := \inf\{ \int \int_{\mathbb{TR}^n \times \mathbb{TR}^n} W_\tau(\mathbf{x}, \mathbf{y}) \pi(d\mathbf{x}, d\mathbf{y}) \ : \ \pi \in \Pi(\mu_1, \mu_2)\}. \tag{12.15}$$

where $\Pi(\mu_1, \mu_2)$ is the set of all product measures with marginals μ_1 and μ_2.

Remark 12.2.4. This framing of the problem relies on the assumption that the transport map X_τ pushes forward an original $\rho_0 \in \mathcal{P}_2(\mathbb{R}^n)$ to ρ_1 which also lies in $\mathcal{P}_2(\mathbb{R}^n)$. For this to be the case the velocity field has to satisfy a non-degeneracy condition such as, for $A \in \mathcal{B}(\mathbb{R}^n)$ if $\rho_0(A) = 0$ then $\rho_0(X_\tau^{-1}(A)) = 0$ too [27, Eq. 2.5]. This is Brenier's polar factorisation theorem [74, p.112].

To see the importance of this condition, consider an initial velocity field $u(x) = -\frac{x}{\tau}$, the map X_τ would collapse the whole measure onto the point $x = 0$ and a pushforward via this map would therefore give the Dirac delta δ_0.

12.2.3 On the independence of the velocity distribution marginal

A result due to Gangbo and Westdickenberg [27, Prop. 4.5] establishes that, given $\mu_1 \in \mathcal{P}(\mathbb{TR}^n)$ with marginal $\mu_1|_1 = \rho_1$ and a second marginal $\rho_2 \in \mathcal{P}(\mathbb{R}^n)$ then the minimiser

$$\inf\{\mathbf{W}_\tau(\mu_1, \mu)^2 \ : \ \mu \in \mathcal{P}(\mathbb{TR}^n),\ \mu|_1 = \rho_2\} = \frac{3}{4\tau^2} W_2(\rho_1, \rho_2)^2, \tag{12.16}$$

where $W_2(\rho_1, \rho_2)$ is the Wasserstein distance. This condition is essentially saying that the minimisation process is independent of the initial velocity distribution.

Lemma 12.2.5. *[27, Prop. 4.5] If γ is the transport plan which minimises $W_\tau(\mu_1, \mu)$, let $\gamma' \in \mathcal{P}(\mathbb{TR} \times \mathbb{R})$ be the marginal of the (x_1, u_1, x_2) coordinates. Then γ has a disintegration $\gamma = \gamma' \sigma_2$ and Equation* (12.16) *implies that the minimisation has no u_2 dependence. Thus σ_2 is a Dirac measure and the minimising velocity is $u_2 = \beta(x_1, u_1, x_2)$, where*

$$\beta(x_1, u_1, x_2) = u_1 + \frac{3}{2}\frac{x_2 - x_1}{\tau}. \tag{12.17}$$

Proof. For any test function ϕ

$$\int_{\mathbb{TR}^n \times \mathbb{TR}^n} \phi(x_1, u_1, x_2, u_2) \gamma(dx_1, du_1, dx_2, du_2) = \int_{\mathbb{TR}^n \times \mathbb{R}^n} \left(\int_{T_{x_2}\mathbb{R}} \phi(x_1, u_1, x_2, u_2) \sigma_2(du_2 | x_1, u_1, x_2) \right) \gamma'(dx_1, du_1, dx_2)$$

The fact minimising $W_\tau(\mu_1, \mu)$is independent of u_2 implies that,

$$\int_{T_{x_2}\mathbb{R}} W_\tau(x_1, u_1, x_2, u_2) \sigma_2(du_2 | x_1, u_1, x_2) = \inf\{ \int_{T_{x_2}\mathbb{R}} W_\tau(x_1, u_1, x_2, u_2) \sigma(du_2) \mid \sigma \in \mathcal{P}(\mathbb{R})\}. \tag{12.18}$$

Then the map $u_2 \mapsto W_\tau(x_1, u_1, x_2, u_2)$ is convex and thus Jensen's inequality implies

that

$$\int_{T_{x_2}\mathbb{R}} W_\tau(x_1, u_1, x_2, u_2)\sigma(du_2) \geq \tag{12.19}$$

$$W_\tau\left(x_1, u_1, x_2, \int_{T_{x_2}\mathbb{R}} y\, \sigma(dy)\right). \tag{12.20}$$

If the measure σ is a Dirac distribution, then the inequality becomes an equality, and in some sense the Dirac distribution has a variance which is the smallest limit of variances of any distribution which is absolutely continuous with respect to Lebesgue. (though Dirac delta itself is not actually abs cont with respect to Lebesgue right)

Therefore a choice of $\sigma(dy) = \delta_a(dy)$ will minimise the infimum in Equation (12.18), and convert the problem of finding σ_2 into finding the location, a, of the delta function. In other words, simply minimising $W_\tau(x_1, u_1, x_2, a)$ □

12.3 Minimising the functional, existence and uniqueness.

The intuition behind this method is that of a Hamiltonian action functional, the minimiser of the action functional is the solution under the principle of least action. We have constructed a functional resembling the action of the Euler dynamical system based on minimising the acceleration cost plus the potential energy within the system.

The minimisation problem is to minimise the functional $\mathbf{A}_\tau(\mu, \mu^*)^2 + \mathcal{U}(\mu^*) \ : \ \mu^* \in \mathcal{P}(\mathbb{TR}^n)$, and the solution is given in the Proposition by Gangbo:

Proposition 12.3.1. *[27, Prop. 4.5] Consider a measure $\mu \in \mathcal{P}(\mathbb{TR}^n)$, defined by a density and velocity field (ρ, u), ρ is the density of an absolutely continuous measure with respect to Lebesgue, and u satisfies the non-degeneracy condition of Remark 12.2.4. As part of the definition $\mu|_1 = \rho(x)dx$ and $\mu|_2 = u\#(\rho(x)dx)$. If μ^τ is given by*

$$\mu^\tau = \arg\min_{\mu^*}\{\mathbf{A}_\tau(\mu, \mu^*)^2 + \mathcal{U}(\mu^*) \ : \ \mu^* \in \mathcal{P}(\mathbb{TR}^n)\}. \tag{12.21}$$

Then the pushforward map for the density, $\rho^\tau \in \mathcal{P}_2(\mathbb{R}^n)$, is given implicitly by

$$\rho^\tau dx = \left((Id + \frac{2\tau^2}{3}\nabla U'(\rho^\tau))^{-1} \circ (Id + \tau u) \right) \# \rho dx, \tag{12.22}$$

and the velocity is given by $u^\tau \in L^2(\mathbb{R}^n, \rho^\tau)$

$$u^\tau = u \circ (Id + \tau u)^{-1} \circ \left(Id + \frac{2\tau^2}{3}\nabla U'(\rho^\tau) \right) - \tau \nabla U'(\rho^\tau) \tag{12.23}$$

Before proving this proposition, the convexity of the functional needs to be evaluated, and the crucial subproblem is discussed in the following section. For consistent notation, the initial measure μ will be referred to as μ_1 with marginals $(\rho_1, u_1 \# \rho_1)$ and then the *free transport* of this measure is μ_2 with marginals $(\rho_2, u_2 \# \rho_2)$ where $\rho_2 = (Id + \tau u_1) \# \rho_1$ and $u_2 \# \rho_2 = u_1 \# \rho_1$.

12.3.1 Existence and uniqueness

In this section we aim to establish the existence and uniqueness of a minimiser of the functional

$$G(\rho) := \frac{1}{2\tau} W_2(\rho, \rho_0)^2 + \mathcal{U}(\rho), \tag{12.24}$$

on the set $\mathcal{P}_2(\mathbb{R}^n) \cap L^\gamma(\mathbb{R}^n)$. Proving these statements generally relies on the functional being convex over some sort of compact set. In this situation there are a few choices of structure upon which the set $\mathcal{P}_2(\mathbb{R}^n)$ can be convex. This idea will be delved into in the next section. For this section, classical convexity of the functional is all that is needed.

For any two measures $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R}^n)$ the *linear interpolant* of the measures is defined $\rho_t := t\rho_1 + (1-t)\rho_2$.

Definition 12.3.2 (Linear Convexity)**.** (i) A set of measures $P \subset \mathcal{P}_2(\mathbb{R})$ is *convex* if the linear interpolant between any two measures in the set, denoted ρ_t, produces another measure in the set (for all $t \in [0, 1]$).

(ii) A functional, G, defined on a convex subset of $\mathcal{P}_2(\mathbb{R}^n)$ is convex if the map $t \mapsto$

$G(\rho_t(x))$ is convex for all $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R}^n)$. Explicitly,

$$G(t\rho_1 + (1-t)\rho_2) \leq tG(\rho_1) + (1-t)G(\rho_2) \tag{12.25}$$

As a direct consequence the set $\mathcal{P}_2(\mathbb{R}^n) \cap L^\gamma(\mathbb{R}^n)$ is clearly convex.

Lemma 12.3.3. *The functional $G(\rho)$ given by Equation (12.24) is convex on $\mathcal{P}_2(\mathbb{R}^n)$ according to the definition of linear convexity.*

Proof. Consider the internal energy term, $\mathcal{U}(\rho_t)$ and express it using the general form of the internal energy $U(\rho) = \rho^\gamma$. Then the internal energy is related to the L^γ norm as,

$$\|\rho_t\|_\gamma = \left(\int_{\mathbb{R}^n} \rho_t^\gamma dx\right)^{\frac{1}{\gamma}} = (\mathcal{U}(\rho_t))^{\frac{1}{\gamma}}. \tag{12.26}$$

The coefficient $\gamma \in (1, 2]$ and so $f(x) = x^\gamma$ is convex, and the L^γ norm is convex by Minkowski's inequality. The composition of these two functions gives $\mathcal{U}$ and the composition of convex functions is convex.

Regarding the Wasserstein distance term, let the optimal map between $\tilde{\rho}$ and ρ_1 be denoted π_1 and likewise, the optimal map between $\tilde{\rho}$ and ρ_2 is denoted π_2. Then the density ρ_t can be expressed as a pushforward of $\tilde{\rho}$ by $\rho_t = (t\pi_1 + (1-t)\pi_2)\#\tilde{\rho}$. A simple application of the triangle inequality gives the convexity of the Wasserstein distance,

$$W_2(\rho_t, \tilde{\rho}) \leq \left(\int_{\mathbb{R}^n} \|x - t\pi_1(x) - (1-t)\pi_2(x)\|^2 \tilde{\rho} dx\right)^{\frac{1}{2}}, \tag{12.27}$$

$$\leq \left(\int_{\mathbb{R}^n} \|tx - t\pi_1(x)\|^2 \tilde{\rho} dx\right)^{\frac{1}{2}} + \left(\int_{\mathbb{R}^n} \|(1-t)x - (1-t)\pi_2(x)\|^2 \tilde{\rho} dx\right)^{\frac{1}{2}} \tag{12.28}$$

$$= tW_2(\rho_1, \tilde{\rho}) + (1-t)W_2(\rho_2, \tilde{\rho}). \tag{12.29}$$

□

Definition 12.3.4. Define the set $\mathcal{P}^G(\mathbb{R}^n) \subset \mathcal{P}^{\gamma,K_0}_{2,K_2}(\mathbb{R}^n)$ where $\mathcal{P}^{\gamma,K_0}_{2,K_2}(\mathbb{R}^n)$ is defined in Definition 3.1.11. For any $\rho \in \mathcal{P}^G(\mathbb{R}^n)$,

(i)

$$G(\rho) \leq K_0$$

(ii)

$$\int \|x\|^2 \rho(x) dx \leq K_2$$

.

The set $\mathcal{P}^G(\mathbb{R}^n)$ is sequentially compact by Lemma 3.3.7, by virtue of being a subset of $\mathcal{P}^{\gamma,K_0}_{2,K_2}(\mathbb{R}^n)$.

Lemma 12.3.5 (Fatou's Lemma)**.** *For non-negative functions f_n on a measure space with measure μ,*

$$\int_A \liminf_{n\to\infty} f_n d\mu \leq \liminf_{n\to\infty} \int_A f_n d\mu \tag{12.30}$$

for any measurable set A.

Lemma 12.3.6. *There exists a minimiser to the functional $G(\rho)$ on the set $\mathcal{P}^G(\mathbb{R}^n)$. In other words, the density $\tilde{\rho}$ such that*

$$G(\tilde{\rho}) := \inf\{G(\rho) : \rho \in \mathcal{P}^G(\mathbb{R}^n)\} \tag{12.31}$$

exists and lies within $\mathcal{P}^G(\mathbb{R}^n)$.

Proof. The functional is finite so let K_0 denote its upper bound, then $0 \leq G(\rho) \leq K_0$ for all $\rho \in \mathcal{P}^G(\mathbb{R}^n)$. All limit points of $\mathcal{P}^G(\mathbb{R}^n)$ are probability density functions in $\mathcal{P}(\mathbb{R})$, this follows from condition (ii) in the definition of the set $\mathcal{P}^G(\mathbb{R}^n)$ in combination with Fatou's Lemma. Thus there exists a sequence $\rho_n(x) \in \mathcal{P}^G(\mathbb{R}^n)$ such that $G(\rho_n)$ tends to the infimum,

$$\lim_{n\to\infty} G(\rho_n(x)) = \inf\{G(\rho) \ : \ \rho \in \mathcal{P}^G(\mathbb{R}^n)\}. \tag{12.32}$$

By the weak compactness of $\mathcal{P}^G(\mathbb{R}^n)$ the sequence ρ_n has a subsequence which weakly converges to the limit $\rho_\infty \in \mathcal{P}^G(\mathbb{R}^n)$. Fatou's lemma again establishes that $\mathcal{U}(\rho_n) \to \mathcal{U}(\rho_\infty)$. Lastly, if $\rho_n \to \rho_\infty$ weakly and

$$\lim_{R\to\infty} \limsup_{n\to\infty} \int_{\|x\|>R} \|x\|^2 \rho_n(x) dx = 0, \tag{12.33}$$

then by Villani [74, Thm. 7.12] the Wasserstein distance $W(\rho_n, \rho_2) \to W(\rho_\infty, \rho_2)$ weakly for any reference measure $\rho_2 \in \mathcal{P}(\mathbb{R}^n)$. And the above condition clearly holds in the case at hand due to the fact that Equation (3.24) holds for all $\rho \in \mathcal{P}^G(\mathbb{R}^n)$ and letting $R \to \infty$.

□

Lemma 12.3.7. *The minimiser to the functional $G(\rho)$ on the set $\mathcal{P}^G(\mathbb{R}^n)$ is unique.*

Proof. Consider two weakly convergent sequences $\rho_n \to \rho_\infty$ and $q_n \to q_\infty$, both tend to minimisers of $G(\rho)$. Let M denote this minimal value, $M := \inf\{G(\rho) : \rho \in \mathcal{P}^G(\mathbb{R}^n)\}$. By the definition of the set $\mathcal{P}^G(\mathbb{R}^n)$, for all $\rho \in \mathcal{P}^G(\mathbb{R}^n)$, $\int_{\mathbb{R}} \rho(x)^\gamma dx < K_0$, and so $\mathcal{P}^G(\mathbb{R}^n)$ is a subset of $L^\gamma(\mathbb{R}^n)$. The Clarkson inequality (Equation (3.16)) implies

$$\left(\left\| \frac{\rho_\infty + q_\infty}{2} \right\|_\gamma^{\gamma'} + \left\| \frac{\rho_\infty - q_\infty}{2} \right\|_\gamma^{\gamma'} \right)^{\frac{\gamma}{\gamma'}} \leq \frac{1}{2}\|\rho_\infty\|_\gamma^{\gamma'} + \frac{1}{2}\|q_\infty\|_\gamma^{\gamma'}. \tag{12.34}$$

Add the terms $\frac{1}{2}W(\rho_\infty, \rho_2)^2 + \frac{1}{2}W(q_\infty, \rho_2)^2$ on both sides. The right hand side is

$$\frac{1}{2}\|\rho_\infty\|_\gamma^{\gamma'} + \frac{1}{2}\|q_\infty\|_\gamma^{\gamma'} + \frac{1}{2}W(\rho_\infty, \rho_2)^2 + \frac{1}{2}W(q_\infty, \rho_2)^2 = \frac{1}{2}G(\rho_\infty) + \frac{1}{2}G(q_\infty) = M.$$

The Wasserstein distance is convex with respect to the linear structure on $\mathcal{P}(\mathbb{R}^n)$, therefore

$$W(\frac{\rho_\infty + q_\infty}{2}, \rho_2)^2 \leq \frac{1}{2}W(\rho_\infty, \rho_2)^2 + \frac{1}{2}W(q_\infty, \rho_2)^2$$

and because norms are non-negative, then somewhat trivially

$$\left\| \frac{\rho_\infty + q_\infty}{2} \right\|_\gamma^{\gamma} \leq \left(\left\| \frac{\rho_\infty + q_\infty}{2} \right\|_\gamma^{\gamma'} + \left\| \frac{\rho_\infty - q_\infty}{2} \right\|_\gamma^{\gamma'} \right)^{\frac{\gamma}{\gamma'}}.$$

Combining the last two inequalities

$$G\left(\frac{\rho_\infty + q_\infty}{2} \right) \leq \left(\left\| \frac{\rho_\infty + q_\infty}{2} \right\|_\gamma^{\gamma'} + \left\| \frac{\rho_\infty - q_\infty}{2} \right\|_\gamma^{\gamma'} \right)^{\frac{\gamma}{\gamma'}} + \frac{1}{2}W(\rho_\infty, \rho_2)^2 + \frac{1}{2}W(q_\infty, \rho_2)^2$$

and applying to Equation (12.34) with the Wasserstein terms added implies $G\left(\frac{\rho_\infty + q_\infty}{2}\right) \leq M$. Due to the convexity of $\mathcal{P}^G(\mathbb{R}^n)$, $\frac{\rho_\infty + q_\infty}{2} \in \mathcal{P}^G(\mathbb{R}^n)$, and so as M is the infimum

the inequality becomes an equality, and implies that

$$\left\| \frac{\rho_\infty - q_\infty}{2} \right\|_\gamma^\gamma = 0,$$

in other words $\rho_\infty = q_\infty$ a.e. and the minimiser is unique. □

12.3.2 Minimise the Wasserstein plus potential functional

The previous section established that there exists a minimiser to Equation (12.24). This section aims to provide a pushforward map that makes calculation of the minimising density possible. For this the idea of a differential of the functional $G(\rho)$ is required, and this needs the functional to be convex.

Displacement Convexity

The functional $G(\rho)$ must be displacement convex to define a subdifferential and attempt to produce a weak Euler–Lagrange equation for the transport map. The term displacement convex shifts the space under consideration from the density functions $\rho \in \mathcal{P}_2(\mathbb{R}^n)$ themselves to the space of transport maps between them. The linear interpolant is replaced with the interpolation function, and before proving the displacement convexity of $G(\rho)$ some features of the interpolation function are established.

Let $\psi(x)$ be the pushforward map between measures $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R}^n)$, so that $\psi(x)\#\rho_1(x)dx = \rho_2(x)dx$. From this, the interpolation function between measures ρ_1 and ρ_2 is defined,

$$\begin{aligned} \psi_t(x) :&[0,1] \to \{\text{pushforward maps } g \mid g : \mathcal{P}_2(\mathbb{R}^n) \to \mathcal{P}_2(\mathbb{R}^n)\} \\ &t \mapsto (1-t)x + t\psi(x). \end{aligned}$$

In addition to this, introduce the probability density $\rho_t := \psi_t\#\rho_1$ which is the pushforward of ρ_1 by ψ_t.

Lemma 12.3.8. *If $\psi(x)$ is the optimal map between two measures in $\mathcal{P}_2(\mathbb{R})$, then the function $\psi_t(x) : \mathbb{R}^n \to \mathbb{R}^n$ is monotone.*

Proof. By Brenier's theorem (Thm 4.1.12), the optimal map between measures in

$\mathcal{P}_2(\mathbb{R}^n)$ is the gradient of a convex function. Consider said convex function $\Phi : \mathbb{R}^n \to \mathbb{R}^n$, where $\nabla\Phi(x) = \psi(x)$. Then the function

$$\frac{1}{2}(1-t)\|x\|^2 + t\Phi(x)$$

is also convex. Take its gradient, $(1-t)x + t\nabla\Phi(x)$, which is monotone by properties of convex functions (Definition 4.1.14) and compare it with the definition of $\psi_t(x) = (1-t)x + t\psi(x)$. By the condition that $\nabla\Phi(x) = \psi(x)$, it is established that $\psi_t(x)$ is monotone. □

Lemma 12.3.9. *The determinant of the Jacobian of the interpolation function ψ_t is always positive.*

Proof. Recall Φ as defined in Lemma 12.3.8, the Jacobian of $\psi_t(x)$ can be expressed as $J_{\psi_t}(x) = (1-t)I + t\text{Hess }\Phi$ where Hess denotes the Hessian matrix of second order partial derivatives. Then the identity matrix is positive definite, and the Hessian of a convex function is positive semi-definite, that makes $J_{\psi_t}(x)$ positive definite. A positive definite matrix has a positive determinant. □

Lemma 12.3.10. *Let ψ_t be the interpolation function defined in Lemma 12.3.8 and let $\Delta(x,t) = det(J_{\psi_t}(x))$ denote the determinant of the Jacobian of ψ_t. Then for $\gamma \in (1,2]$, the function $t \mapsto \frac{1}{\Delta(x,t)^{\gamma-1}}$ is convex, and specifically*

$$\gamma\frac{\dot{\Delta}^2}{\Delta^2} - \frac{\ddot{\Delta}}{\Delta} \geq 0. \tag{12.35}$$

Proof. This proof is adapted from a preprint by Blower. The second derivative of $\frac{1}{\Delta(x,t)^{\gamma-1}}$ with respect to time is

$$\frac{d^2}{dt^2}\frac{1}{\Delta(x,t)^{\gamma-1}} = \frac{\gamma-1}{\Delta^{\gamma-1}}\left(\gamma\frac{\dot{\Delta}^2}{\Delta^2} - \frac{\ddot{\Delta}}{\Delta}\right). \tag{12.36}$$

The aim is to show this is positive, to that goal, the constant $(\gamma-1)$ is positive, as is the determinant of a monotone function. The function ψ_t is monotone by Lemma 12.3.8 so Δ is positive. All that remains is the term in brackets.

Recall $\psi(x)$ is the gradient of the convex function Φ, so the Jacobian of $\psi_t(x)$ can be expressed as $J_{\psi_t}(x) = (1-t)I + t\text{Hess}\Phi$ where Hess denotes the Hessian matrix

of second order partial derivatives. To take the derivative of Δ consider the relation between the trace and determinant of the logarithm of a matrix [23, p.1029],

$$\log \det A = \log \left(\prod_i \lambda_i \right) = \sum_i \log(\lambda_i) = \text{trace}(\log A)$$

where $A \in M_n(\mathbb{R})$ with eigenvalues λ_i. Employing this relation for $A = \Delta$,

$$\begin{aligned} \log(\Delta) &= \text{trace} \log \left((1-t)I + t\text{Hess}\Phi \right), \\ \frac{\dot{\Delta}}{\Delta} &= \text{trace} \left(\left((1-t)I + t\text{Hess}\Phi \right)^{-1} \left(\text{Hess}\Phi - I \right) \right). \end{aligned}$$

One can take the second derivative similarly, where the notation $J_{\psi_t}(x)$ is again employed for visual clarity,

$$\frac{\ddot{\Delta}}{\Delta} - \frac{\dot{\Delta}^2}{\Delta^2} = -\text{trace} \left(J_{\psi_t}^{-1}(x) \frac{dJ_{\psi_t}(x)}{dt} J_{\psi_t}^{-1}(x) \frac{dJ_{\psi_t}(x)}{dt} \right).$$

The relation $\text{trace}(ABAB) = \text{trace}(A^{\frac{1}{2}}BABA^{\frac{1}{2}}) = \text{trace}\left(A^{\frac{1}{2}}BA^{\frac{1}{2}}\right)^2$ is predicated on the postive semidefinite-ness of the matrix A, by Lemma 12.3.9 Δ is positive definite,

$$\frac{\ddot{\Delta}}{\Delta} - \frac{\dot{\Delta}^2}{\Delta^2} = -\text{trace} \left(J_{\psi_t}^{-1}(x)^{\frac{1}{2}} \frac{dJ_{\psi_t}(x)}{dt} J_{\psi_t}^{-1}(x)^{\frac{1}{2}} \right)^2.$$

The term is squared and thus the trace is positive. The relation,

$$\begin{aligned} \left(\gamma \frac{\dot{\Delta}^2}{\Delta^2} - \frac{\ddot{\Delta}}{\Delta} \right) &= (\gamma - 1) \frac{\dot{\Delta}^2}{\Delta^2} - \left(\frac{\ddot{\Delta}}{\Delta} - \frac{\dot{\Delta}^2}{\Delta^2} \right) \\ &= (\gamma - 1) \frac{\dot{\Delta}^2}{\Delta^2} + \text{trace} \left(J_{\psi_t}^{-1}(x)^{\frac{1}{2}} \frac{dJ_{\psi_t}(x)}{dt} J_{\psi_t}^{-1}(x)^{\frac{1}{2}} \right)^2 \geq 0, \end{aligned}$$

establishes the positivity of the term in brackets in (12.36) and concludes the proof. □

Definition 12.3.11 (Displacement convex)**.** (i) A set of measures $P \subset \mathcal{P}_2(\mathbb{R}^n)$ is *displacement convex* if the interpolation function between any two measures in the set, denoted ψ_t, produces another measure in the set (for all $t \in [0,1]$) when acting as a pushforward.

(ii) A functional, G, defined on a displacement convex subset of $\mathcal{P}_2(\mathbb{R}^n)$ is displacement convex if the map $t \mapsto G(\rho_t(x))$ is convex for all $\rho_1, \rho_2 \in \mathcal{P}_2(\mathbb{R}^n)$. The measure $\rho_t(x) = \psi_t \# \rho_1(x)$ where ψ_t interpolates between ρ_1 and ρ_2.

Lemma 12.3.12. *The set $\mathcal{P}_2(\mathbb{R}^n)$ is displacement convex.*

Proof. The pushforward measure $\psi_t \# \rho_1$ where $\rho_1 \in \mathcal{P}_2(\mathbb{R}^n)$ is in $\mathcal{P}_2(\mathbb{R}^n)$. Consider any $A \in \mathcal{B}(\mathbb{R}^n)$, then $(\psi_t \# \rho_1)(A) = \rho_1(\psi_t^{-1}(A))$, the preimage of A under ψ_t is the set $\{x \in \mathbb{R} : (1-t)x + t\psi(x) \in A\}$ and this set is a translation of the preimage of A under ψ which is in $\mathcal{B}(\mathbb{R}^n)$ by the definition of $\rho_2 \in \mathcal{P}_2(\mathbb{R}^n)$. The sigma algebra $\mathcal{B}(\mathbb{R}^n)$ is closed under translations. □

Lemma 12.3.13. *Consider the function $U(\rho) = \rho^\gamma$ with $\gamma \in (1, 2]$. If $\psi_t(x)$ is the interpolation function as defined in Lemma 12.3.8 and $\rho_t = \psi_t \# \rho_1$ for $\rho_0 \in \mathcal{P}_2(\mathbb{R}^n)$ then the functional $\mathcal{U}(\rho) = \int U(\rho)\mu(dx)$ is displacement convex on $\mathcal{P}_2(\mathbb{R}^n)$.*

Proof. Establish the displacement convexity of $\mathcal{U}$ by showing $t \mapsto \mathcal{U}(\rho_t)$ is convex, or $\frac{d^2\mathcal{U}(\rho_t)}{dt^2} \geq 0$ as discussed in Lemma 4.1.15. Consider $\rho \in \mathcal{P}_2(\mathbb{R}^n)$ and let this represent the initial probability density for $\mathcal{U}$, Consider $\psi_t(x) : \mathbb{R}^n \to \mathbb{R}^n$, the interpolation function between ρ and ρ_1. Introduce a new function Θ, a rescaling of the derivative of $U(x)$, $x\Theta(x) = U(x)$, making $\Theta(x) = x^{\gamma-1}$, and so $\Theta : [0, \infty) \to [0, \infty)$ is increasing. This allows $\mathcal{U}$ to be expressed as an integral with respect to the measure ρ_t and therefore one can apply a pushforward to the measures,

$$\begin{aligned}
\mathcal{U}(\rho_t) &= \int \Theta(\rho_t(x))\rho_t(x)dx, \\
&= \int \Theta(\rho_t(\varphi_t(x)))\rho_1(x)dx, \\
&= \int \Theta\left(\frac{\rho_0(x)}{\Delta(x,t)}\right)\rho_1(dx).
\end{aligned}$$

The second step was a Jacobian change of variables which was applied under the caveats of Lemma 4.1.18. Calculate the first order derivative,

$$\frac{d}{dt}\mathcal{U}(\rho_t) = \int -\Theta'\left(\frac{\rho_1(x)}{\Delta(x,t)}\right)\frac{\rho_1(x)\dot{\Delta}(x,t)}{\Delta(x,t)^2}\rho_1(dx).$$

For the second order derivative the arguments are supressed.

$$\frac{d^2}{dt^2}\mathcal{U}(\rho_t) = \int \Theta''\left(\frac{\rho_1}{\Delta}\right)\left(\frac{\rho_1\dot{\Delta}}{\Delta^2}\right)^2 \rho_1(dx) + \int \Theta'\left(\frac{\rho_1}{\Delta}\right)\left(\frac{2\rho_1\dot{\Delta}^2}{\Delta^3} - \frac{\rho_1\ddot{\Delta}}{\Delta^2}\right)\rho_1(dx),$$
$$= \int \left(\Theta''\left(\frac{\rho_1}{\Delta}\right)\frac{\rho_1}{\Delta} + \Theta'\left(\frac{\rho_1}{\Delta}\right)\right)\frac{\rho_1\dot{\Delta}^2}{\Delta^3} + \Theta'\left(\frac{\rho_1}{\Delta}\right)\frac{\rho_1}{\Delta}\left(\frac{\dot{\Delta}^2}{\Delta^2} - \frac{\rho_1\ddot{\Delta}}{\Delta^2}\right)\rho_1(dx).$$

The rest of the proof is to establish the positivity of this integral. The function $\Theta(e^r)$ is convex with respect to r, as $\frac{d^2}{dr^2}\Theta(e^r) = (\gamma-1)^2 e^{r(\gamma-1)} \geq 0$. If $\Theta(e^r)$ is convex, then $\frac{d^2}{dr^2}\Theta(e^r) = \Theta''(e^r)e^r - \theta'(e^r) \geq 0$ and this holds for all r in $\mathbb{R}$. Furthermore, as ρ is a density function of a probability measure, $\rho : \mathbb{R}^n \to [0,\infty)$ and so for all values of ρ, $\Theta''(\rho)\rho - \Theta'(\rho) \geq 0$.

The determinant of the Jacobian of a monotone function is positive, see Lemma 12.3.8. Thus, $\frac{\rho_1}{\Delta}$ is positive. The map $t \mapsto \frac{1}{\Delta(x,t)^{\gamma-1}}$ is convex by Lemma 12.3.10, and this implies

$$\left(\gamma\frac{\dot{\Delta}^2}{\Delta^2} - \frac{\ddot{\Delta}}{\Delta^2}\right) \geq 0,$$

and so is the term in brackets. This makes each term in the integral positive, implying the second order derivative of $\mathcal{U}(\rho_t)$ is positive and so $\mathcal{U}(\rho_t)$ is displacement convex. □

The variational derivative of $G(\rho)$

With the convexity of the functional $\mathcal{U}(\rho)$ now established, under suitable assumptions the variational derivative of $G(\rho)$ can be calculated. The Wasserstein distance can be differentiated by the following theorem.

Theorem 12.3.14 (Differentiability of W^2)**.** *Villani [74, p268]. Given $\rho_2 \in \mathcal{P}_2(\mathbb{R}^n)$ and the path $\rho_t : [0,1] \to \mathcal{P}_2(\mathbb{R}^n)$ which is absolutely continuous and twice differentiable with respect to t and is a weak solution to*

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t u_t) = 0, \tag{12.37}$$

where $u_t(x)$ is C^1 in x and t. Then if $\rho_0 \in \mathcal{P}_2(\mathbb{R}^n)$, the variational derivative of the

functional $F(\rho) = W(\rho, \rho_2)^2$ *is given by*

$$\begin{aligned}\delta F(\rho; u) &= 2\int \langle \nabla\varphi(x) - x, u_t \circ \nabla\varphi(x)\rangle d\rho_2, \\ &= 2\int \langle y - \nabla\varphi^*(y), u_t(y)\rangle d\rho_0\end{aligned}$$

where $\nabla\varphi$ *is the optimal map between* ρ_2 *and* ρ_0*, and* φ^* *denotes the conjugate* φ *in the sense of the Legendre transform.*

As established in in Section 12.3.1, there exists a minima to $G(\rho)$. Defining $G(\rho)$ in terms of $\rho_2 \in K$,

$$G(\rho) := \frac{1}{2\tau} W_2(\rho, \rho_2)^2 + \mathcal{U}(\rho). \tag{12.38}$$

Then this minima is denoted ρ^τ,

$$\rho^\tau = \arg\min_\rho \left\{ G(\rho) \ : \ \rho \in K \right\}. \tag{12.39}$$

If we take a variation around ρ^τ of $G(\rho)$ then we can determine some conditions on the transportation map.

Proposition 12.3.15. *Let* $\nabla\varphi$ *denote the optimal transport map between* ρ_2 *and* ρ^τ*. Take the variation of* $G(\rho)$ *along the path* $\rho_t : [0,1] \to \mathcal{P}_2(\mathbb{R}^n)$ *as specified in Theorem 12.3.14 starting at the minimiser,* $\rho_t|_{t=0} = \rho^\tau$*. The condition that* $\delta G(\rho) = 0$ *produces the Euler–Lagrange equation,*

$$\frac{1}{\tau}(Id - \nabla\varphi^*)(x) + \nabla\frac{\delta\mathcal{U}(\rho)}{\delta\rho} = 0 \tag{12.40}$$

where the second term is interpreted weakly.

To explain, notationally, to calculate the variational derivative of $\mathcal{U}$ along the path ρ_t, essentially a variation along $\rho + \epsilon\frac{\partial\rho_t}{\partial t}$ will give you

$$\delta\mathcal{U}(\rho) = \int \frac{\delta\mathcal{U}}{\delta\rho}\frac{\partial\rho_t}{\partial t}\Big|_{t=0} dx = \int \frac{\delta\mathcal{U}}{\delta\rho}\nabla\cdot(\rho_t u_t)\Big|_{t=0} dx = \int \langle \nabla\frac{\delta\mathcal{U}}{\delta\rho}, u_t\rangle d\rho_0 \tag{12.41}$$

So formally,

$$\delta G(\rho, u) = \frac{1}{\tau}\int \langle y - \nabla\varphi^*(y), u_t(y)\rangle d\rho_0 + \int \langle \nabla \frac{\delta \mathcal{U}}{\delta \rho}, u_t\rangle d\rho_0 \tag{12.42}$$

The Euler Lagrange equations for this functional give conditions for which the value of ρ will minimise $G(\rho)$, and come about by setting $\delta G(\rho, u) = 0$. Thus Equation (12.40) holds in a weak sense.

Therefore, everywhere except on sets of zero measure, the map, denoted $\nabla\varphi^*$, which pushes ρ^τ forward to ρ_2 can be defined by Equation (12.40). Rearranging that equation gives $\nabla\varphi^* = Id + \tau\nabla\delta\mathcal{U}/\delta\rho$, an implicit map for this pushforward. Note however that it is the dual map, the map which pushes the optimal ρ forward to ρ_2. Therefore the optimal ρ is given by:

$$\rho^\tau = (\nabla\varphi)\#\rho_2 = \left(Id + \tau\nabla\frac{\delta\mathcal{U}}{\delta\rho}\right)^{-1}\#\rho_2 \tag{12.43}$$

In the case of interest, we seek to minimise the functional with a scale factor.

Lemma 12.3.16. *Let the constant τ be rescaled to $2\tau^2/3$, then consider the new functional*

$$\frac{3}{4\tau^2}W_2(\rho, \rho_2)^2 + \mathcal{U}(\rho) \tag{12.44}$$

with a potential $U(\rho) = \rho^\gamma$ for some power of $\gamma \in (1, 2]$. The density which minimises this functional is

$$\rho^\tau = (\nabla\varphi)\#\rho_2 = \left(Id + \frac{2\tau^2}{3}\nabla U'\right)^{-1}\#\rho_2 \tag{12.45}$$

The proof of the lemma follows from the discussion of the functional in this and the preceding section. Rescaling the constant changes none of the analysis, and the functional derivative is treated in the following lemma.

Lemma 12.3.17. *The functional $\mathcal{U}(\rho)$ with the potential $U(\rho) = \rho^\gamma$,*

$$\mathcal{U}(\rho) = \int \rho^\gamma dx, \tag{12.46}$$

has a functional derivative, $\frac{\delta\mathcal{U}}{\delta\rho} = U'(\rho) = \gamma\rho^{(\gamma-1)}$.

Proof. The Gateaux derivative of $\mathcal{U}$ in the direction of an arbitrary test function h is

$$\begin{aligned}\delta\mathcal{U} &= \lim_{\epsilon\to 0}\frac{\mathcal{U}(\rho+\epsilon h)-\mathcal{U}(\rho)}{\epsilon}\\ &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int\left((\rho+\epsilon h)^{\gamma}-\rho^{\gamma}\right)dx\\ &= \lim_{\epsilon\to 0}\int\left(\gamma\rho^{(\gamma-1)}h+o(\epsilon)\right)dx\\ &= \int\gamma\rho^{(\gamma-1)}h dx.\end{aligned}$$

From which the functional derivative is $\gamma\rho^{(\gamma-1)} = U'(\rho)$. □

12.3.3 The velocity update

Here we will prove proposition 12.3.1, which seeks to minimise the functional:

$$\mu^{\tau} = \arg\min_{\mu^*}\{\mathbf{A}_{\tau}(\mu,\mu^*)^2 + \mathcal{U}(\mu^*) \ : \ \mu^* \in \mathcal{P}(\mathbb{TR}^n)\}. \tag{12.47}$$

Lemma 12.3.18. *Under the assumptions of Proposition* (12.3.1)*, the density ρ^{τ} which is the marginal of*

$$\mu^{\tau} = \arg\min_{\mu^*}\{\mathbf{A}_{\tau}(\mu,\mu^*)^2 + \mathcal{U}(\mu^*) \ : \ \mu^* \in \mathcal{P}(\mathbb{TR}^n)\}. \tag{12.48}$$

is given by

$$u^{\tau} = u\circ(Id+\tau u)^{-1}\circ\left(Id+\frac{2\tau^2}{3}\nabla U'(\rho^{\tau})\right) - \tau\nabla U'(\rho^{\tau}). \tag{12.49}$$

Proof. The first step in the proof is to change the functional A_{τ} into W_{τ} by use of the transport map X_{τ} given in Equation (12.12). Define $\mu_2 = X_{\tau}\#\mu$ and then $W_{\tau}((X_{\tau}\#\mu),\mu^*) = W_{\tau}(\mu_2,\mu^*) = A_{\tau}(\mu,\mu^*)$. By definition, μ_2 has marginal $\mu_2|_1 = (Id+\tau u)\#\rho := \rho_2$. An application of Equation (12.16) allows the minimisation prob-

lem to be expressed in terms of densities only,

$$\arg\min_{\rho^*}\{\frac{3}{2\tau^2}\mathbf{W}_\tau(\rho_2,\rho^*)^2+\mathcal{U}(\rho^*)\ :\ \rho^*\in\mathcal{P}(\mathbb{R}^n)\}. \tag{12.50}$$

Note that $\mathcal{U}(\mu)$ has no dependence on a velocity coordinate, so it can be expressed as $\mathcal{U}(\rho)$ without any loss of generality.

The functional given in Equation 12.50 has a unique minimiser, ρ^τ as established in Section 12.3.1, and thus the use of an arg min is a defined mathematical operation. An application of Lemma 12.3.13 to this problem gives the pushforward of the measure ρ_2 to ρ_τ, the measure that minimises Equation (12.50),

$$\rho^\tau=\left(Id+\frac{2\tau^2}{3}\nabla U'(\rho^\tau)\right)^{-1}\#\rho_2. \tag{12.51}$$

The proof of Equation (13.5) finishes with a composition of this pushforward with the pushforward of ρ to ρ_2 given earlier in the proof $(Id+\tau u)\#\rho=\rho_2$.

The updated velocity requires Proposition 12.2.5, which shows that the minimising velocity is a function β of the original density and velocity and the new density. Having now developed an analysis of the pushforward maps (in some cases optimal) to map between the spaces referred to in Proposition 12.2.5 by the coordinates (x_1,u_1,x_2) the function $\beta(x_1,u_1,x_2)$ must be expressed as a function of coordinates in the same measure space. As in said Proposition, assume $\gamma(dx_1,du_1,dx_2,du_2)$ is a transport plan which minimises $W_\tau(\mu_2,\mu^*)$. By Lemma 12.3.16, the first marginal of the minimising μ^τ is ρ^τ. Therefore, $W_\tau(\mu,\mu^\tau)$ can be expressed in terms of the transport maps. Consider W_τ as in the Proposition,

$$W_\tau(\mu_2,\mu^\tau)=\int_{\mathbb{TR}^n\times\mathbb{TR}^n}W_\tau(x_1,u_1,x_2,u_2)\gamma(dx_1,du_1,dx_2,du_2)$$
$$W_\tau(\mu_2,\mu^\tau)=\int_{\mathbb{TR}^n\times\mathbb{TR}^n}W_\tau(\nabla\varphi^*(x),(u\circ\nabla\varphi^*)(x),x,\beta(\nabla\varphi^*(x),(u\circ\nabla\varphi^*)(x),x))\rho^\tau(x)dx.$$

where $\nabla\varphi^*$ is the optimal pushforward map $\nabla\varphi^*\#\rho^\tau=\rho_2$. As such, the velocity update β needs to be understood in composition with each of the pushforward maps. Furthermore, as the final velocity marginal should be a transport map $u^\tau\#\rho^\tau$ the integral

should be in terms of ρ^τ, and then the optimal pushforward of ρ^τ to the x_2 coordinate is given by,

$$u^\tau(x) = \beta\left((\nabla\varphi^*)(x), (u_2 \circ \nabla\varphi^*)(x), (Id)(x)\right) \tag{12.52}$$

$$= (u_2 \circ \nabla\varphi^*)(x) + \frac{3}{2\tau}(Id - \nabla\varphi^*)(x) \tag{12.53}$$

Thus, the definition of $\nabla\varphi^*$ is applied from Lemma 12.3.16.

$$u^\tau = (u_2 \circ \left(Id + \frac{2\tau^2}{3}\nabla U'(\rho^\tau)\right)) + \frac{3}{2\tau}(Id - \left(Id + \frac{2\tau^2}{3}\nabla U'(\rho^\tau)\right)) \tag{12.54}$$

$$u^\tau = (u_2 \circ \left(Id + \frac{2\tau^2}{3}\nabla U'(\rho^\tau)\right)) - \tau\nabla U'(\rho^\tau). \tag{12.55}$$

The final correction needed is for the initial distribution μ_2 to be recognised as the free transportation of the original distribution μ_1, in other words moving from the functional $\mathbf{W}_\tau$ to $\mathbf{A}_\tau$. Defined in terms of densities, $mu_2 = (Id \times u_2)\#\rho_2$ whereas $\mu_1 = (Id \times u_1)\#\rho_1$. Now, u_2 and u_1 are the exact same function, just acting as pushforwards on different measures.

The final correction needed is for the velocity u_2 to be defined in terms of the initial velocity of the original distribution ρ_1, denoted u_1. The velocity field remains unchanged, it is just evaluated at different positions, $u_1\#\rho_1 = u_2\#\rho_2$. Therefore, $(u_2)\#\rho_2 = (u_2 \circ (Id + \tau u_1))\#\rho_1 = (u_1)\#\rho_1$, so that $u_2 \circ (Id + \tau u_1) = u_1$ or equivalently $u_2 = u_1 \circ (Id + \tau u_1)^{-1}$ which holds ρ_1 a.e.

□

12.4 Limitations

The largest limitation of the method is in the assumption that the free flow of the fluid will produce a second measure in $\mathcal{P}_2(\mathbb{R}^n)$. See Remark 12.2.4, but this free flow can allow the fluid to pass through itself in non physical ways without penalty. Whereas as soon as a trajectory crosses, the transport map between measures can no longer be invertible. There is an interesting work-around discussed by Westdickenberg and Wilkening [75] in which the velocity is replaced by a different velocity distribution which transports the measure to the same locations, only optimally.

Remark 12.4.1. A limitation of the transport method is around the support of the measure. A requirement for Brenier's theorem is that the measures be absolutely continuous with respect to Lebesgue measure on the same domain, so if the domain expands or bifurcates this is lost. To get around this problem, and others, a minimum density is sometimes defined over the whole domain of interest, so while the majority of the fluid is lying on a smaller subsection of the domain the density is non-zero almost everywhere. This is not a huge limitation for rarefied gas dynamics, though modelling shock waves due to explosions for instance there are moments of extremely low densities.

Remark 12.4.2. Another physically realistic problem which cannot be modelled by this simulation is Cavitation [70]. Cavitation occurs when fluid is moving at sufficiently high velocities to produce gas pockets of low density, for example when ship propellers are given too much torque relative to the area of displaced water. In simulation this implies high velocities may change the nature of the internal energy of the fluid (when it becomes gas from liquid), and then also changes the support of the measure as in the previous remark.

Despite the limitations, the method described expands on the class of solution that can be modelled compared to classical numerical methods, and the approach of minimising an energy functional has had much success as an analytic framework to view the dynamics of systems in Physics.

Chapter 13

A one dimensional transport algorithm

In this chapter I describe an implementation I have made of the numerical method introduced by Gangbo [27] and discussed in the previous chapter. The dam break problem is chosen for its analytical solutions, as discussed in Chapter 7, and so comparison between the numerical solution and a known analytical solution to a system of PDEs can be made.

13.1 Numerical method to solve Euler equations

One step of the algorithm proposed by Westdickenburg and Wilkening [75] is given below. The accompanying theory is supplied in the preceding chapter.

(1) Begin with the data from the last step $(\rho^n, \mathbf{u}^n)$ and choose a timestep $\tau \in [h/2, h]$ which allows the update of the density according to the motion induced by the velocity field,

$$(\mathrm{Id} + \tau \mathbf{u}^n)\#(\rho^n dx) = \hat{\rho}^n dx, \tag{13.1}$$

such that $\hat{\rho}^n dx$ is absolutely continuous with respect to Lebesgue measure.

(2) Update the velocity so that the new velocity, $\hat{\mathbf{u}}^n$ induces an optimal transport map

$(\mathrm{Id} + \tau\hat{\mathbf{u}}^n)$ between ρ and $\hat{\rho}$. In other words, find $\hat{\mathbf{u}}^n \in L^2(\mathbb{R}^n, \rho^n)$ such that

$$(\mathrm{Id} + \tau\hat{\mathbf{u}}^n)\#(\rho^n dx) = \hat{\rho}^n dx \quad \text{and} \tag{13.2}$$

$$\int_{\mathbb{R}^n} |\tau\hat{\mathbf{u}}^n|^2 \rho dx = W(\rho^n, \hat{\rho}^n)^2. \tag{13.3}$$

Equation (13.3) is the condition for which $(\mathrm{Id} + \tau\hat{\mathbf{u}}^n)$ is the optimal map between ρ^n and $\hat{\rho}^n$. Existence and uniqueness of this map follows from Brenier's theorem [74] (Theorem 4.1.12).

(3) Update the density by solving

$$\rho^{n+1} = \mathrm{argmin}_\rho \{\frac{3}{4\tau^2} W(\rho, \hat{\rho}^n)^2 + \mathcal{U}(\rho)\} \tag{13.4}$$

as a convex optimisation problem. ρ^{n+1} is uniquely determined, and can be expressed implicitly as a pushforward of ρ^n by,

$$\rho^{n+1} dx = \left((Id + \frac{2\tau^2}{3} \nabla U'(\rho^{n+1}))^{-1} \circ (Id + \tau u) \right) \# \rho^n dx. \tag{13.5}$$

(4) The velocity is updated analogously,

$$u^{n+1} = u^n \circ (Id + \tau u^n)^{-1} \circ \left(Id + \frac{2\tau^2}{3} \nabla U'(\rho^{n+1}) \right) - \tau \nabla U'(\rho^{n+1}). \tag{13.6}$$

(5) The previous steps are repeated with ρ^{n+1} and u^{n+1} as initial data.

Step (2) of this proposed algorithm is additional to the theory discussed in the previous chapter. In the step, the initial velocity field is adapted according to Dafermos' entropy rate admissibility criterion [17]. This condition essentially says that the total energy of the system should be dissipated as quickly as possible. In this step, the velocity field which produces the fluid density ρ_2 is replaced with the velocity field that would induce the optimal transport of the density ρ to the new ρ_2. The new velocity field will minimise the kinetic energy of the fluid, as shown in Equation (13.3).

This extra condition on the choice of admissible solutions to the problem circumvents a theoretical issue with the choice of acceleration cost metric. Namely, that the

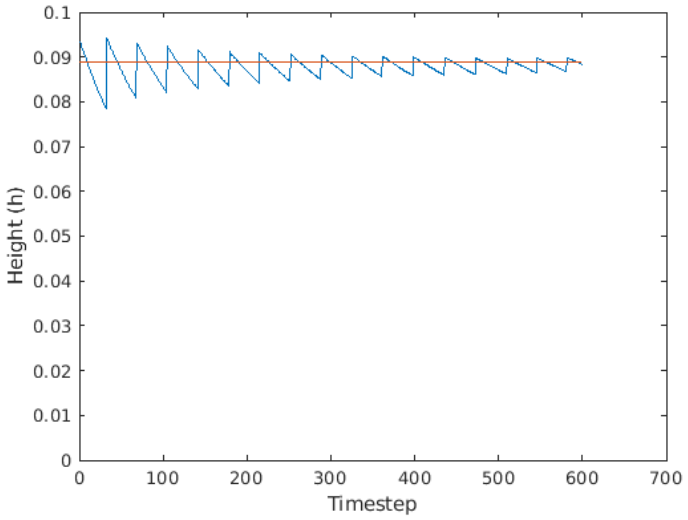

Figure 13.1: The plot shows the height of the water at timesteps 100 to 700 given by the numerical approximation (Blue) and the predicted height given by the Ritter solution (Red).

acceleration cost metric does not penalise fluid travelling directly through itself (with no simulated collision). This is in contrast with the Wasserstein metric with the $2-$norm, in which, provided the two densities in question are absolutely continuous with respect to Lebesgue measure, the optimal map between them is monotone (Theorem 4.1.12).

Implementation

The implementation of this method in one dimension carried out for this analysis is described by the following steps.

The first step is to define the probability density function. Let the vector $x_n = (x_0, x_1, x_2, ..., x_N)$ denote the positions of the edges of each of the intervals. Let the total mass be normalised and distributed equally over each interval. The number of intervals is one less than the number of edges of intervals and thus,

$$\frac{1}{N} = \int_{x_{i-1}}^{x_i} \rho(x) dx \tag{13.7}$$

over each interval. The density can therefore be approximated by a step function,

$$\rho(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{x_i - x_{i-1}} \mathbb{I}_{(x_{i-1}, x_i]} \tag{13.8}$$

The initial velocities are also approximated as a step function with $N+1$ steps and thus expressed as a vector $u_n = (u_0, ...u_N)$.

The algorithm is as follows:

(1) Starting with positions $x_n^0 = (x_1^0, x_2^0, ..., x_N^0)$ which define a piecewise constant density of the fluid $\rho(x)$, and an initial velocity field $u_n^0 = (u_1^0, u_2^0, ..., u_N^0)$.

(2) Free transport the density according to the pushforward map $(\mathrm{Id} + h u_n^0)$ by defining a new set of endpoints to intervals, $y_i = x_i + h u_i$. Then calculate $\hat{\mathbf{u}}$ as defined in Equation (13.3), n one dimension this amounts to sorting the y_i into ascending order; if $\sigma(i)$ is the permutation of indices for which y_i is ascending then $\hat{x}_i := y_{\sigma(i)}$ and $\hat{u}_i := (\hat{x}_i - x_i)/h$.

(3) Next the positions z_i (that define $\hat{\rho}$) which minimise the energy functional given in Equation (13.4) are found via convex optimisation. Given the piecewise constant nature of ρ and $\hat{\rho}$ in this case, the functional reduces to the function

$$F(z) = \frac{3}{4h^2}\sum_{i=1}^{N} \|z_i - \hat{x}_i\|^2 + \sum_{i=1}^{N} U(\frac{1}{z_i - z_{i-1}})(z_i - z_{i-1}). \tag{13.9}$$

This equation further reduces when $U(\rho) = \rho^\gamma$. The paper advises solving this equation by a trust region newton method.

(4) Then the minimiser $z = (z_1, z_2, ..., z_N)$ defines the new positions x_n^1 and the subsequent velocity is also updated by

$$u^{k+1} = u^k + \frac{3}{2h}(x^{k+1} - \hat{x}^k), \tag{13.10}$$

and the previous steps are thus repeated with x_n^1 and u^1 to increment the solution.

Details of the implemented algorithm for one dimension with $\gamma = 2$ can be found in Algorithm 13.1.

Algorithm 13.1 One dimensional Transport

1: $N \leftarrow$ number of intervals to discretise to.
2: $h \leftarrow$ length of a timestep.
3: $T_s \leftarrow$ number of timesteps.
4: $x \leftarrow (x_1^0, x_2^0, ..., x_N^0)$, the initial density.
5: $u \leftarrow (u_1^0, u_2^0, ..., u_N^0)$ the initial velocity field.
6: $F_x(z)$ a function of $z \in \mathbb{R}^n$,

$$F(z) = \frac{3}{4h^2} \sum_{i=1}^{N} \|z_i - x_i\|^2 + \sum_{i=1}^{N} \frac{1}{z_i - z_{i-1}}. \tag{13.11}$$

7: **for** condition $k = 1 : T_s$ **do**
8: $y \leftarrow x + h * u$ as a vector operation.
9: $y \leftarrow \text{sort}(y)$ where the sort function sorts y into ascending order.
10: $v \leftarrow \frac{y-x}{h}$.
11: $x \leftarrow \min_z(F_y(z))$ using trust region method.
12: $u \leftarrow v + \frac{3}{2h}(x - y)$.
13: **end for**

13.2 Comparison with the dam break problem

A useful method to check the accuracy of a numerical method is to compare the method with some sort of known solution. The dam break problem offers a problem for which there is a known solution, see Chapter 7.

It is well known that the Ritter solution to the Saint-Venant equations is

$$u = \frac{2}{3}\left(\frac{x}{t} + c_0\right) \tag{13.12}$$

$$h = \frac{1}{9g}\left(2c_0 - \frac{x}{t}\right)^2, \tag{13.13}$$

with the constant c_0 defined to be $c_0 = \sqrt{gh_0}$, where h_0 is the initial height of the reservoir, and again g is the gravitational constant.

Equation (13.13) gives the profile of the water in the immediacy of the dam at each point in time. Our numerical solution to the problem (as outlined in section 13.1) relies on a step function which is characterised by the end points of each interval. At $t = 0$ the water is all behind the dam, Equation (13.13) is only defined for $t > 0$. As such, to compare the Ritter solution with our numeric simulation, a minimum number of steps

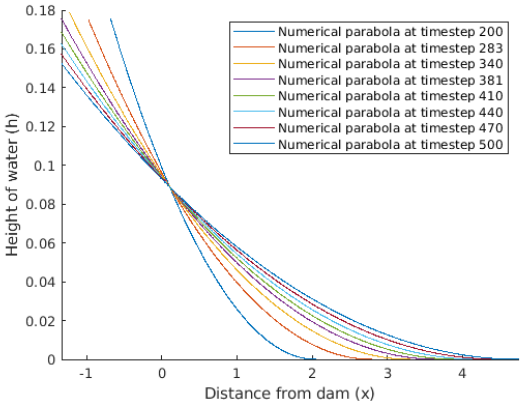

Figure 13.2: The plot shows the height of the water over a relevant section of the domain, with $x = 0$ being the former location of the dam. The flow profile is given for a number of timesteps, with a timestep of 10^{-3} this means timesteps have units of milliseconds (ms).

have to pass the dam before any inferences about the shape of the curve can be made.

One notable attribute of the Ritter solution to the dam break problem is the constant height of the water at the location of the dam. The height at $x = 0$ is equal to $h = \frac{4}{9}h_0$ at all times, which can be seen from Equation (13.13). Figure 13.1 shows how the height of water at the dam in the numerical simulation compares to the theoretical value. The resemblance to a sawtooth function comes from the evaluation of a step function at $x = 0$ as time increases and the steps get wider; the jumps are when a step passes the $x = 0$ mark. The reducing amplitude is a product of the flow profile flattening. The fact that the numerical data is always close to the theoretical value does suggest that the numerical value may converge to $\frac{4}{9}h_0$ if one was to increase the number of steps ad infinitum.

Another attribute of the Ritter solution is how at each fixed time, the profile of the water is a parabola on the x interval $[0, 2c_0t]$. The numerical method used seems to distort the spacial scale of the problem, but this relationship persists. Figure 13.2 shows the profile of the water at 8 timesteps chosen so that the sawtooth (blue) and constant (red) line intersect in Figure 13.1.

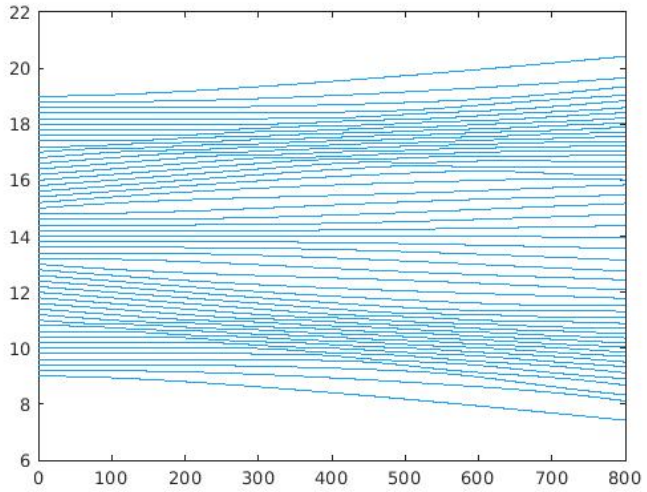

Figure 13.3: The graph shows the trajectories of a one dimensional flow, the x axis being time. The flow starts at uniform density and stationary apart from two equally sized steps. The velocity of the fluid on these two steps is equal and pointing outwards.

13.2.1 Numerical solutions for varying initial conditions

The figures in this section include lines which are the pathlines of the fluid. Pathlines show the trajectory of a particle located at the origin of said pathline. The previous section established that the algorithm proposed for solving the isentropic Euler equations behaves correctly on the initial data given by the dam break problem. Therefore an exploration of other similar initial conditions which do not have closed form solutions is motivated. I have run the algorithm on initial data resembling a dam break at two ends of an interval, and explored subintervals of non zero initial velocity.

The figures of this section show that if the timestep is chosen appropriately, then pathlines of the fluid do not cross even when the free transport of fluid parcels along their current trajectories would imply they should. Though this is done simply by reallocating velocities for these fluid parcels so that they end up in same final configuration without crossing. The internal energy is not used in this step of the algorithm even though it is precisely the mechanism by which the fluid is restricted from piling up on sets of measure zero. This is established in Section 12.3.1; the probability measure lives in the compact space K, but this requires initial velocity fields which are non-degenerate as in Remark 12.2.4.

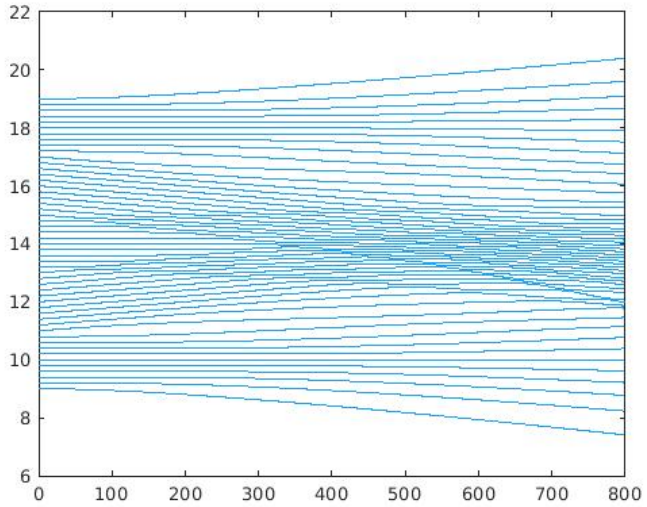

Figure 13.4: The graph shows the trajectories of a one dimensional flow, the x axis being time. The flow starts at uniform density and stationary apart from two equally sized steps. The velocity of the fluid on these two steps is equal and pointing inwards.

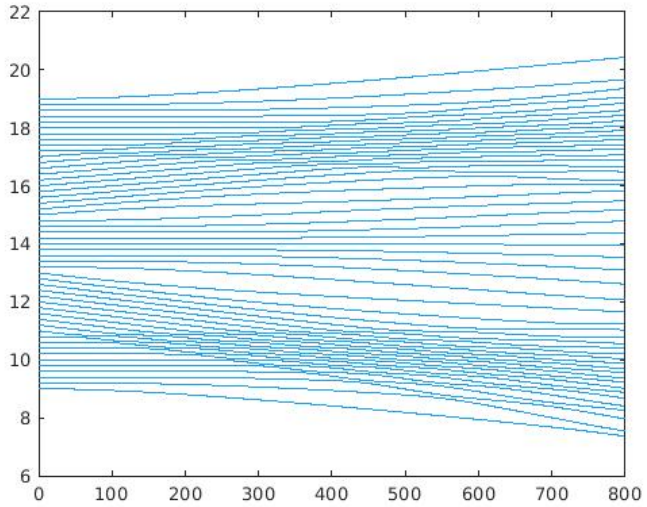

Figure 13.5: The graph shows the trajectories of a one dimensional flow, the x axis being time. The flow starts at uniform density and stationary apart from two equally sized steps. The velocity of the fluid on these two steps is pointing outwards and the step starting at higher y values is travelling faster than the lower one.

If the timestep is not chosen so that $(\mathrm{Id}+\tau u_0)\#\rho_0$ is absolutely continuous then the density will immediately degenerate, to circumvent this a new timestep $\tau_1 < \tau$ can be chosen so that $(\mathrm{Id}+\tau_1 u_0)\#\rho_0$ is absolutely continuous. This adaptive timestep condition can lead to the numerical algorithm incrementing by smaller and smaller timesteps each iteration. A method to circumvent this problem could be attempting to find a $\tau_1 > \tau$ for which the new measure is absolutely continuous. This would imply that between timesteps the system degenerated, however it is possible to run the numerical algorithm in this way. Comparison of these two approaches for a dynamic system in which the physically observed solution is known would make for an interesting direction of further research.

Bibliography

[1] P.-A. Absil. *Optimization Algorithms on Matrix Manifolds*. Ed. by R Mahony and Rodolphe Sepulchre. Description based on publisher supplied metadata and other sources. Princeton: Princeton University Press, 2008. 1240 pp. ISBN: 9781400830244.

[2] Luigi Ambrosio. *Gradient Flows. In Metric Spaces and in the Space of Probability Measures*. Ed. by Nicola Gigli and Giuseppe Savaré. 2nd ed. Lectures in mathematics ETH Zürich. Description based on publisher supplied metadata and other sources. Basel: Birkhäuser Boston, 2008. 1339 pp. ISBN: 9783764387228.

[3] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, July 1999. ISBN: 9780470316962. DOI: 10.1002/9780470316962.

[4] Garrett Birkhoff and Gian-Carlo Rota. *Ordinary differential equations*. Wiley, 1969. ISBN: 0471000329.

[5] Gordon Blower. "A logarithmic Sobolev inequality for the invariant measure of the periodic Korteweg-de Vries equation". In: *Stochastics* 84.4 (Sept. 2011), pp. 533–542. ISSN: 1744-2516. DOI: 10.1080/17442508.2011.597860.

[6] Gordon Blower. "Concentration of the invariant measures for the periodic Zakharov, KdV, NLS and Gross-Piatevskii equations in 1D and 2D". In: *Journal of Mathematical Analysis and Applications* 438.1 (June 2016), pp. 240–266. ISSN: 0022-247X. DOI: 10.1016/j.jmaa.2016.01.080.

[7] Gordon Blower. *Random Matrices: High Dimensional Phenomena*. Cambridge University Press, Oct. 2009. ISBN: 9781139107129. DOI: 10.1017/cbo9781139107129.

[8] Gordon Blower, Azadeh Khaleghi, and Moe Kuchemann-Scales. “Hasimoto frames and the Gibbs measure of the periodic nonlinear Schrödinger equation”. In: *Journal of Mathematical Physics* 65.2 (Feb. 2024). ISSN: 1089-7658. DOI: 10.1063/5.0169792.

[9] J. Bourgain. “Periodic nonlinear Schrödinger equation and invariant measures”. In: *Communications in Mathematical Physics* 166.1 (Dec. 1994), pp. 1–26. ISSN: 1432-0916. DOI: 10.1007/bf02099299.

[10] Jean Bourgain. *Global solutions of nonlinear Schrödinger equations*. AMS ebook collection. Includes bibliographical references and index. - Electronic reproduction; Providence, Rhode Island; American Mathematical Society; 2012. - Description based on print version record. Providence, R.I: American Mathematical Society, 2012. 1182 pp. ISBN: 9781470431921.

[11] Stephen P. Boyd. *Convex optimization*. Ed. by Lieven Vandenberghe. Version 29. first published 2004. Cambridge: Cambridge University Press, 2023. 716 pp. ISBN: 0521833787.

[12] D. L. Burkholder. “Martingale Transforms”. In: *The Annals of Mathematical Statistics* 37.6 (1966), pp. 1494–1504. ISSN: 00034851. URL: http://www.jstor.org/stable/2238766 (visited on 12/30/2024).

[13] Keith Burns and Marian Gidea. *Differential Geometry and Topology: With a View to Dynamical Systems*. Chapman and Hall/CRC, May 2005. ISBN: 9780429124006. DOI: 10.1201/9781420057539.

[14] R. H. Cameron and W. T. Martin. “Transformations of Weiner Integrals Under Translations”. In: *The Annals of Mathematics* 45.2 (Apr. 1944), p. 386. ISSN: 0003-486X. DOI: 10.2307/1969276.

[15] Oscar Castro-Orgaz and Hubert Chanson. “Ritter’s dry-bed and dam-break flows: positive and negative wave dynamics”. In: *Environmental Fluid Mechanics* 17.4 (2017), pp. 665–694. ISSN: 1573-1510. DOI: 10.1007/s10652-017-9512-5. URL: https://doi.org/10.1007/s10652-017-9512-5.

[16] D. Chen G; Wang. *Handbook of mathematical fluid dynamics*. Ed. by Susan Friedlander and Denis Serre. 1st ed. Includes bibliographical references and indexes. Amsterdam ; Elsevier, 2002. 816 pp. ISBN: 9780080532929.

[17] Constantine M Dafermos. “The entropy rate admissibility criterion for solutions of hyperbolic conservation laws”. In: *Journal of Differential Equations* 14.2 (Sept. 1973), pp. 202–212. ISSN: 0022-0396. DOI: 10.1016/0022-0396(73)90043-0.

[18] Persi Diaconis and Peter J. Forrester. “Hurwitz and the origins of random matrix theory in mathematics”. In: *Random Matrices: Theory and Applications* 06.01 (Jan. 2017), p. 1730001. ISSN: 2010-3271. DOI: 10.1142/s2010326317300017.

[19] J. J. Dongarra et al. “A Set of Level 3 Basic Linear Algebra Subprograms”. In: *ACM Trans. Math. Softw.* 16.1 (Mar. 1990), pp. 1–17. ISSN: 0098-3500. DOI: 10.1145/77626.79170. URL: https://doi.org/10.1145/77626.79170.

[20] R.F. Dressler. “Hydraulic resistance effect upon the dam-break functions”. In: *Journal of Research of the National Bureau of Standards* 49.3 (Sept. 1952), p. 217. ISSN: 0091-0635. DOI: 10.6028/jres.049.021.

[21] Richard M. Dudley. *Real analysis and probability*. 2. ed., reprint. Cambridge studies in advanced mathematics 74. Zuerst erschienen 1989 bei Wadsworth. Cambridge [u.a.]: Cambridge Univ. Press, 2008. 555 pp. ISBN: 0521007542.

[22] Nelson Dunford and Jacob T. Schwartz. *Linear Operators. Nelson*. Vol. 1. General theory. New York: Wiley Interscience Publ., 1988. 858 pp. ISBN: 9780471608486.

[23] Nelson Dunford and Jacob T. Schwartz. *Linear Operators. Nelson*. Vol. 2. Spectral theory. New York: Wiley Interscience Publ., 1988. 859192317 pp. ISBN: 0471608475.

[24] Richard Durrett. *Brownian motion and martingales in analysis*. Wadsworth mathematics series. Literaturverz. S. 313 - 324. Belmont, Cal.: Wadsworth Advanced Books [and] Software, 1984. 328 pp. ISBN: 0534030653.

[25] Adam Eisner and Bruce Turkington. “Nonequilibrium statistical behavior of nonlinear Schrödinger equations”. In: *Physica D: Nonlinear Phenomena* 213.1 (2006), pp. 85–97. ISSN: 0167-2789. DOI: https://doi.org/10.1016/j.physd.2005.11.002. URL: https://www.sciencedirect.com/science/article/pii/S016727890500477X.

[26] Lawrence C. Evans. *Partial differential equations*. Second edition. Graduate studies in mathematics 19. Literaturverzeichnis: Seite 689-701. - Index. Providence, Rhode Island: American Mathematical Society, 2022. 712 pp. ISBN: 9780821849743.

[27] Wilfrid Gangbo and Michael Westdickenberg. "Optimal Transport for the System of Isentropic Euler Equations". In: *Communications in Partial Differential Equations* 34.9 (Aug. 2009), pp. 1041–1073. ISSN: 1532-4133. DOI: 10.1080/03605300902892345.

[28] Gisele Ruiz Goldstein. *Semigroups of linear and nonlinear operations and applications. Conference on Semigroups of Operators and Applications*. Literaturangaben. Dordrecht: Kluwer, 1993. 283 pp. ISBN: 0792325605.

[29] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Third. The Johns Hopkins University Press, 1996.

[30] Markus Grasmair. *Basic properties of convex functions*. Institutt for matematiske fag. URL: https://wiki.math.ntnu.no/_media/tma4180/2016v/note2.pdf.

[31] Arthur Gretton. "Consistent Nonparametric Tests of Independence". In: *Journal of Machine Learning Research* 11.46 (2010), pp. 1391–1423. URL: http://jmlr.org/papers/v11/gretton10a.html.

[32] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration*. Second. Vol. 31. Springer Series in Computational Mathematics. Structure-preserving algorithms for ordinary differential equations. Springer-Verlag, Berlin, 2006, pp. xviii+644. ISBN: 3-540-30663-3; 978-3-540-30663-4.

[33] Brian C Hall. *Lie groups, Lie algebras, and representations an elementary introduction*. eng. Second edition. Graduate texts in mathematics ; 222. Cham ; New York: Springer, 2015. ISBN: 9783319134666.

[34] Godfrey H. Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. 1. paperback ed., repr., transferred to digital printing. Cambridge mathematical library. Cambridge [u.a.]: Cambridge Univ. Press, 2001. 324 pp. ISBN: 0521052068.

[35] Hidenori Hasimoto. "A soliton on a vortex filament". In: *Journal of Fluid Mechanics* 51.3 (Feb. 1972), pp. 477–485. ISSN: 1469-7645. DOI: 10.1017/s0022112072002307.

[36] Nicholas J. Higham. "The Scaling and Squaring Method for the Matrix Exponential Revisited". In: *SIAM Journal On Matrix Analysis and Applications* 26.4 (2005), pp. 1179–1193. URL: http://eprints.maths.manchester.ac.uk/634/.

[37] Hille. *Methods in classical and functional analysis*. Addison-Wesley, 1972.

[38] Einar Hille. *Ordinary differential equations in the complex domain*. Pure and applied mathematics. New York, NY [u.a.]: Wiley, 1976. 484 pp. ISBN: 0471399647.

[39] The MathWorks Inc. *MATLAB*. Version 9.13.0 (R2022b). 2022. URL: https://uk.mathworks.com/products/matlab.html.

[40] Wittawat Jitkrittum, Zoltan Szabo, and Arthur Gretton. "An Adaptive Test of Independence with Analytic Kernel Embeddings". In: (Oct. 15, 2016). DOI: 10.48550/ARXIV.1610.04782. arXiv: 1610.04782 [stat.ML].

[41] Olav Kallenberg. *Foundations of Modern Probability*. Springer International Publishing, 2021. ISBN: 9783030618711. DOI: 10.1007/978-3-030-61871-1.

[42] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer New York, 1998. ISBN: 9781461209492. DOI: 10.1007/978-1-4612-0949-2.

[43] Peter E. Kloeden. *Numerical solution of stochastic differential equations*. Ed. by Eckhard Platen. Corrected third printing. Applications of mathematics 23. Berlin: Springer, 1999. 636 pp. ISBN: 9783540540625.

[44] Moe Kuchemann-Scales. *NLS_Stochastic Numerical solver*. GitHub Repository. 2022. URL: https://github.com/MoeK-S/NLS_stochastic.

[45] Pijush K. Kundu. *Fluid mechanics*. Ed. by Ira M. Cohen, David R. Dowling, and Gretar Tryggvason. Sixth Edition. Amsterdam: Elsevier, Academic Press, 2016. 921 pp. ISBN: 9780124059351.

[46] Hui-Hsiung Kuo. "Gaussian measures in Banach spaces". In: (1975), pp. 1–109. ISSN: 1617-9692. DOI: https://doi.org/10.1007/BFb0082008.

[47] Peter Lancaster. *The theory of matrices*. Vol. 23. 108. JSTOR, Oct. 1969, p. 886. DOI: 10.2307/2004986.

[48] Joel L. Lebowitz, Harvey A. Rose, and Eugene R. Speer. "Statistical mechanics of the nonlinear Schrödinger equation". In: *Journal of Statistical Physics* 50.3-4 (Feb. 1988), pp. 657–687. ISSN: 1572-9613. DOI: 10.1007/bf01026495.

[49] Wilhelm Magnus. "On the exponential solution of differential equations for a linear operator". In: *Communications on Pure and Applied Mathematics* 7.4 (Nov. 1954), pp. 649–673. ISSN: 1097-0312. DOI: 10.1002/cpa.3160070404.

[50] Goran Marjanovic and Victor Solo. "Numerical Methods for Stochastic Differential Equations in Matrix Lie Groups Made Simple". In: *IEEE Transactions on Automatic Control* 63.12 (Dec. 2018), pp. 4035–4050. ISSN: 2334-3303. DOI: 10.1109/tac.2018.2798703.

[51] H. P. McKean. "Statistical mechanics of nonlinear wave equations (4): Cubic Schrödinger". In: *Communications in Mathematical Physics* 168.3 (Apr. 1995), pp. 479–491. ISSN: 1432-0916. DOI: 10.1007/bf02101840.

[52] Henry P. McKean. *Stochastic Integrals*. Ed. by Z. W. Birnbaum and E. Lukacs. Description based upon print version of record. Burlington: Elsevier Science, 2014. 157 pp. ISBN: 9781483230542.

[53] Henry P. McKean and Victor Moll. *Elliptic curves. Function theory, geometry, arithmetic*. 1. paperback ed. (with corr.) Hier auch später erschienene, unveränderte Nachdrucke. Cambridge [u.a.]: Cambridge Univ. Press, 1999. 280 pp. ISBN: 0521658179.

[54] Rupert G. Jr Miller. *Simultaneous Statistical Inference*. 2nd ed. Springer Series in Statistics Ser. Description based on publisher supplied metadata and other sources. New York, NY: Springer New York, 1991. 1311 pp. ISBN: 9781461381228. DOI: 10.1007/978-1-4613-8122-8.

[55] Hans Munthe-Kaas. "Runge-Kutta methods on Lie groups". In: *BIT Numerical Mathematics* (Mar. 1998). DOI: 10.1007/BF02510919.

[56] Met Office, ed. *Surface Pressure Charts*. Apr. 30, 2025. URL: https://weather.metoffice.gov.uk/maps-and-charts/surface-pressure.

[57] F. Otto and C. Villani. "Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality". In: *Journal of Functional Analysis* 173.2 (June 2000), pp. 361–400. ISSN: 0022-1236. DOI: 10.1006/jfan.1999.3557.

[58] M. J. Piggott and V. Solo. "Geometric Euler–Maruyama Schemes for Stochastic Differential Equations in SO(n) and SE(n)". In: *SIAM Journal on Numerical Analysis* 54.4 (Jan. 2016), pp. 2490–2516. ISSN: 1095-7170. DOI: 10.1137/15m1019726.

[59] Marc J. Piggott and Victor Solo. "Stochastic numerical analysis for Brownian motion on SO(3)". In: *53rd IEEE Conference on Decision and Control*. IEEE, Dec. 2014, pp. 3420–3425. DOI: 10.1109/cdc.2014.7039919.

[60] Andrew Pressley. *Elementary Differential Geometry*. Springer London, 2010. ISBN: 9781848828919. DOI: 10.1007/978-1-84882-891-9.

[61] Andrew Pressley and Graeme Segal. *Loop groups*. Repr. (with corr.) Oxford science publications. Oxford [u.a.]: Clarendon Press, 2003. 318 pp. ISBN: 019853535X.

[62] E. B. Saff and R. S. Varga. "On the zeros and poles of Padé approximants to exp(z) III". In: *Numerische Mathematik* 30.3 (Sept. 1978), pp. 241–266. ISSN: 0945-3245. DOI: 10.1007/bf01411842.

[63] Laurent Schwartz. *Radon measures on arbitrary topological spaces and cylindrical measures*. @Tata Institute of Fundamental Research studies in mathematics 6. London: Oxford Univ. Press, 1973. 393 pp. ISBN: 0195605160.

[64] Igor R. Shafarevich. *Basic Algebraic Geometry 1: Varieties in Projective Space*. Springer Berlin Heidelberg, 2013. ISBN: 9783642379567. DOI: 10.1007/978-3-642-37956-7.

[65] Ichiro Shigekawa. "Transformations of the Brownian Motion on the Lie Group". In: *Stochastic Analysis, Proceedings of the Taniguchi International Symposium on Stochastic Analysis*. Elsevier, 1984, pp. 409–422. ISBN: 9780444875884. DOI: 10.1016/s0924-6509(08)70402-1.

[66] George Finlay Simmons. *Differential equations with applications and historical notes*. Third edition, first edition in paperback. A Chapman and Hall book. Includes index. Boca Raton: CRC Press, Taylor and Francis Group, 2022. 740 pp. ISBN: 1032477148.

[67] George Finlay Simmons. *Introduction to topology and modern analysis*. Includes bibliographical references (p. 355-357) and index. Malabar, Fla: Krieger Pub. Co, 2003. 372 pp. ISBN: 1575242389.

[68] John Stillwell. *Naive Lie Theory*. Springer New York, 2008. ISBN: 9780387782157. DOI: 10.1007/978-0-387-78214-0.

[69] Karl-Theodor Sturm. "On the geometry of metric measure spaces". In: *Acta Mathematica* 196.1 (2006), pp. 65–131. ISSN: 0001-5962. DOI: 10.1007/s11511-006-0002-8.

[70] Amy Tikkanen. *Cavitation*. Ed. by Brittanica Encyclopeadia. URL: https://www.britannica.com/science/cavitation.

[71] Edward C. Titchmarsh. *The theory of functions*. 2nd ed. Oxford science publications. Literaturverz. S. [445] - 452. Oxford [u.a.]: Oxford Univ. Press, 2002. 454 pp. ISBN: 0198533497.

[72] G. M. Tuynman. "The Derivation of the Exponential Map of Matrices". In: *The American Mathematical Monthly* 102.9 (1995), pp. 818–820. ISSN: 00029890, 19300972. URL: http://www.jstor.org/stable/2974511 (visited on 10/27/2022).

[73] Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. ISBN: 9783540710509. DOI: https://doi.org/10.1007/978-3-540-71050-9.

[74] Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, Mar. 2003. ISBN: 9781470418045. DOI: 10.1090/gsm/058.

[75] Michael Westdickenberg and Jon Wilkening. "Variational particle schemes for the porous medium equation and for the system of isentropic Euler equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 44.1 (Dec. 2009), pp. 133–166. ISSN: 1290-3841. DOI: 10.1051/m2an/2009043.