Analysis and Implementation of Numerical Methods for Simulating Dilute Polymeric Fluids

David Knezevic

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2008

Abstract

Analysis and Implementation of Numerical Methods for Simulating Dilute Polymeric Fluids

David Knezevic — Doctor of Philosophy
Balliol College — Michaelmas Term 2008

In this thesis we develop, analyse and implement a number of numerical methods for simulating dilute polymeric fluids. We use a well-known model in which the polymeric fluid is represented by a suspension of dumbbells in a Newtonian solvent. This model is governed by a coupled Navier–Stokes–Fokker–Planck system of partial differential equations, in which the Fokker–Planck equation is posed on a high-dimensional domain.

We first thoroughly analyse a Galerkin spectral method for the Fokker–Planck equation in *configuration space*, before combining this method with a finite element scheme in *physical space* to obtain an alternating-direction method for the high-dimensional Fokker–Planck equation. Alternating-direction methods have been considered previously in the literature for this problem (*e.g.* see [23, 24, 60]), but this approach has not been subject to rigorous numerical analysis before. We develop many theoretical results for our numerical algorithms, and we focus particularly on establishing stability and convergence estimates. The numerical methods we develop are fully-practical, and we present a range of numerical results demonstrating their accuracy and efficiency.

We also introduce a coupled numerical algorithm for the Navier–Stokes–Fokker–Planck system, which we use to simulate polymeric fluid flow problems of physical interest. The numerical method for the high-dimensional Fokker–Planck equation is the most computationally intensive part of this coupled algorithm, but it is well suited to implementation on a parallel computer, and we exploit this fact to make large-scale computations feasible.

Acknowledgements

First of all, I would like to thank my D.Phil. supervisor, Professor Endre Süli. It has been a joy to work with Professor Süli these past three years. I have learnt so much from his deep analytical insights and the incredible breadth of his mathematical knowledge. Professor Süli's infectious passion for numerical analysis will stay with me for a long time to come.

I would also like to thank all of the members of the Numerical Analysis Group at Oxford University. It has been wonderful to have been a part of the active and social environment in the group, and I have relished the opportunity to learn from everyone, both faculty and students, whether it be in our frequent seminars, or during morning coffee breaks. I'd especially like to thank Martin Stoll; it has been great to share coffees and lunchbreaks with him and to talk about mathematics and about everything else.

I am extremely grateful to the folks at Professor Carey's CFDLab at the University of Texas at Austin. My passion for finite element methods was first ignited when I spent some time in the CFDLab as an undergraduate on student exchange, and I have continued to learn from Professor Carey and the CFDLab group members ever since. I'd especially like to thank John Peterson, who has helped me many times with practical problems related to finite element computing, and also to all the developers of the `libMesh` software package, which has been so influential in my research.

Of course, I also owe a special "thank you" to my family. My Mum and Dad and my brother and sister have always offered complete encouragement and support in my studies, and in everything else I have been involved in. Finally, I'd especially like to thank my lovely fiancée, Olivia. The energy she puts into all of her activities has been a constant inspiration to me, and being with her has made my life in Oxford truly wonderful.

Contents

List of Figures

List of Tables

List of Symbols

d	space dimension
$\mathbb{R}$	set of all real numbers
$\mathrm{L}^p(\Omega)$	Lebesgue space of order p on $\Omega \subset \mathbb{R}^d$
$\mathrm{W}^{k,p}(\Omega)$	Sobolev space of order $k,\ p$ on $\Omega \subset \mathbb{R}^d$
$\mathrm{H}^k(\Omega)$	Sobolev space of order k (with $p = 2$) on $\Omega \subset \mathbb{R}^d$
$\underset{\sim}{x}$, $\underset{\sim}{q}$	physical and configuration space variables
t, T	time variable, $t \in [0, T] \subset \mathbb{R}$
Ω, D	physical and configuration space domains
$\underset{\sim}{n}$	outward unit normal to Ω
$\underset{\sim}{u}$, p	velocity and pressure for Newtonian solvent
$\underset{\approx}{\kappa}$	velocity gradient tensor, $\nabla_x \underset{\sim}{u}$
$\underset{\approx}{\tau}$	polymeric extra-stress tensor
b	non-dimensional dumbbell extension parameter
$\underset{\sim}{F}$, U	force law and potential for dumbbell spring
M	Maxwellian for the force law $\underset{\sim}{F}$
ψ	solution of Fokker–Planck equation
$\hat{\psi}$	transformed solution, *i.e.* $\psi / M^{2s/b},\ s \in (1/2, \infty)$
$\hat{\psi}_N$, $\hat{\psi}_{h,N}$	numerical solution on D and $\Omega \times D$, respectively
μ_s, μ_p, μ	solvent viscosity, polymeric viscosity, total viscosity ($\mu := \mu_s + \mu_p$)
ρ	solvent density
ν_s, ν_p, ν	kinematic viscosities, *i.e.* $\nu := \mu / \rho$
γ	ratio of solvent to total viscosity, *i.e.* $\gamma := \nu_s / \nu$
Re	Reynolds number
Wi	Weissenberg number, ratio of microscopic to macroscopic time-scales
$\varrho(\underset{\sim}{x}, t)$	$\int_D \psi(\underset{\sim}{x}, \underset{\sim}{q}, t)\, \mathrm{d}\underset{\sim}{q}$
δ	extension rate in an extensional flow
$\mathcal{P}_N(D)$	finite dimensional spectral space on D
$\mathcal{T}_h$	finite element triangulation of $\overline{\Omega}$
V_h	$\mathrm{H}^1(\Omega)$-conforming finite element space with respect to $\mathcal{T}_h$
Δt	time-step size for temporal discretisation, $t^n = n \Delta t$
N_T	total number of discrete time-steps, $N_T = T / \Delta t$
δ_{ij}	Kronecker delta

Chapter 1

Introduction

The study of the dynamics of polymeric fluids has been an area of active research since the 1950's and has undergone significant evolution since that time. In the early work in this field, analytical techniques were developed with the goal of deriving exact solutions for idealised flow problems. With the increasing availability of computational power in subsequent years, it was natural for researchers to apply numerical methods to more complicated flow problems for polymeric fluids (and non-Newtonian fluids in general) than were tractable with analytical methods. This line of research, known as *computational rheology*, took root in the 1970's and it remains an exciting and challenging area of scientific computing today.

In this thesis we investigate a particular problem from computational rheology: the simulation of dilute polymeric fluids using deterministic multiscale numerical methods. We focus our attention on the rigorous analysis of the numerical methods developed here and we also present a wide array of computational results, which demonstrate the effectiveness of the methods in practice.

The essence of the subject of modelling dilute polymeric fluids is encapsulated in the coupled Navier–Stokes–Fokker–Planck system (this is discussed in detail in Section 1.3). This system of equations is often referred to as the "micro-macro" model to emphasise that it is fundamentally multiscale in nature. It is worth highlighting at the outset that there is an extensive literature on numerical methods for simulating polymeric fluids, but most of the previous work uses either a fully macroscopic approach in order to circumvent the multiscale nature of the Navier–Stokes–Fokker–Planck system (see the text [69] for an overview of this field) or a stochastic approach in which the micro-macro system is treated using Monte-Carlo-type methods (*cf.* [68]). The direction pursued in this thesis is rather different; our goal is to solve the micro-macro system using deterministic methods (*e.g.* finite element or spectral methods). This will subsequently be referred to as the *deterministic multiscale* approach. The

various advantages and disadvantages of fully macroscopic, stochastic and deterministic multiscale methods will be discussed in detail later, but it should be noted at the outset that the deterministic multiscale method has received far less attention in the literature than the other approaches, probably because this approach can be highly computationally intensive. The central goal of this thesis, therefore, is to develop multiscale numerical methods for the micro-macro model of dilute polymeric fluids and to address some of the questions related to numerical analysis of such methods, which, up to now, have not been considered in the literature.

In this introductory chapter, we discuss background material on the mathematical modelling of polymer fluids. Newtonian fluids are briefly considered in Section 1.1, and then in Section 1.2 some "coarse-grained" mechanical models for polymer molecules are introduced. Next, in Section 1.3, we derive the Fokker–Planck equation and define the coupled Navier–Stokes–Fokker–Planck system. Section 1.4 contains a literature review of the many and varied numerical methods that have been used for simulating polymeric fluids (these methods fall into the three categories mentioned in the previous paragraph), and the chapter concludes with an overview of the outlook and goals of this thesis.

1.1 Overview of Newtonian fluid dynamics

The success of classical fluid dynamics in accurately describing the properties of a wide range of fluids (typically with low molecular weight, *e.g.* water) using macroscopic continuum models is well established. We begin with a very brief review of some principles of classical fluid dynamics (for a full discussion see [11]) as this will be useful for elucidating important ideas in the theory of polymeric fluids.

In the case of Newtonian fluids it has been experimentally established that in a shear flow, *i.e.* $u = u(y), v = 0$ where u and v are the components of a two-dimensional velocity field $\underset{\sim}{u} = (u, v)$, the fluid stress can be related to shear rate by "Newton's law of viscosity":

$$\sigma_{yx} = \mu \frac{du}{dy}, \tag{1.1}$$

where σ_{yx} denotes the force per unit area acting in the x-direction, on a surface normal to the y-direction. That is, stress is proportional to shear rate and the viscosity, μ, is the constant of proportionality. This relationship can be generalised to a tensor equation for the stress tensor, $\underset{\approx}{\sigma}$, and the strain tensor as follows:

$$\underset{\approx}{\sigma} = -p\underset{\approx}{I} + \mu \left(\underset{\sim}{\nabla} \underset{\sim}{u} + (\underset{\sim}{\nabla} \underset{\sim}{u})^T \right). \tag{1.2}$$

This equation provides a relationship between the stress and strain of a fluid (in this case, a simple linear equation) and is known as a *constitutive equation*.

Combining the Newtonian constitutive equation (1.2) with the equations of conservation of mass:

$$\underset{\sim}{\nabla} \cdot \underset{\sim}{u} = 0, \tag{1.3}$$

and momentum:

$$\rho \left(\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla} \underset{\sim}{u} \right) = \underset{\sim}{\nabla} \cdot \underset{\approx}{\sigma}, \tag{1.4}$$

where ρ is the fluid density (assumed to be constant), gives rise to the Navier–Stokes equations for an incompressible, viscous, isothermal fluid:

$$\begin{aligned} \frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla} \underset{\sim}{u} - \nu \Delta \underset{\sim}{u} + \underset{\sim}{\nabla} p &= 0, & (1.5) \\ \underset{\sim}{\nabla} \cdot \underset{\sim}{u} &= 0, & (1.6) \end{aligned}$$

where the momentum equation has been divided through by ρ, the pressure in (1.5) has implicitly been rescaled by ρ and $\nu := \mu/\rho$ is the kinematic viscosity. These equations (which involve only macroscopic quantities) form the cornerstone of classical fluid dynamics.

The situation with polymeric fluids, however, is quite different. In general the contributions to the stress tensor $\underset{\approx}{\sigma}$ from microscopic polymer molecules cannot be averaged out into purely macroscopic quantities and therefore in order to faithfully simulate a polymeric fluid, the microscopic and macroscopic length scales must be coupled together. This coupling is achieved by the Navier–Stokes–Fokker–Planck system alluded to above.

In the next section, mechanical models (*i.e.* systems containing masses, rigid rods and/or springs) for microscopic polymer molecules are considered. From the perspective of polymer fluid dynamics, the purpose of these models is to capture the most important characteristics of polymer molecules in systems with many fewer degrees of freedom and in order to yield mathematical models for polymeric fluids that are analytically and computationally tractable.

1.2 Modelling polymeric fluids

Polymer molecules consist of long chains of repeated basic structural units, or *monomers*. Polymers of interest typically contain on the order of 10^3 to 10^6 monomers and the presence of these long chain molecules in a fluid can dramatically affect the fluid's macroscopic properties. In particular, polymer molecules introduce elastic properties

and, as a result, polymeric fluids are often described as *viscoelastic*. Viscoelasticity gives rise to an exotic range of phenomena, such as shear-thinning, rod-climbing, the "tubeless siphon", and elastic recoil [17].

Most approaches to the mathematical modelling of polymeric fluids are based on kinetic theory, in which the behaviour of the microscopic polymer molecules is characterised in a statistical sense. The starting point in deriving kinetic–theory–based equations is to propose a simple mechanical model that represents an individual polymer molecule. A mechanical model that would faithfully capture the microscopic properties of an actual polymer would be extremely complicated, with a very high number of degrees of freedom, and would be prohibitively difficult to deal with and as a result, a range of simplifications and idealisations have been proposed.

The following "coarse-grained" models for polymer molecules are discussed below: the freely rotating chain model; the bead-rod chain model; the bead-spring chain model; and the dumbbell model (see Chapter 10 of Bird *et. al.* [18] for more details on each of these). This hierarchy of models is depicted in Figure 1.1(a).

The Freely Rotating Chain Model

It was observed by Flory [33] that bond angles between monomers in a polymer chain are restricted to quite narrow ranges about their average values (up to $\sim 3\%$ deviation). This motivated the freely rotating chain model which represents each monomer unit as a bead, where adjacent beads are joined by a rigid, massless rod and where rods are set at a fixed angle (the average bond angle) but are free to rotate. This model has been used in a number of kinetic theory studies by Kirkwood [48]. For the purposes of multiscale computations, though, this model is far too complex. It requires one degree of freedom for each monomer, so that the number of degrees of freedom in a single chain would be on the order of 10^3 to 10^6.

The Bead-Rod Chain Model

The bead-rod chain model is significantly simpler. It lumps a group of monomers into a single bead and adjacent beads are connected by a massless rod. The constraint on bond angle is dropped so that this model is referred to as "freely jointed". The number of degrees of freedom for this model is typically around 100. The bead-rod chain was first analysed in a seminal paper by Kramers in 1944 [51], and the model is often referred to as a *Kramers chain*. While clearly a considerable simplification from the freely rotating chain, this model still reflects a number of the important characteristics of a polymer molecule – in particular the bead-rod chain has a large number of internal

degrees of freedom, it can be oriented and deformed by a flow and it has a constant contour length.

The Bead-Spring Chain Model

The bead-spring chain is a yet coarser model; a polymer is modelled by a chain of typically around 10 beads joined by springs. The model is completed by specifying a force law for the springs (see below). This model has been the basis of a number of kinetic-theory-based investigations of polymer fluids, *e.g.* the seminal papers of Rouse and Zimm [71, 86].

The Dumbbell Model

The dumbbell model is the simplest in the hierarchy of coarse-grained mechanical models for polymers; it consists of only two masses, which are connected by a spring (or sometimes a rigid rod, although we only consider the spring case in this thesis). A dumbell is fully specified by the position of its centre of mass, $\underset{\sim}{x}$, and its configuration (or end-to-end) vector, $\underset{\sim}{q}$ (see Figure 1.1(b)). Despite the simplicity of the dumbbell model, it is still very useful for simulating polymeric fluids in many flow regimes because dumbbells can be stretched and oriented by a flow, and these two actions determine the main contributions from polymer molecules to the macroscopic properties of a viscoelastic fluid.

Spring Force Laws

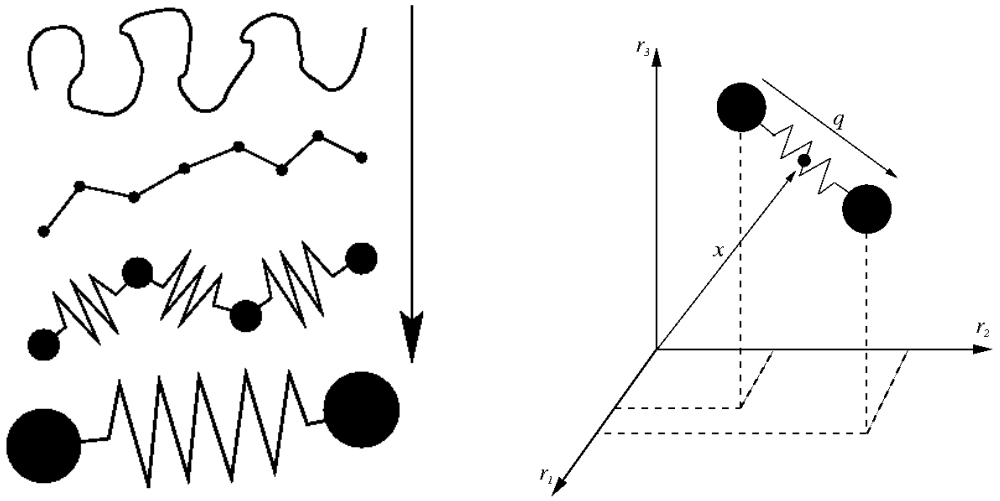
As indicated above, a force law, $\underset{\sim}{F}$, must also be defined for the coarse-grained models that contain one or more springs. In general, the elastic force is assumed to be defined by a (sufficiently smooth) potential $U : \mathbb{R}_{\geq 0} \to \mathbb{R}$ via

$$\underset{\sim}{F}(\underset{\sim}{q}) = H\, U'(\tfrac{1}{2}|\underset{\sim}{q}|^2)\underset{\sim}{q}, \tag{1.7}$$

where $\underset{\sim}{q}$ is the configuration vector (as illustrated in Figure 1.1(b)) of a given spring and $H \in \mathbb{R}_{>0}$ is the spring constant. The simplest force law is that of a Hookean spring:

$$U(s) = s \qquad \text{and} \qquad \underset{\sim}{F}(\underset{\sim}{q}) = H\underset{\sim}{q}. \tag{1.8}$$

Many interesting analytical results have been derived for dilute solutions of Hookean dumbbells; indeed the simple linear relationship in (1.8) makes this model attractive from the mathematical point of view. For example, it is well known that the Oldroyd-B macroscopic model for dilute polymeric fluids (originally derived from continuum mechanics considerations [67]) is equivalent to the Hookean dumbbell micro-macro



(a) (b)

Figure 1.1: (a) Diagram of the hierarchy of mechanical models for polymer molecules, descending from a polymer molecule with on the order of 10^3 to 10^6 monomers to the dumbbell model, containing only two masses connected by a spring. (b) A more detailed depiction of the dumbbell model. The state of a dumbbell is defined by the position of its centre of mass, $\underset{\sim}{x}$, and its configuration (or end-to-end) vector, $\underset{\sim}{q}$. The dumbbell shown in this schematic can move in $\mathbb{R}^3$, and therefore has six degrees-of-freedom.

model (*e.g.* see [6]). However, due to the physically unrealistic ability of Hookean springs to be infinitely stretched these models can break down in certain cases, such as strong extensional flows. A remedy is to use the Finitely Extensible Non-linear Elastic (FENE) force law, suggested by Warner [82], for which we have,

$$U(\tfrac{1}{2}|\underset{\sim}{q}|^2) = -\frac{l_{\max}^2}{2}\ln\left(1-\frac{|\underset{\sim}{q}|^2}{l_{\max}^2}\right) \qquad \text{and} \qquad \underset{\sim}{F}(\underset{\sim}{q}) = \frac{H\underset{\sim}{q}}{1-|\underset{\sim}{q}|^2/l_{\max}^2}. \tag{1.9}$$

As the name suggests, FENE springs can only be stretched a finite amount because the spring potential is unbounded as $|\underset{\sim}{q}| \to l_{\max}$. Unlike with Hookean springs, there is no equivalent macroscopic formulation for suspensions of FENE dumbbells; the FENE dumbbell model requires a truly multiscale approach. Note also that for $|\underset{\sim}{q}| < l_{\max}$ fixed, the FENE force converges to the Hookean spring force as $l_{\max} \to \infty$.

In this thesis, the focus is on developing deterministic multiscale methods for simulating the flow of a suspension of FENE dumbbells[1] in a Newtonian solvent. This is an imposing challenge in itself because (as discussed in Section 1.3) for a d-dimensional flow, the Fokker–Planck equation is posed in $2d$ spatial dimensions, where we consider

[1]Although, in Chapter 2, we consider a more general class of spring potentials that include the FENE potential as a special case.

$d = 2$ or 3. Solving this high-dimensional equation is a large-scale computational problem, which requires highly specialised numerical methods. Replacing dumbbells with bead-spring chains would clearly make the problem far more challenging still. The development of methods to treat the bead-spring chain case efficiently using deterministic algorithms (as opposed to Monte Carlo approaches) has received attention in the literature recently (see Section 1.4). Extending the work in this thesis to the bead-spring case is a goal of future research.

1.3 The micro-macro model

With the background material developed in the previous two sections it is now possible to derive the Navier–Stokes–Fokker–Planck model for dilute polymeric fluids. As indicated above, we consider a dilute solution of polymer chains suspended in a Newtonian solvent, and we assume that individual polymer chains do not interact with one another, but can be convected, stretched and oriented by the macroscopic velocity field, and are also subject to thermal agitation due to the motion of the solvent molecules.

Suppose the fluid is confined to a physical domain Ω, assumed to be a bounded open set in $\mathbb{R}^d$, $d = 2$ or 3, and that appropriate boundary conditions are imposed on $\partial\Omega$. The conservation equations for polymeric fluids are the same as for the Newtonian case, but the presence of polymer molecules contributes a *polymeric extra-stress*, represented by the tensor $\underset{\approx}{\tau}$. That is, the total stress tensor $\underset{\approx}{\sigma}$ is given by

$$\underset{\approx}{\sigma} = -p\underset{\approx}{I} + \mu_s(\underset{\sim}{\nabla}\underset{\sim}{u} + (\underset{\sim}{\nabla}\underset{\sim}{u})^T) + \underset{\approx}{\tau}, \tag{1.10}$$

where in this case the viscosity is labelled with a subscript s to indicate that it comes from the solvent. Combining (1.10) with the conservation of mass and momentum equations yields a modified form of the Navier–Stokes equations in which the divergence of $\underset{\approx}{\tau}$ arises as a source term. Thus, the model problem takes the following form:

Find $\underset{\sim}{u} : (\underset{\sim}{x}, t) \in \Omega \times \mathbb{R} \to \underset{\sim}{u}(\underset{\sim}{x}, t) \in \mathbb{R}^d$ and $p : (\underset{\sim}{x}, t) \in \Omega \times \mathbb{R} \to p(\underset{\sim}{x}, t) \in \mathbb{R}$ such that

$$\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla}\underset{\sim}{u} - \nu_s \Delta \underset{\sim}{u} + \underset{\sim}{\nabla} p = \frac{1}{\rho}\underset{\sim}{\nabla} \cdot \underset{\approx}{\tau} \qquad \text{in } \Omega \times (0, T], \tag{1.11}$$

$$\underset{\sim}{\nabla} \cdot \underset{\sim}{u} = 0 \qquad \text{in } \Omega \times (0, T], \tag{1.12}$$

$$\underset{\sim}{u}(\underset{\sim}{x}, 0) = \underset{\sim}{u}_0(\underset{\sim}{x}) \qquad \forall \underset{\sim}{x} \in \Omega, \tag{1.13}$$

where ν_s is the kinematic solvent viscosity, $\nu_s := \mu_s/\rho$. The system is completed by specifying appropriate boundary conditions on $\partial\Omega$.

The system (1.11)–(1.13) models the macroscopic flow of a polymeric fluid, and the contributions of microscopic polymer molecules enter through the extra-stress tensor, $\underset{\approx}{\tau}$. In the case that the polymers are represented by coarse-grained models (*e.g.* dumbbells), it turns out that $\underset{\approx}{\tau}$ can be computed in terms of a statistical averaging of the probability density function describing the distribution of configurations of polymer molecules within the fluid.[2] The probability density function for dumbbell configurations will henceforth be denoted ψ, and the idea of the deterministic multiscale method is to compute ψ directly by solving a partial differential equation (the high-dimensional Fokker–Planck equation alluded to above) so that $\underset{\approx}{\tau}$ can be computed and fed into the macroscopic system (1.11)–(1.13).

1.3.1 Derivation of the Fokker–Planck equation

In this section the Fokker–Planck equation for polymeric fluids that governs ψ is derived from first principles. For the purposes of the derivation, it suffices to consider the general spring force law (1.7). Similar derivations can be found in Bird *et. al.* [18], the Ph.D. thesis of Lozinski [59] or the paper by Barrett & Süli [10].

First of all, consider an isolated dumbell immersed in a Newtonian solvent with fluid velocity given by $\underset{\sim}{u}(\underset{\sim}{x},t)$. Denote by $\underset{\sim}{r}_1(t), \underset{\sim}{r}_2(t) \in \Omega \subset \mathbb{R}^d$ the position vectors of the two masses of the dumbbell at time t, where Ω is referred to as *physical space*. For the purpose of this derivation we assume that $\Omega = \mathbb{R}^d$; this allows us to avoid complications associated with the behaviour of dumbbells at the domain boundary. From Section 1.3.2 onwards, we shall assume Ω is a bounded subset of $\mathbb{R}^d$.

As in Figure 1.1(b), the centre of mass, $\underset{\sim}{x}(t)$, and configuration vector, $\underset{\sim}{q}(t)$, are defined as:

$$\underset{\sim}{x}(t) = \left(\underset{\sim}{r}_1(t) + \underset{\sim}{r}_2(t)\right)/2 \qquad \text{and} \qquad \underset{\sim}{q}(t) = \underset{\sim}{r}_2(t) - \underset{\sim}{r}_1(t). \tag{1.14}$$

Assuming Ω is convex then $\underset{\sim}{x}(t) \in \Omega$. Also, let *configuration space* be the set of all admissible configuration vectors (which we assume to be a time-invariant domain), *i.e.*

$$D = \{\underset{\sim}{q} \in \mathbb{R}^d : \underset{\sim}{q} = \underset{\sim}{r}_2 - \underset{\sim}{r}_1, \text{ for all admissible } \underset{\sim}{r}_1, \underset{\sim}{r}_2 \in \Omega\}.$$

For example, for Hookean dumbbells, configuration space is all of $\mathbb{R}^d$, whereas for FENE dumbbells, we have $D = B(0, l_{\max})$, where $B(0,s) \subset \mathbb{R}^d$ is the ball centered at the origin with radius s. It is more natural to treat the Fokker–Planck equation

[2] The precise equation for computing $\underset{\approx}{\tau}$ is known as *Kramers expression*, and it is discussed below in Section 1.3.3.

in $(\underset{\sim}{x}, \underset{\sim}{q})$-coordinates than in $(\underset{\sim}{r}_1, \underset{\sim}{r}_2)$-coordinates because with the FENE model for example, for a given $\underset{\sim}{r}_1$, we have $\underset{\sim}{r}_2 \in B(\underset{\sim}{r}_1, l_{\max})$, *i.e.* in contrast to the vectors $(\underset{\sim}{x}, \underset{\sim}{q}) \in \Omega \times D$, the domains of $\underset{\sim}{r}_1$ and $\underset{\sim}{r}_2$ cannot be decoupled in this case.

Considering an isolated dumbbell, Newton's Second Law can be applied to the i^{th} bead such that $\underset{\sim}{F}_i^{\text{total}} = m_i \underset{\sim}{a}_i$, where $\underset{\sim}{a}_i$ is the acceleration of bead $i = 1, 2$ and $\underset{\sim}{F}_i^{\text{total}}$ is the sum of the following components:

- $\underset{\sim}{F}_i^{\text{drag}}$: Drag force due to bead i moving through the viscous solvent.
- $\underset{\sim}{B}_i$: Brownian force due to random collisions of solvent molecules with bead i.
- $\underset{\sim}{F}_i$: The spring force on bead i, *e.g.* (1.9).

Hence, we have the following force balance equations for beads 1 and 2:

$$\begin{aligned} m_1 \underset{\sim}{a}_1(t) &= \underset{\sim}{F}_1^{\text{drag}}(t) + \underset{\sim}{B}_1(t) + \underset{\sim}{F}(\underset{\sim}{r}_2(t) - \underset{\sim}{r}_1(t)), \\ m_2 \underset{\sim}{a}_2(t) &= \underset{\sim}{F}_2^{\text{drag}}(t) + \underset{\sim}{B}_2(t) + \underset{\sim}{F}(\underset{\sim}{r}_1(t) - \underset{\sim}{r}_2(t)). \end{aligned}$$

We model the hydrodynamic drag force, $\underset{\sim}{F}^{\text{drag}}$, using Stokes' law for the viscous drag on a sphere at low Reynolds number [1], *i.e.*

$$\underset{\sim}{F}_i^{\text{drag}} = \zeta \left(\underset{\sim}{u}(\underset{\sim}{r}_i(t), t) - \frac{\mathrm{d}\underset{\sim}{r}_i}{\mathrm{d}t}(t) \right),$$

where the term inside the brackets is the velocity of bead i relative to the velocity of the solvent, and ζ is the friction coefficient.

Following Schieber & Öttinger [72] we consider the zero-mass limit for the dumbbell beads and therefore multiplying through by $\mathrm{d}t$ we obtain the following two equations:

$$\zeta \left(\mathrm{d}\underset{\sim}{r}_1(t) - \underset{\sim}{u}(\underset{\sim}{r}_1(t), t)\, \mathrm{d}t \right) = \underset{\sim}{B}_1(t)\, \mathrm{d}t + \underset{\sim}{F}(\underset{\sim}{r}_2(t) - \underset{\sim}{r}_1(t))\, \mathrm{d}t, \tag{1.15}$$

$$\zeta \left(\mathrm{d}\underset{\sim}{r}_2(t) - \underset{\sim}{u}(\underset{\sim}{r}_2(t), t)\, \mathrm{d}t \right) = \underset{\sim}{B}_2(t)\, \mathrm{d}t + \underset{\sim}{F}(\underset{\sim}{r}_1(t) - \underset{\sim}{r}_2(t))\, \mathrm{d}t. \tag{1.16}$$

Equations (1.15) and (1.16) are referred to as Langevin's equations [26] for the dumbbell. The Brownian force is defined as,

$$\underset{\sim}{B}_i(t)\, \mathrm{d}t := \sqrt{2 k_B \mathcal{T} \zeta}\, \mathrm{d}\underset{\sim}{W}_i(t), \tag{1.17}$$

where $\underset{\sim}{W}_i(t)$ is a d-component Wiener process [70] and $k_B = 1.38 \times 10^{-23}\,\mathrm{m^2 kg\,s^{-2} K^{-1}}$ is Boltzmann's constant and $\mathcal{T}$ is the absolute temperature measured in Kelvin, K. The coefficient $\sqrt{2 k_B \mathcal{T} \zeta}$ in (1.17) is due to the Einstein–Smoluchowski relation, which

determines the diffusion coefficient in Brownian motion [65]. Therefore, (1.15), (1.16) can be rewritten as follows:

$$\mathrm{d}\begin{bmatrix} \chi_1(t) \\ \chi_2(t) \end{bmatrix} = \begin{bmatrix} \upsilon(\chi_1(t),t) + \zeta^{-1}\digamma(\chi_2(t)-\chi_1(t)) \\ \upsilon(\chi_2(t),t) + \zeta^{-1}\digamma(\chi_1(t)-\chi_2(t)) \end{bmatrix}\mathrm{d}t + \sqrt{\frac{2k_B\mathcal{T}}{\zeta}}\,\mathrm{d}\begin{bmatrix} \mathcal{W}_1(t) \\ \mathcal{W}_2(t) \end{bmatrix}. \quad (1.18)$$

Defining

$$X(t) := \begin{bmatrix} \chi_1(t) \\ \chi_2(t) \end{bmatrix}, \qquad W(t) := \begin{bmatrix} \mathcal{W}_1(t) \\ \mathcal{W}_2(t) \end{bmatrix}, \qquad \sigma := \sqrt{\frac{2k_B\mathcal{T}}{\zeta}}\underset{\approx}{I},$$

$$b(X(t)) := \begin{bmatrix} \upsilon(\chi_1(t),t) + \zeta^{-1}\digamma(\chi_2(t)-\chi_1(t)) \\ \upsilon(\chi_2(t),t) + \zeta^{-1}\digamma(\chi_1(t)-\chi_2(t)) \end{bmatrix},$$

(1.18) can be written as the following stochastic differential equation:

$$\mathrm{d}X(t) = b(X(t)) + \sigma(X(t))\,\mathrm{d}W(t), \qquad X(0) = X. \quad (1.19)$$

We can now use the *forward Kolmogorov equation* to obtain a partial differential equation for the evolution of the probability density function of the stochastic process $t \mapsto X(t)$ (see Corollary 5.2.10 in [53]).

Theorem 1.1 Forward Kolmogorov (Fokker–Planck) equation. *Let the random variable $X(t)$ have a density function $(\underset{\sim}{z},t) \mapsto \psi(\underset{\sim}{z},t)$ of class $\mathrm{C}^{2,1}(\mathbb{R}^d \times \mathbb{R}^d, [0,\infty))$ (*i.e. *twice continuously differentiable with respect to $\underset{\sim}{z} \in \mathbb{R}^d \times \mathbb{R}^d$ and once with respect to t), and let $X(0) = X$ be a square-integrable random variable with density function $\psi_0 \in \mathrm{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$. Also, suppose that b and σ in* (1.19) *are globally Lipschitz continuous, and $a(\underset{\sim}{z}) = \sigma(\underset{\sim}{z})\sigma(\underset{\sim}{z})^T$. Then,*

$$\frac{\partial\psi}{\partial t} + \sum_{j=1}^{2d}\frac{\partial}{\partial z_j}\left(b_j\psi\right) = \frac{1}{2}\sum_{i,j=1}^{2d}\frac{\partial^2}{\partial z_i \partial z_j}(a_{ij}\psi), \quad (1.20)$$

in $\mathbb{R}^{2d} \times [0,\infty)$ where $\psi(\underset{\sim}{z},0) = \psi_0(\underset{\sim}{z})$ for $\underset{\sim}{z} \in \mathbb{R}^d$.

Remark 1.2 *The Hookean spring force satisfies the global Lipschitz continuity assumption in Theorem 1.1, whereas the FENE force does not. Indeed, the FENE force is only locally Lipschitz on D, and it is not defined on all of $\mathbb{R}^d$. Nevertheless, we shall proceed based on the conjecture that Theorem 1.1 applies in the FENE case also.*

Applying Theorem 1.1 to (1.19) yields:

$$\begin{aligned}\frac{\partial\psi^{12}}{\partial t} &+ \underset{\sim}{\nabla}_{r_1}\cdot\left[\underset{\sim}{u}(\underset{\sim}{r}_1,t)\psi^{12}+\frac{1}{\zeta}\underset{\approx}{F}(\underset{\sim}{r}_2-\underset{\sim}{r}_1)\psi^{12}\right] \\ &+ \underset{\sim}{\nabla}_{r_2}\cdot\left[\underset{\sim}{u}(\underset{\sim}{r}_2,t)\psi^{12}+\frac{1}{\zeta}\underset{\approx}{F}(\underset{\sim}{r}_1-\underset{\sim}{r}_2)\psi^{12}\right]=\frac{k_B\mathcal{T}}{\zeta}\Delta_{r_1}\psi^{12}+\frac{k_B\mathcal{T}}{\zeta}\Delta_{r_2}\psi^{12},\end{aligned} \tag{1.21}$$

where ψ^{12} denotes the probability density function with respect to $(\underset{\sim}{r}_1,\underset{\sim}{r}_2)$-coordinates. Changing to $(\underset{\sim}{x},\underset{\sim}{q})$-coordinates and letting $\psi(\underset{\sim}{x},\underset{\sim}{q},t):=\psi^{12}(\underset{\sim}{r}_1,\underset{\sim}{r}_2,t)$, we obtain

$$\begin{aligned}\frac{\partial\psi}{\partial t} &+ \underset{\sim}{\nabla}_q\cdot\left(\left[\underset{\sim}{u}(\underset{\sim}{x}+\underset{\sim}{q}/2,t)-\underset{\sim}{u}(\underset{\sim}{x}-\underset{\sim}{q}/2,t)\right]\psi-\frac{2}{\zeta}\underset{\approx}{F}(\underset{\sim}{q})\psi\right) \\ &+ \underset{\sim}{\nabla}_x\cdot\left(\frac{\underset{\sim}{u}(\underset{\sim}{x}-\underset{\sim}{q}/2,t)+\underset{\sim}{u}(\underset{\sim}{x}+\underset{\sim}{q}/2,t)}{2}\psi\right)=\frac{k_B\mathcal{T}}{2\zeta}\Delta_x\psi+\frac{2k_B\mathcal{T}}{\zeta}\Delta_q\psi,\end{aligned} \tag{1.22}$$

where we have used the fact that $\underset{\sim}{F}(\underset{\sim}{q})=-\underset{\sim}{F}(-\underset{\sim}{q})$ (*cf.* (1.7)).

In order to simplify (1.22) further, we adopt the *local homogeneity assumption*, which states that $\underset{\sim}{u}$ and ψ are linear in $\underset{\sim}{x}$ on the length scale of a dumbbell. This is a plausible assumption because the dumbbell length scale is typically orders of magnitude smaller than the macroscopic length scale. Using linear expansions of $\underset{\sim}{u}(\underset{\sim}{x}+\underset{\sim}{q}/2)$ and $\underset{\sim}{u}(\underset{\sim}{x}-\underset{\sim}{q}/2)$ in (1.22) yields:

$$\frac{\partial\psi}{\partial t}+\underset{\sim}{\nabla}_x\cdot(\underset{\sim}{u}\psi)+\underset{\sim}{\nabla}_q\cdot\left(\underset{\approx}{\kappa}\underset{\sim}{q}\psi-\frac{2}{\zeta}\underset{\approx}{F}(\underset{\sim}{q})\psi\right)=\frac{k_B\mathcal{T}}{2\zeta}\Delta_x\psi+\frac{2k_B\mathcal{T}}{\zeta}\Delta_q\psi, \tag{1.23}$$

where $\underset{\approx}{\kappa}:=\underset{\sim}{\nabla}_x\underset{\sim}{u}$ is a standard short-hand notation for $\underset{\sim}{\nabla}_x\underset{\sim}{u}$. Note that by incompressibility of $\underset{\sim}{u}$, $\mathrm{tr}(\underset{\approx}{\kappa})=0$.

The next step is to put (1.23) into non-dimensional form by scaling as follows:

$$\underset{\sim}{x}:=L_0\hat{\underset{\sim}{x}},\qquad \underset{\sim}{q}:=l_0\hat{\underset{\sim}{q}},\qquad \underset{\sim}{u}:=U_0\hat{\underset{\sim}{u}},\qquad t:=L_0/U_0\hat{t},\qquad \psi:=\hat{\psi}/l_0^d, \tag{1.24}$$

where $l_0:=\sqrt{k_B\mathcal{T}/H}$ is the characteristic length-scale of a dumbbell and L_0, U_0 are the characteristic length and velocity of the macroscopic flow, respectively.

Applying (1.24) to (1.23) yields:

$$\frac{U_0}{L_0}\frac{\partial\psi}{\partial t}+\frac{U_0}{L_0}\underset{\sim}{\nabla}_x\cdot(\underset{\sim}{u}\psi)+\underset{\sim}{\nabla}_q\cdot\left(\frac{U_0}{L_0}\underset{\approx}{\kappa}\underset{\sim}{q}\psi-\frac{1}{2\lambda}\underset{\sim}{F}(\underset{\sim}{q})\psi\right)=\frac{1}{2\lambda}\Delta_q\psi+\frac{1}{8\lambda}\left(\frac{l_0}{L_0}\right)^2\Delta_x\psi, \tag{1.25}$$

where $\lambda:=\zeta/4H$ is the characteristic relaxation time of a dumbbell, and the hat superscripts have been dropped in (1.25) for notational convenience.

Note that for the FENE case, $|\hat{\underset{\sim}{q}}|\in[0,\sqrt{b})$ where $b:=Hl_{\max}^2/k_B\mathcal{T}$ and therefore configuration space in non-dimensional form is $D=B(0,\sqrt{b})\subset\mathbb{R}^d$, and (1.9) becomes:

$$U(\tfrac{1}{2}|\underset{\sim}{q}|^2):=-\frac{b}{2}\ln\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right),\qquad \underset{\sim}{F}(\underset{\sim}{q})=\frac{\underset{\sim}{q}}{1-|\underset{\sim}{q}|^2/b}. \tag{1.26}$$

The dimensionless parameter b is typically in the range $[10, 100]$. In [43], Jourdain, Lelièvre and Le Bris showed that for the stochastic differential equation modelling a suspension of FENE dumbbells (which corresponds to the deterministic Fokker–Planck-based model considered here), the solution exists and has trajectorial uniqueness if, and only if, $b > 2$ (*cf.* also Example 1.2 in [9]). Hence, throughout the rest of this thesis, we assume that $b \in (2, \infty)$ for the FENE potential.

Multiplying (1.25) through by L_0/U_0 gives:

$$\frac{\partial \psi}{\partial t} + \nabla_x \cdot (\underset{\sim}{u}\psi) + \nabla_q \cdot \left(\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \psi - \frac{1}{2\mathrm{Wi}} \underset{\sim}{F}(\underset{\sim}{q})\psi \right) = \frac{1}{2\mathrm{Wi}}\, \Delta_q \psi + \frac{1}{8\mathrm{Wi}} \left(\frac{l_0}{L_0} \right)^2 \Delta_x \psi, \quad (1.27)$$

where $\mathrm{Wi} := \lambda U_0 / L_0$ is the non-dimensional *Weissenberg number*, which is the ratio of the microscopic to macroscopic time-scales, and is typically on the order of 1 or 10.

Equation (1.27) contains an $\underset{\sim}{x}$-diffusion term. The standard approach in the literature has been to discard this term outright because its coefficient is typically on the order of 10^{-8} [15]. However, it has been recognised by Barrett & Süli [10] that, from the point of view of analysis, this simplification is counterproductive because when the $\underset{\sim}{x}$-diffusion term is neglected (1.27) becomes a degenerate parabolic equation that exhibits hyperbolic behaviour in physical space. Nevertheless, the focus of this thesis is on developing a computational framework for the coupled micro-macro system and, due to its small coefficient, the physical space diffusion term would have a negligible effect in such a framework. Hence, from now on we consider the Fokker–Planck equation with no $\underset{\sim}{x}$-diffusion, *i.e.*

$$\frac{\partial \psi}{\partial t} + \nabla_x \cdot (\underset{\sim}{u}\psi) + \nabla_q \cdot \left(\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \psi - \frac{1}{2\mathrm{Wi}} \underset{\sim}{F}(\underset{\sim}{q})\psi \right) = \frac{1}{2\mathrm{Wi}}\, \Delta_q \psi. \quad (1.28)$$

Notice that (at least in the case of FENE or Hookean dumbbells) the Fokker–Planck equation (1.28) contains an unbounded advection coefficient $\underset{\sim}{F}$. This is inconvenient from the point of view of analysis. Therefore we will focus on the following Kolmogorov symmetrisation [50] of the Fokker–Planck equation, in which the spring force, $\underset{\sim}{F}$, has been absorbed into a weighted diffusion term,

$$\frac{\partial \psi}{\partial t} + \nabla_x \cdot (\underset{\sim}{u}\psi) + \nabla_q \cdot (\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \psi) = \frac{1}{2\mathrm{Wi}}\, \nabla_q \cdot \left(M \nabla_q \left(\frac{\psi}{M} \right) \right), \quad (1.29)$$

where M is the (normalised) *Maxwellian* defined by

$$\underset{\sim}{q} \mapsto M(\underset{\sim}{q}) := \frac{1}{C} \exp\left(-U(\tfrac{1}{2}|\underset{\sim}{q}|^2) \right) \in \mathrm{L}^1(D), \qquad C := \int_D \exp\left(-U(\tfrac{1}{2}|\underset{\sim}{q}|^2) \right) \mathrm{d}\underset{\sim}{q}. \quad (1.30)$$

The Maxwellian transformation used in (1.29) allows us to circumvent analytical difficulties introduced by the unbounded convection term, $\underset{\sim}{F}$. In Chapter 2, we will also

consider an alternative transformation of (1.28) due to Chauvière & Lozinski [24] that allows us to deal with the unbounded convection term in a different manner, and hence a range of theoretical results can be proved for the Chauvière–Lozinski transformed equation also.

The function $(\underset{\sim}{x}, \underset{\sim}{q}, t) \mapsto \psi(\underset{\sim}{x}, \underset{\sim}{q}, t)$ represents the probability, at time t, of finding a dumbbell with center of mass in the volume element $\underset{\sim}{x} + \mathrm{d}\underset{\sim}{x}$ and orientation vector in the element $\underset{\sim}{q} + \mathrm{d}\underset{\sim}{q}$. Recall that the above derivation of the Fokker–Planck equation assumed that $\Omega = \mathbb{R}^d$, but we shall henceforth assume that Ω is a bounded subset of $\mathbb{R}^d$. Also, it is crucial to note that (1.29) is posed in $2d$ spatial dimensions, plus time. Since the computational complexity of classical numerical methods grows exponentially with the dimension of the spatial domain, the high-dimensionality of (1.29) poses a significant computational challenge. Developing a fully practical computational framework for this high-dimensional equation is a central goal of this thesis.

1.3.2 Properties of the probability density function

Since ψ is a probability density function (pdf) for each $\underset{\sim}{x} \in \Omega$, the initial condition should be non-negative:

$$\psi(\underset{\sim}{x}, \underset{\sim}{q}, 0) = \psi_0(\underset{\sim}{x}, \underset{\sim}{q}) \geq 0, \qquad \text{for a.e. } (\underset{\sim}{x}, \underset{\sim}{q}) \in \Omega \times D, \tag{1.31}$$

and should also satisfy the following normalisation property:

$$\int_D \psi_0(\underset{\sim}{x}, \underset{\sim}{q}) \, \mathrm{d}\underset{\sim}{q} = 1, \qquad \text{for a.e. } \underset{\sim}{x} \in \Omega. \tag{1.32}$$

We now show that (1.32) is preserved for $t \in (0, T]$ for solutions of (1.29). In Chapter 2, the function space $\mathfrak{K}_0$ is introduced as the space in which weak solutions of the Fokker–Planck equation in configuration space are sought and, by definition, $\sqrt{M}\mathrm{C}_0^\infty(D)$ is dense in $\mathfrak{K}_0$. We defer further discussion of $\mathfrak{K}_0$ until Chapter 2. Suppose now that $\psi(\underset{\sim}{x}, \cdot, t) \in \sqrt{M}\mathrm{C}_0^\infty(D) \subset \mathfrak{K}_0$. Then, integrating (1.29) in configuration space and applying the divergence theorem gives:

$$\begin{aligned}
&\frac{\partial}{\partial t} \int_D \psi \, \mathrm{d}\underset{\sim}{q} + \underset{\sim}{\nabla}_x \cdot \left(\underset{\sim}{u} \left(\int_D \psi \, \mathrm{d}\underset{\sim}{q} \right) \right) \\
&\qquad = \int_D \underset{\sim}{\nabla}_q \cdot \left(-\underset{\approx}{\kappa} \, \underset{\sim}{q} \, \psi + \frac{1}{2\mathrm{Wi}} M \underset{\sim}{\nabla}_q \left(\frac{\psi}{M} \right) \right) \mathrm{d}\underset{\sim}{q} \\
&\qquad\qquad = \int_{\partial D} \left(-\underset{\approx}{\kappa} \, \underset{\sim}{q} \, \psi + \frac{1}{2\mathrm{Wi}} M \underset{\sim}{\nabla}_q \left(\frac{\psi}{M} \right) \right) \cdot \underset{\sim}{n} \, \mathrm{d}s = 0,
\end{aligned}$$

where $\underset{\sim}{n}$ is the outward unit normal on ∂D, and the boundary terms vanish due to the compact support of ψ in D. This result extends to all $\mathfrak{K}_0$ by density. Let $\varrho(\underset{\sim}{x},t)$ be defined as follows:

$$\varrho(\underset{\sim}{x},t) := \int_D \psi(\underset{\sim}{x},\underset{\sim}{q},t)\,\mathrm{d}\underset{\sim}{q}. \tag{1.33}$$

Then (1.33) can be rewritten as,

$$\frac{\partial}{\partial t}\varrho(\underset{\sim}{x},t) + \underset{\sim}{\nabla}_x \cdot (\underset{\sim}{u}\,\varrho(\underset{\sim}{x},t)) = 0, \quad \text{for all } \psi \in \mathfrak{K}_0.$$

It follows from the Reynolds transport theorem, that

$$\frac{\partial}{\partial t}\int_{V(t)} \varrho(\underset{\sim}{x},t)\,\mathrm{d}\underset{\sim}{x} = 0, \qquad t \in (0,T), \tag{1.34}$$

for an arbitrary material volume $V(t)$ and hence the following result has been established.

Lemma 1.3 *Let $V(t) \subset \Omega$ be an arbitrary material volume for $t \in [0,T]$, and let $\varrho_0(\underset{\sim}{x},t) = \int_D \psi_0(\underset{\sim}{x},\underset{\sim}{q},t)$. Then*

$$\int_{V(t)} \varrho(\underset{\sim}{x},t)\,\mathrm{d}\underset{\sim}{x} = \int_{V(0)} \varrho_0(\underset{\sim}{x},t)\,\mathrm{d}\underset{\sim}{x}. \tag{1.35}$$

for all $\psi \in \mathfrak{K}_0$.

An important consideration that will be returned to in subsequent chapters is whether results analogous to Lemma 1.3 can be established for solutions (both continuous and discrete) based on the weak formulation of (1.29).

It is also desirable to preserve the property (1.31) for $t \in (0,T]$. This non-negativity property is considered for weak solutions of the Fokker–Planck equation (*cf.* Lemma 2.7) as well as for approximate solutions obtained via a Galerkin spectral approach (*cf.* Remark 2.20) in Chapter 2.

1.3.3 Polymeric extra-stress

As indicated above, in the context of the coupled Navier–Stokes–Fokker–Planck system, the purpose of solving (1.29) is so that the polymeric extra-stress tensor, $\underset{\approx}{\tau}$, can be computed and fed into the right-hand side of (1.11). The polymeric extra-stress tensor is determined by the following equation, known as *Kramers expression*:

$$\underset{\approx}{\tau}(\underset{\sim}{x},t) = n_p\left(\int_D \underset{\sim}{F}(\underset{\sim}{q}) \otimes \underset{\sim}{q}\,\psi\,\mathrm{d}\underset{\sim}{q} - \underset{\approx}{I}\right), \qquad (\underset{\sim}{x},t) \in \Omega \times (0,T], \tag{1.36}$$

where n_p is the polymer number density, *i.e.* the number of polymer molecules per unit volume. For a derivation of (1.36), see, for example, [55]. Note that it follows from (1.36) that $\underset{\approx}{\tau}$ is symmetric. Since $\underset{\approx}{\tau}$ enters into (1.11) only via its divergence, the constant $n_p\underset{\approx}{I}$ in (1.36) has no effect in the coupled system and therefore we ignore it from now on. Non-dimensionalising (1.36) according to (1.24) gives,

$$\underset{\approx}{\tau}(\underset{\sim}{x},t) = n_p k_B \mathcal{T} \int_D \underset{\sim}{F}(\underset{\sim}{q}) \otimes \underset{\sim}{q}\, \psi(\underset{\sim}{x},\underset{\sim}{q},t)\, \mathrm{d}\underset{\sim}{q}. \tag{1.37}$$

At this point we make the specific assumption that $\underset{\sim}{F}$ is the FENE force in order to derive the full Navier–Stokes–Fokker–Planck system, in non-dimensional form, for a suspension of FENE dumbbells.

It can be shown that for a dilute solution of FENE dumbbells in shear flow, the $(1,2)$-component of $\underset{\approx}{\tau}$ is approximated by,

$$\tau_{12} \approx \dot{\gamma}\lambda n_p k_B \mathcal{T} \left(\frac{b}{b+d+2}\right), \tag{1.38}$$

where $\dot{\gamma}$ is the shear rate (see [18]). Equation (1.38) is an asymptotic expression for τ_{12} that is valid for small $\dot{\gamma}$. Therefore, by analogy with Newtonian fluids, the *polymeric viscosity*, μ_p, for FENE dumbbell suspensions is defined as,

$$\mu_p := \lambda n_p k_B \mathcal{T} \left(\frac{b}{b+d+2}\right), \tag{1.39}$$

so that (1.37) can be rewritten:

$$\frac{1}{\rho}\underset{\approx}{\tau}(\underset{\sim}{x},t) = \frac{\nu_p}{\lambda}\frac{b+d+2}{b} \int_D \underset{\sim}{F}(\underset{\sim}{q}) \otimes \underset{\sim}{q}\, \psi(\underset{\sim}{x},\underset{\sim}{q},t)\, \mathrm{d}\underset{\sim}{q}, \tag{1.40}$$

where the equation has been divided through by the density ρ as in (1.11), and $\nu_p := \mu_p/\rho$.

Equation (1.40) provides a bridge between the Fokker–Planck equation and the Navier–Stokes equation. The full coupled form of the micro-macro system is discussed in the next section.

1.3.4 The coupled Navier–Stokes–Fokker–Planck system

The Fokker–Planck equation and Kramers expression have been written in terms of non-dimensional variables in (1.29) and (1.40), respectively. Therefore, it remains to non-dimensionalise the Navier–Stokes equations, (1.11), (1.12), in the same manner. The mass conservation equation, (1.12), contains only one non-zero term and therefore

rescaling is trivial. Applying (1.24) in the momentum equation, letting $\nu = \nu_s + \nu_p$, rescaling pressure as $p = U_0^2\,\hat{p}$ and using (1.40) yields

$$\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\underset{\sim}{u} + \underset{\sim}{\nabla}_x p = \frac{\gamma}{\mathrm{Re}}\Delta_x\underset{\sim}{u} + \frac{b+d+2}{b}\frac{1-\gamma}{\mathrm{Re\,Wi}}\underset{\sim}{\nabla}_x\cdot\underset{\approx}{\tau}, \tag{1.41}$$

where $\mathrm{Re} := L_0U_0/\nu$ (*i.e.* the Reynolds number) and $\gamma := \nu_s/\nu$ are non-dimensional parameters.[3] Note that we have absorbed the coefficients on the right-hand side of (1.40) into the momentum equation in order to perform non-dimensionalisation.

Combining the equations heretofore derived gives the following system:

$$\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\underset{\sim}{u} + \underset{\sim}{\nabla}_x p = \frac{\gamma}{\mathrm{Re}}\Delta_x\underset{\sim}{u} + \frac{b+d+2}{b}\frac{1-\gamma}{\mathrm{Re\,Wi}}\underset{\sim}{\nabla}_x\cdot\underset{\approx}{\tau}, \qquad (\underset{\sim}{x},t)\in\Omega\times(0,T], \tag{1.42}$$

$$\underset{\sim}{\nabla}_x\cdot\underset{\sim}{u} = 0, \qquad (\underset{\sim}{x},t)\in\Omega\times(0,T], \tag{1.43}$$

$$\frac{\partial \psi}{\partial t} + \underset{\sim}{\nabla}_x\cdot(\underset{\sim}{u}\psi) + \underset{\sim}{\nabla}_q\cdot(\underset{\approx}{\kappa}\,\underset{\sim}{q}\,\psi) = \frac{1}{2\mathrm{Wi}}\,\underset{\sim}{\nabla}_q\cdot\left(M\underset{\sim}{\nabla}_q\frac{\psi}{M}\right), \qquad (\underset{\sim}{x},\underset{\sim}{q},t)\in\Omega\times D\times(0,T], \tag{1.44}$$

$$\underset{\approx}{\tau}(\underset{\sim}{x},t) = \int_D \underset{\sim}{F}\otimes\underset{\sim}{q}\,\psi(\underset{\sim}{x},\underset{\sim}{q},t)\,\mathrm{d}\underset{\sim}{q}, \qquad (\underset{\sim}{x},t)\in\Omega\times(0,T], \tag{1.45}$$

$$\underset{\sim}{u}(\underset{\sim}{x},0) = \underset{\sim}{u}_0(\underset{\sim}{x}), \quad \underset{\sim}{x}\in\Omega, \qquad \psi(\underset{\sim}{x},\underset{\sim}{q},0) = \psi_0(\underset{\sim}{x},\underset{\sim}{q}), \quad (\underset{\sim}{x},\underset{\sim}{q})\in\Omega\times D. \tag{1.46}$$

Equations (1.42)–(1.46) are the coupled Navier–Stokes–Fokker–Planck model for dilute polymeric fluids. Note that the non-dimensionalisation used above is the same as the one introduced on page 8 of [55]. In Chapter 4, we also consider a Stokes–Fokker–Planck model in which (1.42) is replaced by a simpler linear equation (*cf.* (4.4)) that is relevant for modelling creeping flows, *i.e.* in the limit $\mathrm{Re}\to 0_+$.

In the discussion above, we have assumed that both Ω and D are domains in $\mathbb{R}^d$ so that the Fokker–Planck equation is posed on $\Omega\times D\subset\mathbb{R}^{2d}$. However, it is not essential that this is the case and, for example, in [23] the authors considered a micro-macro model in which $\Omega\subset\mathbb{R}^2$ and $D\subset\mathbb{R}^3$. No significant complications are introduced from the theoretical or implementational point of view by allowing the dimensionality of D and Ω to be different, but for the rest of this dissertation we will restrict our attention to the case when these domains have the same dimensionality.

1.4 Literature review of numerical methods for simulating polymeric fluids

As indicated in the opening of this chapter, the techniques for numerically simulating polymeric fluids can be grouped into three categories: fully macroscopic methods,

[3] Hat superscripts have again been dropped in (1.41) for notational simplicity; the variables are to be understood as non-dimensional.

stochastic multiscale methods and deterministic multiscale methods. A survey of some of the key literature for each method is presented below.

Fully macroscopic methods

Continuum numerical simulations of polymeric fluids have been popular since the 1970's. In some sense, this is the most natural approach to simulating polymeric fluids because by avoiding consideration of the microscopic length scales, one can save an enormous amount of computational effort. However, except in certain simple cases (*e.g.* a suspension of Hookean dumbbells, see Section 1.2) in order to derive a closed form macroscopic model for a polymeric fluid, it is necessary to resort to an ad hoc "closure approximation", and the shortcomings of such approximations are well documented [45, 56, 84]. Nevertheless, in many situations, macroscopic models are sufficiently accurate to capture the relevant characteristics of polymer flows and in such cases these methods are preferable to using multiscale methods.

A macroscopic computation typically employs standard tools of computational fluid dynamics, such as finite elements, finite volumes or spectral methods, but specialised considerations are usually necessary in practice in order to ensure convergence. The challenges of developing continuum numerical methods for polymeric fluids are epitomised by the well-known "high Weissenberg number" problem, which refers to the difficulty of developing numerical methods that remain stable as Wi is increased. The development of macroscopic numerical methods for polymer fluids is clearly a very important field of research; a vast literature has been developed and yet there remain many unresolved issues in this area that are the focus of ongoing research. However, since the focus of this thesis is on multiscale methods, we will not consider fully macroscopic methods any further here (for a detailed discussion, see the book by Owens & Phillips [69]).

Stochastic multiscale methods

An alternative approach that has gained popularity since the early 1990's is to treat the micro-macro model directly by solving the stochastic differential equation (1.19) using Monte Carlo-type methods and coupling with deterministic numerical methods for solving the Navier–Stokes equations (1.11), (1.12). The Monte Carlo method involves distributing a large number of model polymer molecules throughout the computational domain and tracking their motion as they are convected along streamlines and stretched and oriented by a flow. The stress field, $\underset{\approx}{\tau}$, can then be determined by computing ensemble averages, so that the Navier–Stokes equations can then be solved (with source

term $\nabla_x \cdot \underset{\sim}{\tau}$) to determine the macroscopic velocity field, typically using finite-elements or some other standard CFD method. In 1992 Öttinger & Laso [54] proposed the first scheme of this type, which is referred to by the acronym CONNFFESSIT for "Calculation of Non-Newtonian Flow: Finite Elements and Stochastic Simulation Technique". Many other flavours of stochastic multiscale methods have subsequently been developed, such as the method of Brownian configuration fields [40] and the Lagrangian particle method [36]. Note also that there has been a lot of interest in the mathematical properties of multiscale stochastic methods. For example, existence and uniqueness of solutions have been established for stochastic simulations of suspensions of Hookean and FENE dumbbells in papers by Jourdain, Lelièvre & Le Bris [41, 42, 43].

The stochastic multiscale approach is a computationally intensive procedure – it is little wonder, therefore, that there was no work done in this direction prior to the 1990's. Moreover, a drawback of the stochastic approach is that it introduces a slowly decaying stochastic error (typically $\mathcal{O}(N^{-1/2})$ as $N \to \infty$). Variance reduction techniques were developed to ameliorate this error term and reduce the number of polymer molecules one must track in order to compute an ensemble average to within a given error tolerance (see [46] for an overview of variance reduction in this context). However, even with variance reduction techniques, the presence of stochastic error is a significant limitation of the stochastic approaches and circumventing this is an important motivation for moving to deterministic methods. On the other hand, an important advantage of the stochastic approach is that it scales well with the number of degrees of freedom in the polymer model – this ensures that stochastic methods remain effective when applied to bead-spring chain polymer models [46].

Deterministic multiscale methods

As indicated earlier, the deterministic multiscale approach involves solving the coupled Navier–Stokes–Fokker–Planck system directly. This approach has received comparatively little attention, most likely because solving the high-dimensional Fokker–Planck equation is an imposing computational challenge. Nevertheless, literature on this method extends back to the 1970's although the early works in which the Fokker–Planck equation was solved directly were not truly multiscale since simplified flow regimes were considered for which ψ was assumed to be a function of $\underset{\sim}{q}$ and t only (problems in which ψ does not depend on $\underset{\sim}{x}$ are often referred to as *homogeneous flows*). For example, Stewart & Sørensen in 1972 [76] used spherical harmonics to solve the Fokker–Planck equation for a steady shear flow of a dilute suspension of rigid dumbbells. Warner [82] applied a similar approach to the study of shear flows of FENE

dumbbells. The first work in which a deterministic approach was utilised to simulate a non-homogeneous velocity flow was by Fan in 1989 [32], who considered a planar channel flow using a rigid dumbbell polymer model, and also made the simplifying assumption that the physical space convection term, $\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \psi$, vanished. Fan's work was subsequently built upon by Nayak [66] and Grosso *et al.* [35] who eliminated this assumption on $\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \psi$.

Recently, the deterministic multiscale approach has been further developed by Lozinski, Chauviére and collaborators, who proposed a spectral method for simulating the micro-macro model for dilute solutions of FENE dumbbells [23, 24, 59, 60, 61]. Similarly, Helzel & Otto [38] solved the micro-macro model arising in the simulation of suspensions of rod-like polymers using finite difference and finite volume methods.

In the papers of Lozinski, Chauviére *et al.* and Helzel & Otto, the authors decomposed the Fokker–Planck equation (1.28) (*i.e.* in the non-symmetrised form) according to

$$\frac{\partial \psi}{\partial t} + (L_x + L_q)\,\psi = 0, \tag{1.47}$$

where

$$L_q \psi = \underset{\sim}{\nabla}_q \cdot (\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \psi) - \frac{1}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_q \cdot \underset{\sim}{F}(\underset{\sim}{q}) \psi + \Delta_q \psi \right), \tag{1.48}$$

$$L_x \psi = \underset{\sim}{\nabla}_x \cdot (\underset{\sim}{u} \psi), \tag{1.49}$$

and then they used an alternating-direction approach based on the operators L_q and L_x to compute numerical solutions.

That is, suppose that $0 = t^0 < t^1 < \cdots < t^n < \cdots \leq T$ is a uniform partition of spacing Δt of the interval $[0, T]$. A (two-stage) alternating-direction scheme involves approximating the solution, ψ, by ψ_2 in the following manner: given $\psi_2(t^n)$, $n \geq 0$, with $\psi_2(t^0) = \psi_0$, find ψ_1 and ψ_2 such that,

$$\frac{\partial \psi_1}{\partial t} + L_q \psi_1 = 0, \qquad t \in (t^n, t^{n+1}], \qquad \psi_1(t^n) = \psi_2(t^n), \tag{1.50}$$

$$\frac{\partial \psi_2}{\partial t} + L_x \psi_2 = 0, \qquad t \in (t^n, t^{n+1}], \qquad \psi_2(t^n) = \psi_1(t^{n+1}). \tag{1.51}$$

A practical alternating-direction numerical method is based on spatial and temporal discretisation of (1.50) and (1.51).

In the case of the Fokker–Planck equation, (1.50) is a convection-diffusion equation posed on D and (1.51) is a first-order hyperbolic equation on Ω. After discretising in space and time, the two-stage scheme described above can be implemented by alternating between applying L_x to Ω cross sections of $\Omega \times D$ and L_q to D cross-sections

of $\Omega \times D$. This type of scheme is also referred to as a dimension splitting or operator splitting approach. We will use the three terms (*i.e.* alternating direction/dimension splitting/operator splitting) interchangeably in this thesis, but our preference will be for the name 'alternating-direction method', since we believe it is more descriptive than the alternatives.

Using this operator-splitting, the "curse of dimensionality" associated with the numerical solution of the Fokker–Planck equation in $2d$ dimensions is ameliorated, as the splitting leads to a sequence of d-dimensional solves at each time step rather than a single $2d$-dimensional solve. Also, this splitting of L allows different numerical methods to be used in Ω and D. In Chapter 3 we consider alternating-direction numerical methods for the FENE Fokker–Planck equation on $\Omega \times D$ and we use a heterogeneous alternating-direction method based on a finite-element method in Ω and a single-domain Galerkin spectral in D. These are appropriate choices because a finite-element method is flexible enough to deal with the general domain Ω, whereas D is always a ball in $\mathbb{R}^d$ and is therefore the L_q operator is well suited to a spectral discretisation via a polar or spherical co-ordinate transformation to a cartesian product domain. Note also that we shall primarily focus is on the Maxwellian transformed Fokker–Planck equation and therefore instead of L_q as defined in (1.48), we will generally consider the following $\underset{\sim}{q}$-direction operator:

$$L_q\psi = \underset{\sim}{\nabla}_q \cdot (\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \psi) - \frac{1}{2\mathrm{Wi}} \underset{\sim}{\nabla}_q \cdot \left(M \underset{\sim}{\nabla}_q \left(\frac{\psi}{M} \right) \right). \tag{1.52}$$

The operators (1.48) and (1.52) are identical. However, as we discuss in Chapter 2, the natural weak formulation of (1.52), in which we use test functions φ/M, is not identical to the standard weak formulation of (1.48) in which unweighted test functions, φ, are used.

Lozinski & Chauvière [23,24,60] demonstrated that compared to a stochastic method for the FENE dumbbell model, their deterministic multiscale scheme was more efficient in terms of computational cost, and was also more accurate due to the absence of stochastic error for the benchmark problem of laminar flow around a cylindrical obstacle in a channel.

A further interesting observation by Lozinski & Chauvière was that the direct discretisation of (1.28) did not lead to a stable numerical method, and instead they used a substitution of the form $\psi/(1 - |\underset{\sim}{q}|^2/b)^s$, for some s that is chosen on computational grounds (for example, the authors recommended $s = 2$ and $s = 2.5$ for $d = 2$ and $d = 3$, respectively [23, 24]). We return to this point in Section 2.2.1 where we show

that the bilinear form corresponding to the Chauvière–Lozinski-transformed FENE Fokker–Planck equation is coercive for $s > 1/2$, hence it is not surprising that Lozinski & Chauvière's method was unstable when no substitution was used.

Based on the results of Lozinski & Chauvière, it is clear that the deterministic multiscale approach can be effective for models with low-dimensional configuration space. However, it is still an open question whether this approach can be extended to bead-spring chain dumbbell models in which configuration space has dimension greater than three. There has been some recent work in this direction using numerical methods that were developed for high-dimensional (*i.e.* $d \gg 3$) PDEs. For example, Ammar, Mokdad, Chinesta & Keunings developed a reduced basis approach and used it to solve the Fokker–Planck equation in configuration space of dimension up to 20 [2, 3]. An alternative idea is to use sparse grids, which have been shown to be effective for solving elliptic and parabolic PDEs in high-dimensional domains [73,80]. This idea was applied to the Fokker–Planck equation by Delaunay, Lozinski & Owens [27]. Attempts to solve the Fokker–Planck equation for configuration spaces for $d \gg 3$ are still at an early stage, and indeed the numerical results presented in the literature so far have been for homogeneous flows only. Nevertheless, reduced basis and sparse grid methods appear to be a promising approach for this problem and may enable the development of efficient deterministic multiscale methods for simulating suspensions of bead-spring chains.

Clearly the well-posedness of the Navier–Stokes–Fokker–Planck system is a prerequisite for the success of the deterministic multiscale approach. PDE analysis of the micro-macro model is outside the scope of this thesis, but it is worth noting here that there have been a number of recent papers in which the question of existence of solutions (among many other things) has been considered (*e.g.* see [6, 9, 10, 57, 58]). The review article of Li & Zhang [55] provides an informative overview of this literature.

1.5 Outlook and goals

We are now in a position to give more details on the aims of this thesis. Our focus is on the deterministic multiscale method. As discussed in Section 1.4, several different deterministic multiscale numerical methods have been developed in the literature, but the numerical analysis of these methods has not previously been considered in detail. The central goal of this work, therefore, is to develop rigorous analysis of deterministic multiscale methods in order to ensure that there is a firm theoretical foundation for this approach.

We begin in Chapter 2, by focusing on the analysis of a Galerkin spectral method for discretising (1.50), *i.e.* the $\underset{\sim}{q}$-direction part of the Fokker–Planck equation (or equivalently, the Fokker–Planck equation for a homogeneous flow problem). The focus in Chapter 2, is on the Maxwellian transformed Fokker–Planck equation (*cf.* (1.52)), but we also consider the transformation proposed by Chauvière & Lozinski for (1.48) in some detail. Numerical methods based on either transformation require careful analysis; the Maxwellian weight arising in the principal part of the symmetrised formulation is degenerate in the sense that it vanishes on ∂D, and the Chauvière–Lozinski-transformed scheme contains the unbounded convection coefficient $\underset{\sim}{F}$. We also pay particular attention to the practical implementation of the spectral method on D, and we present numerical results for the cases $d = 2$ and $d = 3$.

In Chapter 3, the Galerkin spectral method developed in Chapter 2 is combined with a finite element method in Ω to yield the alternating-direction scheme with which we obtain approximate solutions of (1.29). We show that some subtle issues arise in the numerical analysis of such alternating-direction schemes and, as a result, we develop a specialised quadrature-based Galerkin alternating-direction method for the Fokker–Planck equation that is amenable to stability and convergence analysis; this analysis builds upon the arguments in Chapter 2. We also present some computational results in order to provide experimental support for our theoretical results, and to demonstrate the effectiveness of our alternating-direction approach in practice.

The focus in Chapter 4 is on obtaining computational results for the Navier–Stokes–Fokker–Planck system. Our approach is to couple a standard finite element scheme for solving the Navier–Stokes equations with an alternating-direction method from Chapter 3 for the Fokker–Planck equation. Solving the Fokker–Planck equation is the bottleneck step in this algorithm, due to the fact that it is posed on $\Omega \times D$. The numerical results in Chapter 4, and indeed in Chapters 2 and 3 as well, are for the FENE dumbbell case only. However, it would be straightforward to apply the methods developed in this thesis to more general dumbbell spring potentials, such as potentials that satisfy Hypotheses A and B defined in Chapter 2.

Finally, we want to emphasise an important innovation developed in this thesis: the application of parallel computation to alternating-direction numerical methods for the Fokker–Planck equation. Alternating-direction algorithms are well suited to implementation on parallel computers since they involve solving a large number of independent equations in each time-step. We show in Chapters 3 and 4 that our alternating direction approach can be efficiently implemented in parallel, and this enables us to solve large-

scale deterministic multiscale problems that may otherwise have been computationally intractable (*e.g.* an important large-scale case is when $\Omega \times D \subset \mathbb{R}^6$).

Chapter 2

The Fokker–Planck Equation in Configuration Space

This chapter is concerned with the numerical approximation of the d-dimensional Fokker–Planck equation posed in configuration space:

$$\frac{\partial \psi}{\partial t} + \nabla_{\underset{\sim}{q}} \cdot (\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \psi) = \frac{1}{2\mathrm{Wi}}\, \nabla_{\underset{\sim}{q}} \cdot \left(M \nabla_{\underset{\sim}{q}} \frac{\psi}{M} \right), \ (\underset{\sim}{q}, t) \in D \times (0,T], \tag{2.1}$$

where the $d \times d$ tensor $\underset{\approx}{\kappa}$ is assumed to belong to $(\mathrm{C}[0,T])^{d\times d}$ (*i.e.* it is independent of $\underset{\sim}{x}$) and is such that $\mathrm{tr}(\underset{\approx}{\kappa})(t) = 0$ for all $t \in [0,T]$. It will be assumed throughout that (2.1) is supplemented with the following initial and boundary conditions:

$$\psi(\underset{\sim}{q}, 0) = \psi_0(\underset{\sim}{q}), \qquad \text{for all } \underset{\sim}{q} \in D, \tag{2.2}$$

$$\psi(\underset{\sim}{q}, t) = o\left(\sqrt{M(\underset{\sim}{q})} \right), \qquad \text{as } \mathrm{dist}(\underset{\sim}{q}, \partial D) \to 0_+, \ \text{ for all } t \in (0,T]. \tag{2.3}$$

The boundary condition (2.3) follows from the weak formulation of (2.1) developed below (*cf.* (2.5), (2.6)). The initial datum ψ_0 is such that $\psi_0 \geq 0$ and $\int_D \psi_0(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} = 1$, as in (1.31) and (1.32). We will henceforth use the notation $\mathfrak{d}(\underset{\sim}{q}) := \mathrm{dist}(\underset{\sim}{q}, \partial D) = \sqrt{b} - |\underset{\sim}{q}|$.

The motivation for studying this subproblem is that, as indicated in Chapter 1, an efficient approach to the numerical solution of (1.44) in $2d+1$ variables is based on operator-splitting with respect to $(\underset{\sim}{q}, t)$ and $(\underset{\sim}{x}, t)$ as in (1.50), (1.51). Thereby, the resulting time-dependent transport equation with respect to $(\underset{\sim}{x}, t)$ is completely standard, $\psi_t + \nabla_{\underset{\sim}{x}} \cdot (\underset{\sim}{u}(\underset{\sim}{x}, t)\psi) = 0$, while the transport-diffusion equation with respect to $(\underset{\sim}{q}, t)$ is (2.1).

The focus of this chapter is on the analysis and implementation of spectral methods for computing numerical solutions of (2.1). We emphasise rigour in establishing the analytical properties of the weak formulation of (2.1) and also in developing spectral

convergence estimates for the numerical methods based on this weak formulation. Most of the material in this chapter follows the paper [49].

As indicated in Chapter 1, we are primarily interested in solving the micro-macro equations for FENE dumbbells. However, the analysis in this chapter is valid for a more general class of spring force laws. Therefore, the following structural hypotheses, which generalise the relevant properties of the FENE spring potential, are adopted.

Hypothesis A. The spring potential $U \in \mathrm{C}^1([0, \frac{b}{2}))$ is a non-negative monotonic increasing function, with $U(0) = 0$, $\lim_{s\to b/2_-} U(s) = +\infty$, $\lim_{s\to b/2_-} (\frac{b}{2} - s)U'(s) < \infty$. $\diamond$

Hypothesis A is consistent with the physical requirement that, in order to faithfully model *finite* stretching of polymer chains, the spring force $\underset{\sim}{F}(\underset{\sim}{q})$ should have infinite intensity when the maximum admissible elongation $|\underset{\sim}{q}| = \sqrt{b}$ is reached; *i.e.*, the function $\underset{\sim}{q} \mapsto U'(\frac{1}{2}|\underset{\sim}{q}|^2)$ should tend to $+\infty$ as $\mathfrak{d}(\underset{\sim}{q}) \to 0_+$.

Recall the definition of the Maxwellian M for a spring potential U, (1.30). Since, by Hypothesis A, $U(\frac{1}{2}|\underset{\sim}{q}|^2) \to +\infty$ as $\mathfrak{d}(\underset{\sim}{q}) \to 0_+$, it follows that $M(\underset{\sim}{q}) \to 0_+$ as $\mathfrak{d}(\underset{\sim}{q}) \to 0_+$.

Hypothesis B. $\sqrt{M} \in \mathrm{H}^1_0(D)$, and M is a *weight function of type 3* on D in the sense of Triebel [78], p.247, Definition 3.2.1.3c; *i.e.*, there exist positive constants c_1, c_2 and λ, and a positive monotonic increasing function τ defined on the interval $(0, \lambda)$, such that $c_1\, \tau(\mathfrak{d}(\underset{\sim}{q})) \le M(\underset{\sim}{q}) \le c_2\, \tau(\mathfrak{d}(\underset{\sim}{q}))$ for all $\underset{\sim}{q} \in D$ satisfying $\mathfrak{d}(\underset{\sim}{q}) < \lambda$. $\diamond$

Hypotheses A and B will be assumed throughout this chapter.

Example 2.1 *Consider the function U defined by*

$$U(s) := -f(s) \ln\left(1 - \frac{2s}{b}\right), \qquad s \in [0, \tfrac{b}{2}), \qquad \text{with } b > 2,$$

where $f \in \mathrm{C}^\infty[0, \frac{b}{2}]$ is a monotonic nondecreasing function, positive on $(0, \frac{b}{2}]$, with $f(\frac{b}{2}) > 1$; then U and the associated Maxwellian M satisfy hypotheses A and B, respectively. When $f(s) = b/2$, the FENE potential is recovered.

The central difficulty of (2.1), (2.2), (2.3), from both the analytical and the computational point of view, is the presence in (2.1) of the degenerate Maxwellian $M(\underset{\sim}{q})$, with $\lim_{\mathfrak{d}(\underset{\sim}{q}) \to 0_+} M(\underset{\sim}{q}) = 0$.

Most numerical methods developed for the Fokker–Planck equation have been based on the 'original' form of the equation,

$$\frac{\partial \psi}{\partial t} + \underset{\sim}{\nabla}_q \cdot \Big(\underset{\approx}{\kappa}\underset{\sim}{q}\, \psi\Big) = \frac{1}{2\mathrm{Wi}}\, \underset{\sim}{\nabla}_q \cdot \Big(\underset{\sim}{\nabla}_q \psi + \underset{\sim}{F}(\underset{\sim}{q})\, \psi\Big), \tag{2.4}$$

see, for example, [23, 24, 60] or [2, 3]. From the theoretical viewpoint at least, the advantage of (2.1) over (2.4), is that on transformation into weak form the diffusion operator becomes symmetric (see (2.5)), which facilitates the analysis of the Fokker–Planck equation for a general class of Maxwellians. Notwithstanding this potential theoretical advantage, the computational benefits, or otherwise, of discretising (2.1) rather than (2.4) remain to be understood.

The aims of the analysis in this chapter are therefore two-fold:

(a) The principal objective is to develop the mathematical and numerical analysis of equation (2.1) for the class of Maxwellians satisfying Hypotheses A and B. The discretisation of the equation is based on a spectral Galerkin method in the spatial variable $\underset{\sim}{q}$ coupled with backward Euler time-stepping. One can, of course, consider more accurate time discretisation schemes, such as an nth-order backward differentiation formula, BDFn, $n \in \{2, \ldots, 6\}$, for example. High-order time discretisation of the problem is, however, a secondary consideration to the central theme of this chapter, and it is not discussed here.

(b) In the special case of the FENE model, it shall be shown how the results under (a) can be adapted to the case of alternative discretisation proposed by Chauvière & Lozinski [23, 24, 59, 60], which applies a transformation, different from the symmetrising transformation considered under (a), to the 'original' form (2.4) of the Fokker–Planck equation. The transformed equation is then approximated in the same way as in (a), using a spectral Galerkin method in space and a backward Euler discretisation in time.

Since the analytical arguments under (b) are almost identical to those under (a), for the sake of brevity, attention will be focused on (a), but the key adjustments that need to be made in order to obtain the corresponding results under (b) shall be systematically indicated.

First of all, we define the function spaces relevant to the weak formulation of (2.1). Note that since only configuration space is considered in this chapter, $\|\cdot\|$ and $(\cdot,\cdot)$ will denote the $\mathrm{L}^2(D)$ norm and inner-product, respectively. In subsequent chapters when numerical methods for the Fokker–Planck equation on physical space as well as configuration space are considered, the non-subscripted norm and inner-product will imply the domain $\Omega \times D$.

Let

$$\mathfrak{H} := \left\{ \varphi \in \mathrm{L}^2_{\mathrm{loc}}(D) \,:\, \int_D \left(\frac{\varphi}{\sqrt{M}} \right)^2 \mathrm{d}\underset{\sim}{q} < \infty \right\},$$
$$\mathfrak{K} := \left\{ \varphi \in \mathfrak{H} \,:\, \int_D \left(\left(\frac{\varphi}{\sqrt{M}} \right)^2 + \left| \sqrt{M}\, \underset{\sim}{\nabla}_q \left(\frac{\varphi}{M} \right) \right|^2 \right) \mathrm{d}\underset{\sim}{q} < \infty \right\},$$

and define $\mathfrak{K}_0$ as the closure of $\sqrt{M}\mathrm{C}_0^\infty(D)$ in the norm of $\mathfrak{K}$. Taking test functions as φ/M with $\varphi \in \mathfrak{K}_0$, we get the following weak formulation of the initial-boundary-value problem (2.1).

Given $\psi_0 \in \mathfrak{H}$, find $\psi \in \mathrm{L}^\infty(0,T;\mathfrak{H}) \cap \mathrm{L}^2(0,T;\mathfrak{K}_0)$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_D \frac{\psi\,\varphi}{M}\, \mathrm{d}\underset{\sim}{q} - \int_D \underset{\approx}{\kappa}\, \underset{\sim}{q}\, \frac{\psi}{\sqrt{M}} \cdot \sqrt{M}\, \underset{\sim}{\nabla}_q \left(\frac{\varphi}{M} \right) \mathrm{d}\underset{\sim}{q} \tag{2.5}$$
$$+ \frac{1}{2\mathrm{Wi}} \int_D \sqrt{M}\, \underset{\sim}{\nabla}_q \left(\frac{\psi}{M} \right) \cdot \sqrt{M}\, \underset{\sim}{\nabla}_q \left(\frac{\varphi}{M} \right) \mathrm{d}\underset{\sim}{q} = 0 \quad \forall \varphi \in \mathfrak{K}_0,$$

in the sense of distributions on $(0,T)$, and $\psi(\cdot,0) = \psi_0(\cdot)$.

Now, by introducing the notation

$$\hat{\varphi} := \frac{\varphi}{\sqrt{M}} \quad \text{and} \quad \underset{\sim}{\nabla}_M \hat{\varphi} := \sqrt{M}\, \underset{\sim}{\nabla}_q \left(\frac{\hat{\varphi}}{\sqrt{M}} \right)$$

(2.5) can be reformulated on observing that, by the definition of $\mathfrak{K}$, $\varphi \in \mathfrak{K}_0$ if, and only if, $\hat{\varphi} \in \mathrm{H}^1_0(D;M)$, where $\mathrm{H}^1_0(D;M)$ is the closure of $\mathrm{C}_0^\infty(D)$ in the norm of $\mathrm{H}^1(D;M)$, and

$$\mathrm{H}^1(D;M) := \left\{ \zeta \in \mathrm{L}^2(D) : \|\zeta\|^2_{\mathrm{H}^1(D;M)} := \int_D \left(|\zeta|^2 + |\underset{\sim}{\nabla}_M \zeta|^2 \right) \mathrm{d}\underset{\sim}{q} < \infty \right\}.$$

When applied to an element of $\mathrm{H}^1_0(D;M)$ the norm $\|\cdot\|_{\mathrm{H}^1(D;M)}$ will be written $\|\cdot\|_{\mathrm{H}^1_0(D;M)}$. As a matter of fact, it shall be shown in Section 2.1 that $\mathrm{C}_0^\infty(D)$ is dense in $\mathrm{H}^1(D;M)$ and therefore, perhaps somewhat counter-intuitively, $\mathrm{H}^1_0(D;M) = \mathrm{H}^1(D;M)$, and also $\mathfrak{K}_0 = \mathfrak{K}$.

Remark 2.2 *We note in passing that the substitution $\hat{\varphi} = \varphi/\sqrt{M}$ also appears in the recent paper by Du, Liu and Yu [29], though the operator $\underset{\sim}{\nabla}_M$ does not.*

*In the case of the FENE Maxwellian (*cf. *Example 2.1), Chauvière & Lozinski [23, 24, 59, 60] used a spectral method to approximate $\psi/M^{2s/b}$ instead of $\psi/\sqrt{M}$, where s is a parameter that was chosen on the basis of numerical experiments. Clearly, the two expressions coincide when $s = b/4$; on the other hand, the values $s = 2$ and $s = 2.5$ were recommended in [23, 24, 59, 60] on computational grounds for $d = 2$*

and $d = 3$, respectively. More will be said in Sections 2.2, 2.3 and 2.5 about the analytical implications of using, in the special case of the FENE model, the substitution $\hat{\psi} := \psi/M^{2s/b}$ instead of the substitution $\hat{\psi} := \psi/\sqrt{M}$. In particular, we shall show that both substitutions result in unconditionally stable and convergent numerical methods, although in the case of the Chauvière & Lozinski type substitution it will be necessary to assume for this purpose that $b \geq 4s^2/(2s-1)$ with $s > 1/2$, while the symmetrised formulation based on (2.1) will be seen to result in a stable and optimally convergent scheme for all $b > 2$. In Section 2.6 we shall perform quantitative comparisons of the two approaches through numerical experiments. $\diamond$

With these notational conventions, (2.5) has the following form.

Given $\hat{\psi}_0 := \psi_0/\sqrt{M} \in \mathrm{L}^2(D)$, find $\hat{\psi} \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(D)) \cap \mathrm{L}^2(0,T;\mathrm{H}^1_0(D;M))$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_D \hat{\psi}\,\hat{\varphi}\;\mathrm{d}\underset{\sim}{q} - \int_D \underset{\approx}{\kappa}\,\underset{\sim}{q}\hat{\psi}\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\;\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_M\hat{\psi}\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\;\mathrm{d}\underset{\sim}{q} = 0 \qquad \forall\hat{\varphi}\in\mathrm{H}^1_0(D;M), \tag{2.6}$$

in the sense of distributions on $(0,T)$, and $\hat{\psi}(\cdot,0) = \hat{\psi}_0(\cdot)$.

The function space $\mathrm{H}^1(D;M)$ may appear exotic. However, it will be shown in Section 2.1 that, under Hypotheses A and B, $\mathrm{H}^1(D;M) = \mathrm{H}^1_0(D;M)$ and $\mathrm{H}^1_0(D) \subset \mathrm{H}^1_0(D;M)$. The connection between $\mathrm{H}^1_0(D;M)$ and $\mathrm{H}^1_0(D)$ will prove helpful in the development of Galerkin methods for (2.6), since the construction of finite-dimensional subspaces of $\mathrm{H}^1_0(D)$ and the analysis of their approximation properties are well understood.

In Section 2.2 the weak formulation (2.6) of the initial boundary value problem will be revisited. A backward Euler semidiscretisation of the weak formulation shall be constructed, and the unconditional stability of the temporal semidiscretisation in the $\ell^\infty(0,T;\mathrm{L}^2(D))$ and $\ell^2(0,T;\mathrm{H}^1_0(D;M))$ norms shall be established. Also, in the case of the FENE model with $b \geq 4s^2/(2s-1)$ and $s > 1/2$, it will be demonstrated that these results can be carried across, independent of the spatial dimension d, to a weak formulation that results from using the alternative substitution $\hat{\psi} := \psi/M^{2s/b}$; the cases of $s = 2$ and $s = 2.5$ correspond to the methods proposed by Chauvière & Lozinski for $d = 2$ and $d = 3$, respectively.

In Section 2.3 the fully-discrete method is developed and, using the stability results from Section 2.2, a bound on the global error in terms of the approximation error in a suitably defined spectral projection operator is derived.

In Section 2.4, the precise definition of the projection operator is given: its non-standard form stems from a *decomposition lemma*, Lemma 2.14, for elements of the

Sobolev space $\mathrm{H}^1(D)$ transformed to polar coordinates. For ease of presentation, we confine ourselves to the case of two space dimensions ($d = 2$) in Section 2.4; analogous arguments could be developed in the $d = 3$ case.

The convergence analysis is completed in Section 2.5 by showing that, under Hypotheses A and B, the method exhibits optimal-order convergence in the Maxwellian-weighted norm $\|\cdot\|_{\ell^2(0,T;\mathrm{H}^1_0(D;M))}$ with respect to the spatial and temporal discretisation parameters.

Section 2.6 is devoted to numerical experiments that illustrate the performance of the method. We focus solely on the FENE potential in this section. First of all, we discuss the implementation of our Galerkin spectral method for the $d = 2$ case in Section 2.6.1, and we also present a range of computational results in order to illustrate the behaviour of the method in practice, as well as to provide experimental verification of the convergence analysis from Section 2.5. In Section 2.6.2, we compare the behaviour of the numerical method based on the backward Euler temporal discretisation with a semi-implicit scheme in which the transport term in (2.6) is treated explicitly in time. The semi-implicit scheme is used in Chapter 3, and the results of Section 2.6.2 have important implications there. Finally, we consider the implementation of the spectral method in three spatial dimensions in Section 2.6.3 and we demonstrate that, as expected, the behaviour of the Galerkin spectral method is essentially the same as in the $d = 2$ case.

2.1 Properties of Maxwellian-weighted spaces

In this section, density results are derived for the Maxwellian-weighted function spaces that were defined above. Since the density results below are not specific to the FENE model, they shall be stated more generally, for any potential U and associated Maxwellian M that satisfy Hypotheses A and B, respectively.

(a) Suppose that the Maxwellian M satisfies Hypothesis B; M is then a weight-function of Type 3 in the sense of Triebel. According to [78], Theorem 3.2.2a, the weighted Sobolev space $\mathrm{H}^1_M(D) = \{v \in \mathrm{L}^2_M(D) : \nabla_q v \in (\mathrm{L}^2_M(D))^d\}$ is a Hilbert space with respect to the norm $\|\cdot\|_{\mathrm{H}^1_M(D)}$ defined by

$$\|v\|_{\mathrm{H}^1_M(D)} := \left(\|v\|^2_{\mathrm{L}^2_M(D)} + \|\nabla_q v\|^2_{\mathrm{L}^2_M(D)}\right)^{\frac{1}{2}},$$

and $\mathrm{L}^2_M(D) = (1/\sqrt{M}\,)\,\mathrm{L}^2(D)$ is a Hilbert space with norm $\|\cdot\|_{\mathrm{L}^2_M(D)}$ defined by $\|v\|_{\mathrm{L}^2_M(D)} := \|\sqrt{M}v\|$, where $\|\cdot\|$ denotes the $\mathrm{L}^2(D)$ norm induced by the $\mathrm{L}^2(D)$ inner product $(\cdot,\cdot)$. By [78], Theorem 3.2.2c, $\mathrm{C}^\infty(\overline{D})$ is dense in both $\mathrm{H}^1_M(D)$ and $\mathrm{L}^2_M(D)$;

see also Ch. I, Sec. 7, in Kufner [52], or one of [13,14]. Thus, since $v \in \mathrm{H}^1_M(D)$ if and only if $\sqrt{M}\,v \in \mathrm{H}^1(D;M)$, it follows that $\sqrt{M}\,\mathrm{C}^\infty(\overline{D})$ is dense in the Hilbert spaces $\mathrm{H}^1(D;M)$ and $\mathrm{L}^2(D)$, whereby $\mathrm{H}^1(D;M)$ is dense in $\mathrm{L}^2(D)$.

(b) Now suppose that U satisfies Hypothesis A and the associated Maxwellian M satisfies Hypothesis B. It follows from Hardy's inequality (see, for example, [4,63]) that

$$\int_D \left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-2} |\hat{\psi}(\underset{\sim}{q})|^2\,\mathrm{d}\underset{\sim}{q} \le 4b\|\underset{\sim}{\nabla}_q\hat{\psi}\|^2 \qquad \forall\hat{\psi}\in\mathrm{H}^1_0(D). \tag{2.7}$$

Since $\underset{\sim}{\nabla}_M\hat{\psi} = \underset{\sim}{\nabla}_q\hat{\psi} + \frac{1}{2}\underset{\sim}{q}U'\left(\frac{1}{2}|\underset{\sim}{q}|^2\right)\hat{\psi}$, Hypothesis A implies that there exists $C_1 \in \mathbb{R}_{>0}$ (for the FENE model $C_1=1$) such that $(1-|\underset{\sim}{q}|^2/b)^2|U'(\frac{1}{2}|\underset{\sim}{q}|^2)|^2 \le C_1^2$ for all $\underset{\sim}{q}\in D$, whereby

$$\|\underset{\sim}{\nabla}_M\hat{\psi}\| \le (1+C_1b)\|\underset{\sim}{\nabla}_q\hat{\psi}\| \qquad \forall\hat{\psi}\in\mathrm{H}^1_0(D). \tag{2.8}$$

Now, (2.8) implies that $\mathrm{H}^1_0(D)\subset\mathrm{H}^1(D;M)$.

Finally, we show that $\mathrm{H}^1(D;M)=\mathrm{H}^1_0(D;M)$. As $\sqrt{M}\mathrm{C}^\infty(\overline{D})\subset\mathrm{H}^1_0(D)\subset\mathrm{H}^1(D;M)$ and $\sqrt{M}\mathrm{C}^\infty(\overline{D})$ is dense in $\mathrm{H}^1(D;M)$ (*cf.* (a) above), we deduce that $\mathrm{H}^1_0(D)$ is dense in $\mathrm{H}^1(D;M)$. Since $\mathrm{C}^\infty_0(D)$ is dense in $\mathrm{H}^1_0(D)$, it follows from (2.8) that $\mathrm{C}^\infty_0(D)$ is also dense in $\mathrm{H}^1(D;M)$. By definition, $\mathrm{H}^1_0(D;M)$ is the closure of $\mathrm{C}^\infty_0(D)$ in $\mathrm{H}^1(D;M)$; thus $\mathrm{H}^1(D;M)=\mathrm{H}^1_0(D;M)$, and therefore also $\mathfrak{K}=\mathfrak{K}_0$. As $\mathrm{H}^1(D;M)$ is continuously and densely embedded into $\mathrm{L}^2(D)$, it follows that $\mathrm{H}^1_0(D;M)$ is continuously and densely embedded into $\mathrm{L}^2(D)$.

Remark 2.3 *A third hypothesis (referred to as Hypothesis C) was introduced in [49], which enabled the inequalities:*

$$\inf_{c\in Ker(\underset{\sim}{\nabla}_M)}\int_D |\hat{\psi}-c\,|^2\,\mathrm{d}\underset{\sim}{q} \le \int_D |\underset{\sim}{\nabla}_M\hat{\psi}|^2\,\mathrm{d}\underset{\sim}{q}, \tag{2.9}$$

and

$$\inf_{c\in Ker(\underset{\sim}{\nabla}_M)}\int_D \frac{|\hat{\psi}-c\,|^2}{1-\frac{|\underset{\sim}{q}|^2}{b}}\,\mathrm{d}\underset{\sim}{q} \le \frac{b}{b-2}\int_D |\underset{\sim}{\nabla}_M\hat{\psi}|^2\,\mathrm{d}\underset{\sim}{q}, \tag{2.10}$$

to be established for all $\hat{\psi}\in\mathrm{H}^1(D;M)$.

2.2 Analysis of the backward Euler semidiscretisation

As noted in the opening of this chapter, by setting $\hat{\psi}(\cdot,t) := \psi(\cdot,t)/\sqrt{M}$ for $t\in[0,T]$ and $\hat{\varphi} := \varphi/\sqrt{M}$ in (2.5) and writing $\hat{\psi}_0 := \psi_0/\sqrt{M}$, the following weak formulation of the initial-boundary-value problem (2.1), (2.2), (2.3) is obtained:

Given $\hat{\psi}_0 \in \mathrm{L}^2(D)$, find $\hat{\psi} \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(D)) \cap \mathrm{L}^2(0,T;\mathrm{H}^1_0(D;M))$ such that (2.6) holds in the sense of distributions on $(0,T)$, and $\hat{\psi}(\cdot,0) = \hat{\psi}_0(\cdot)$.

The function ψ, representing a weak solution to the problem (2.5), is then recovered from $\hat{\psi}$ through the substitution $\psi := \sqrt{M}\,\hat{\psi}$. Thus, instead of constructing a Galerkin approximation to ψ, the aim is to construct a Galerkin approximation to $\hat{\psi}$ from a finite-dimensional subspace of $\mathrm{H}^1_0(D;M)$, from which an approximation to $\hat{\psi}$ can be obtained straightforwardly.

Let $N_T \geq 1$ be an integer, $\Delta t = T/N_T$, and $t^n = n\Delta t$, for $n = 0, 1, \ldots, N_T$. Discretising (2.6) in time using the backward Euler method yields the following semi-discrete numerical scheme.

Given $\hat{\psi}^0 := \hat{\psi}_0 = \psi_0/\sqrt{M} \in \mathrm{L}^2(D)$, find $\hat{\psi}^{n+1} \in \mathrm{H}^1_0(D;M)$, $n = 0, \ldots, N_T - 1$, such that

$$\int_D \frac{\hat{\psi}^{n+1} - \hat{\psi}^n}{\Delta t}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} - \int_D (\underset{\approx}{\kappa}^{n+1}\,\underset{\sim}{q}\,\hat{\psi}^{n+1}) \cdot \underset{\sim}{\nabla}_M \hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}} \int_D \underset{\sim}{\nabla}_M \hat{\psi}^{n+1} \cdot \underset{\sim}{\nabla}_M \hat{\varphi}\,\mathrm{d}\underset{\sim}{q} = 0, \quad (2.11)$$

for all $\hat{\varphi} \in \mathrm{H}^1_0(D;M)$.

Let us first show that for any Δt, sufficiently small, problem (2.11) has a unique solution. To this end, we consider the bilinear form $\mathfrak{B}(\cdot,\cdot)$ defined on $\mathrm{H}^1_0(D;M) \times \mathrm{H}^1_0(D;M)$ by

$$\mathfrak{B}(\hat{\psi},\hat{\varphi}) := \frac{1}{\Delta t}\int_D \hat{\psi}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} - \int_D (\underset{\approx}{\kappa}^{n+1}\,\underset{\sim}{q}\,\hat{\psi}) \cdot \underset{\sim}{\nabla}_M \hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}} \int_D \underset{\sim}{\nabla}_M \hat{\psi} \cdot \underset{\sim}{\nabla}_M \hat{\varphi}\,\mathrm{d}\underset{\sim}{q},$$

and, for $\hat{\psi}^n \in \mathrm{L}^2(D)$ fixed, we define the linear functional $\ell(\hat{\psi}^n;\cdot)$ on $\mathrm{H}^1_0(D;M)$ by

$$\ell(\hat{\psi}^n;\hat{\varphi}) := \frac{1}{\Delta t}\int_D \hat{\psi}^n\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q}.$$

Clearly,

$$\mathfrak{B}(\hat{\psi},\hat{\psi}) \geq \frac{1}{\Delta t}\left(1 - \Delta t \mathrm{Wi} b \|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)}\right) \int_D |\hat{\psi}|^2\,\mathrm{d}\underset{\sim}{q} + \frac{1}{4\mathrm{Wi}}\int_D |\underset{\sim}{\nabla}_M \hat{\psi}|^2\,\mathrm{d}\underset{\sim}{q},$$

and hence, on assuming that $\Delta t \mathrm{Wi} b \|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)} < 1$ and letting $c_{\Delta t} := \frac{1}{\Delta t}\left(1 - \Delta t \mathrm{Wi} b \|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)}\right)$, we deduce that

$$\mathfrak{B}(\hat{\psi},\hat{\psi}) \geq \min\left(c_{\Delta t}, \frac{1}{4\mathrm{Wi}}\right) \|\hat{\psi}\|^2_{\mathrm{H}^1_0(D;M)}. \quad (2.12)$$

Also, by a simple application of the Cauchy–Schwarz inequality, $\mathfrak{B}(\cdot,\cdot)$ is a bounded bilinear functional on $\mathrm{H}^1_0(D;M) \times \mathrm{H}^1_0(D;M)$ and, for any $\hat{\psi}^n \in \mathrm{L}^2(D)$, $\ell(\hat{\psi}^n;\cdot)$ is a bounded linear functional on $\mathrm{H}^1_0(D;M)$. Since $\mathrm{H}^1_0(D;M)$ is a Hilbert space with

norm $\|\cdot\|_{\mathrm{H}^1_0(D;M)}$, the Lax–Milgram theorem implies the existence of a unique solution $\hat{\psi}^{n+1} \in \mathrm{H}^1_0(D;M)$ such that

$$\mathfrak{B}(\hat{\psi}^{n+1},\hat{\varphi}) = \ell(\hat{\psi}^n;\hat{\varphi}) \qquad \forall \hat{\varphi} \in \mathrm{H}^1_0(D;M), \qquad n = 0,1,\ldots,N_T-1. \tag{2.13}$$

As $\hat{\psi}^0 \in \mathrm{L}^2(D)$, we have thus shown that, for any $\Delta t = T/N_T$ such that $\Delta t \mathrm{Wi}\, b \|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)} < 1$, the problem (2.11) has a unique solution $\{\hat{\psi}^n \in \mathrm{H}^1_0(D;M) \,:\, n = 1,\ldots,N_T\}$.

For the purposes of the convergence analysis that will be carried out below, we consider an extended version of the scheme (2.11) with a nonzero right-hand side:

$$\begin{aligned}\int_D \frac{\hat{\psi}^{n+1}-\hat{\psi}^n}{\Delta t}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} - \int_D (\underset{\approx}{\kappa}^{n+1}\,\underset{\sim}{q}\,\hat{\psi}^{n+1})\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\,\mathrm{d}\underset{\sim}{q}\\ = \int_D \mu^{n+1}\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \int_D \underset{\sim}{\nu}^{n+1}\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} \qquad \forall\hat{\varphi}\in\mathrm{H}^1_0(D;M),\end{aligned} \tag{2.14}$$

for $n = 0,\ldots,N_T-1$, where $\mu^{n+1} \in \mathrm{L}^2(D)$ and $\underset{\sim}{\nu}^{n+1} \in (\mathrm{L}^2(D))^d$ for all $n \geq 0$. We have the following stability result for (2.14).

Lemma 2.4 (The first stability inequality) *Let $\Delta t = T/N_T$, $N_T \geq 1$, $\underset{\approx}{\kappa} \in (\mathrm{C}[0,T])^{d\times d}$, $\hat{\psi}^0 \in \mathrm{L}^2(D)$, and define $c_0 := 1 + 4\mathrm{Wi} b\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)}$. If Δt is such that $0 < c_0\Delta t \leq 1/2$, then we have, for all m such that $1 \leq m \leq N_T$,*

$$\begin{aligned}&\|\hat{\psi}^m\|^2 + \sum_{n=0}^{m-1}\Delta t\left\|\frac{\hat{\psi}^{n+1}-\hat{\psi}^n}{\sqrt{\Delta t}}\right\|^2 + \sum_{n=0}^{m-1}\frac{\Delta t}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2\\ &\qquad\leq \mathrm{e}^{2c_0 m\Delta t}\left\{\|\hat{\psi}^0\|^2 + \sum_{n=0}^{m-1}2\Delta t\left(\|\mu^{n+1}\|^2 + 4\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2\right)\right\}.\end{aligned}$$

Proof. Let $0 \leq n \leq N_T - 1$. Setting $\hat{\varphi} = \hat{\psi}^{n+1}$, we write the first term in (2.14) as

$$\int_D \frac{\hat{\psi}^{n+1}-\hat{\psi}^n}{\Delta t}\,\hat{\psi}^{n+1}\,\mathrm{d}\underset{\sim}{q} = \frac{1}{2\Delta t}\left(\|\hat{\psi}^{n+1}\|^2 - \|\hat{\psi}^n\|^2\right) + \frac{1}{2\Delta t}\|\hat{\psi}^{n+1}-\hat{\psi}^n\|^2$$

using the identity $(\alpha-\beta)\alpha = \frac{1}{2}(\alpha^2-\beta^2) + \frac{1}{2}(\alpha-\beta)^2$.

Applying the Cauchy–Schwarz inequality to the transport term in (2.14), we have

$$\int_D (\underset{\approx}{\kappa}^{n+1}\,\underset{\sim}{q}\,\hat{\psi}^{n+1})\cdot\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\,\mathrm{d}\underset{\sim}{q} \leq \sqrt{b}\,|\underset{\approx}{\kappa}^{n+1}|\,\|\hat{\psi}^{n+1}\|\,\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|.$$

Combining these results and applying the Cauchy–Schwarz inequality to the right-hand side terms in (2.14) gives

$$\begin{aligned}&\|\hat{\psi}^{n+1}\|^2 + \|\hat{\psi}^{n+1}-\hat{\psi}^n\|^2 + \frac{\Delta t}{\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2\\ &\qquad\leq \|\hat{\psi}^n\|^2 + 2\Delta t\sqrt{b}\,|\underset{\approx}{\kappa}^{n+1}|\|\hat{\psi}^{n+1}\|\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|\\ &\qquad\qquad +2\Delta t\|\mu^{n+1}\|\|\hat{\psi}^{n+1}\| + 2\Delta t\|\underset{\sim}{\nu}^{n+1}\|\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|\\ &\qquad=: \|\hat{\psi}^n\|^2 + \mathrm{T}_1 + \mathrm{T}_2 + \mathrm{T}_3.\end{aligned}$$

Using Cauchy's inequality $2\alpha\beta \leq \varepsilon\alpha^2 + \varepsilon^{-1}\beta^2$ with $\varepsilon > 0$ on each of T_1 and T_3, we deduce that

$$\mathrm{T}_1 \leq \varepsilon\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2 + \frac{1}{\varepsilon}\Delta t^2 b|\underset{\approx}{\kappa}^{n+1}|^2\|\hat{\psi}^{n+1}\|^2, \qquad \mathrm{T}_3 \leq \varepsilon\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2 + \frac{1}{\varepsilon}\Delta t^2\|\underset{\sim}{\nu}^{n+1}\|^2.$$

Choosing $\varepsilon = \Delta t/(4\mathrm{Wi})$ then gives

$$\begin{aligned}\|\hat{\psi}^{n+1}\|^2 + \|\hat{\psi}^{n+1} - \hat{\psi}^n\|^2 + \frac{\Delta t}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2 \\ \leq \|\hat{\psi}^n\|^2 + 4\Delta t\mathrm{Wi}b|\underset{\approx}{\kappa}^{n+1}|^2\|\hat{\psi}^{n+1}\|^2 + 4\Delta t\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2 + \mathrm{T}_2.\end{aligned}$$

Similarly, we have $\mathrm{T}_2 \leq \Delta t\|\hat{\psi}^{n+1}\|^2 + \Delta t\|\mu^{n+1}\|^2$, and therefore, on defining $c_0 := 1 + 4\mathrm{Wi}b\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)}$, we get

$$\begin{aligned}(1 - c_0\Delta t)\|\hat{\psi}^{n+1}\|^2 + \|\hat{\psi}^{n+1} - \hat{\psi}^n\|^2 + \frac{\Delta t}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2 \\ \leq \|\hat{\psi}^n\|^2 + \Delta t\|\mu^{n+1}\|^2 + 4\Delta t\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2.\end{aligned}$$

As $c_0\Delta t \leq \frac{1}{2}$, dividing through by $(1 - c_0\Delta t)$ and using the fact that $1 \leq \frac{1}{1-c_0\Delta t} \leq 2$, we have

$$\begin{aligned}&\|\hat{\psi}^{n+1}\|^2 + \|\hat{\psi}^{n+1} - \hat{\psi}^n\|^2 + \frac{\Delta t}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2 \\ &\quad\leq \frac{1}{1 - c_0\Delta t}\left(\|\hat{\psi}^n\|^2 + \Delta t\|\mu^{n+1}\|^2 + 4\Delta t\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2\right) \\ &\quad\leq (1 + 2c_0\Delta t)\|\hat{\psi}^n\|^2 + 2\Delta t\left(\|\mu^{n+1}\|^2 + 4\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2\right). \qquad (2.15)\end{aligned}$$

Summing over $n = 0, \ldots, m-1$ in (2.15) we obtain

$$\begin{aligned}&\|\hat{\psi}^m\|^2 + \sum_{n=0}^{m-1}\Delta t\left\|\frac{\hat{\psi}^{n+1} - \hat{\psi}^n}{\sqrt{\Delta t}}\right\|^2 + \sum_{n=0}^{m-1}\frac{\Delta t}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2 \\ &\quad\leq \left\{\|\hat{\psi}^0\|^2 + \sum_{n=0}^{m-1}2\Delta t\left(\|\mu^{n+1}\|^2 + 4\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2\right)\right\} + 2c_0\sum_{n=0}^{m-1}\Delta t\|\hat{\psi}^n\|^2, \qquad (2.16)\end{aligned}$$

for all $m \in \{1, \ldots, N_T\}$. By induction (or by a discrete Gronwall lemma) we deduce that

$$\begin{aligned}&\|\hat{\psi}^m\|^2 + \sum_{n=0}^{m-1}\Delta t\left\|\frac{\hat{\psi}^{n+1} - \hat{\psi}^n}{\sqrt{\Delta t}}\right\|^2 + \sum_{n=0}^{m-1}\frac{\Delta t}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^{n+1}\|^2 \\ &\quad\leq \mathrm{e}^{2c_0 m\Delta t}\left\{\|\hat{\psi}^0\|^2 + \sum_{n=0}^{m-1}2\Delta t\left(\|\mu^{n+1}\|^2 + 4\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2\right)\right\}, \qquad 1 \leq m \leq N_T,\end{aligned}$$

and that completes the proof. $\square$

Theorem 2.5 *Suppose that* $\hat{\psi}_0 \in \mathrm{L}^2(D)$ *and that* $\underset{\approx}{\kappa} \in (\mathrm{C}[0,T])^{d\times d}$. *Then, there exists a unique function* $\hat{\psi}$ *in* $\mathrm{L}^\infty(0,T;\mathrm{L}^2(D)) \cap \mathrm{L}^2(0,T;\mathrm{H}^1_0(D;M)) \cap \mathrm{C}([0,T];\mathrm{L}^2(D))$, *such that*

$$(\hat{\psi}(\cdot,0) - \hat{\psi}_0, \hat{w}) = 0 \qquad \forall \hat{w} \in \mathrm{L}^2(D)$$

and

$$-(\hat{\psi}_0, \hat{\varphi}(\cdot,0)) - \int_0^T \int_D \hat{\psi} \frac{\partial \hat{\varphi}}{\partial t}\, \mathrm{d}\underset{\sim}{q}\, \mathrm{d}t - \int_0^T \int_D (\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \hat{\psi}) \cdot \underset{\sim}{\nabla}_M \hat{\varphi}\, \mathrm{d}\underset{\sim}{q}\, \mathrm{d}t \tag{2.17}$$
$$+\frac{1}{2\mathrm{Wi}} \int_0^T \int_D \underset{\sim}{\nabla}_M \hat{\psi} \cdot \underset{\sim}{\nabla}_M \hat{\varphi}\, \mathrm{d}\underset{\sim}{q}\, \mathrm{d}t = 0, \quad \forall \hat{\varphi} \in \mathrm{H}^1(0,T;\mathrm{H}^1_0(D;M)), \quad \hat{\varphi}(\cdot,T) = 0.$$

The function $\psi = \sqrt{M}\hat{\psi}$ *will be called the weak solution of the initial-boundary-value problem* (2.1), (2.2), (2.3).

Proof. This theorem is proved in Section 3 of [49], the interested reader is referred to that paper for details. The argument makes use of the stability result in Lemma 2.4 in order to use compactness results for the bounded sequence of solutions to (2.11) as $\Delta t \to 0_+$. $\square$

In the next lemma, a configuration space analogue of Lemma 1.3 is established and also it is shown that a weak form of (1.31) is preserved on D. In the remark below, a result is stated that is necessary for the proof of Lemma 2.7.

Remark 2.6 *Suppose* $\hat{\varphi} \in \mathrm{H}^1_0(D;M)$ *and* $L \geq 0$, *and let* $[\hat{\psi}^n]_-$ *be the pointwise negative part of* $\hat{\psi}^n$, i.e. $[x]_\pm := (x \pm |x|)/2$ *for* $x \in \mathbb{R}$. *Then, it is shown in Lemma 3.5 of [49] that*

$$\underset{\sim}{\nabla}_M [\,\hat{\varphi} - L\sqrt{M}\,]_+ = \begin{cases} \underset{\sim}{\nabla}_M (\,\hat{\varphi} - L\sqrt{M}\,) = \underset{\sim}{\nabla}_M \hat{\varphi} & \text{if } \hat{\varphi} > L\sqrt{M}, \\ 0 & \text{if } \hat{\varphi} \leq L\sqrt{M}; \end{cases} \tag{2.18}$$

and

$$\underset{\sim}{\nabla}_M [\,\hat{\varphi} - L\sqrt{M}\,]_- = \begin{cases} \underset{\sim}{\nabla}_M (\,\hat{\varphi} - L\sqrt{M}\,) = \underset{\sim}{\nabla}_M \hat{\varphi} & \text{if } \hat{\varphi} < L\sqrt{M}, \\ 0 & \text{if } \hat{\varphi} \geq L\sqrt{M}; \end{cases} \tag{2.19}$$

i.e. *that the* $[\cdot]_\pm$ *operators act on functions in* $\mathrm{H}^1_0(D;M)$ *as one would expect. Moreover,* $[\,\hat{\varphi} - L\sqrt{M}\,]_+$ *and* $[\,\hat{\varphi} - L\sqrt{M}\,]_-$ *belong to* $\mathrm{H}^1_0(D;M)$. *The proof of these results is rather technical and is therefore omitted here.*

Lemma 2.7 *Let* $\psi_0 \in \mathfrak{H}$ *and* $\psi = \sqrt{M}\hat{\psi}$ *where* $\hat{\psi} \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(D)) \cap \mathrm{L}^2(0,T;\mathrm{H}^1_0(D;M)) \cap \mathrm{C}([0,T];\mathrm{L}^2(D))$ *is the weak solution to* (2.17) *subject to the initial condition* $\hat{\psi}_0 =$

$\psi_0/\sqrt{M}$ (i.e., *the function ψ is the weak solution of the initial-boundary-value problem* (2.1), (2.2), (2.3)). *Then,*

$$\int_D \psi(\underset{\sim}{q},t)\,\mathrm{d}\underset{\sim}{q} = \int_D \psi_0(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q} \qquad \forall t \in [0,T).$$

Furthermore if $\psi_0 \geq 0$ a.e. on D, then $\psi(\cdot,t) \geq 0$ a.e. on D for all $t \in [0,T]$.

Proof. Fix any $t \in (0,T)$, and let $\varepsilon \in (0, T-t]$. Consider the function $\hat{\varphi}_\varepsilon$ defined by

$$\hat{\varphi}_\varepsilon(\underset{\sim}{q},s) := \begin{cases} \sqrt{M} & \text{for } s \in [0,t], \\ \sqrt{M}(t+\varepsilon-s)/\varepsilon & \text{for } s \in [t,t+\varepsilon), \\ 0 & \text{for } s \in [t+\varepsilon,T]. \end{cases}$$

Clearly, $\hat{\varphi}_\varepsilon \in \mathrm{H}^1(0,T;\mathrm{H}^1_0(D;M))$ and $\hat{\varphi}_\varepsilon(\cdot,T) = 0$. Taking $\hat{\varphi}_\varepsilon$ as test function in (2.17) yields

$$-(\hat{\psi}_0, \sqrt{M}\,) + \frac{1}{\varepsilon}\int_t^{t+\varepsilon} (\hat{\psi}(\cdot,s), \sqrt{M}\,)\,\mathrm{d}s = 0.$$

Passing to the limit $\varepsilon \to 0_+$ yields $-(\hat{\psi}_0, \sqrt{M}\,)+(\hat{\psi}(\cdot,t), \sqrt{M}\,) = 0$, whereby $(\psi(\cdot,t),1) = (\psi_0,1)$, as required, for all $t \in (0,T)$; for $t = 0$ the equality holds trivially.

Now, suppose that $\psi_0 \in \mathfrak{H}$ and $\psi_0 \geq 0$; then, $\hat{\psi}_0 \in \mathrm{L}^2(D)$ and $\hat{\psi}_0 \geq 0$. For Δt as in Lemma 2.4, consider the sequence of functions $(\hat{\psi}^n)_{n=0}^{N_T} \subset \mathrm{H}^1_0(D;M)$ defined by (2.13). Let $[\hat{\psi}^n]_-$ be the pointwise negative part of $\hat{\psi}^n$, where $[x]_\pm := (x \pm |x|)/2$ for $x \in \mathbb{R}$. Then, by Remark 2.6, $([\hat{\psi}^n]_-)_{n=0}^{N_T} \subset \mathrm{H}^1_0(D;M)$. It follows that

$$\mathfrak{B}([\hat{\psi}^{n+1}]_-\,,\,[\hat{\psi}^{n+1}]_-) = \mathfrak{B}(\hat{\psi}^{n+1}\,,\,[\hat{\psi}^{n+1}]_-) = \ell(\hat{\psi}^n;[\hat{\psi}^{n+1}]_-),$$

where the first equality is due to the fact that $[\hat{\psi}^{n+1}]_-$ vanishes when $\hat{\psi}^{n+1} > 0$, and the second equality is due to (2.13). Suppose, for induction, that $\hat{\psi}^n \geq 0$; this is certainly true for $n = 0$, since $\hat{\psi}^0 = \hat{\psi}_0 \geq 0$. Hence,

$$\ell(\hat{\psi}^n;[\hat{\psi}^{n+1}]_-) = \frac{1}{\Delta t}\int_D \hat{\psi}^n(\underset{\sim}{q})[\hat{\psi}^{n+1}(\underset{\sim}{q})]_-\,\mathrm{d}\underset{\sim}{q} \leq 0.$$

Therefore, $\mathfrak{B}([\hat{\psi}^{n+1}]_-\,,\,[\hat{\psi}^{n+1}]_-) \leq 0$; thus, (2.12) implies that $\|[\hat{\psi}^{n+1}]_-\|_{\mathrm{H}^1_0(D;M)} \leq 0$, whereby $[\hat{\psi}^{n+1}]_- = 0$ and hence $\hat{\psi}^{n+1} \geq 0$. By induction, $\hat{\psi}^n \geq 0$ for all $n = 0,1,\ldots,N_T$. Then, passing to the limit $\Delta t \to 0_+$, it follows from Theorem 2.5 that the weak solution $\hat{\psi}$ is non-negative on $D \times [0,T]$ (see [49]). $\square$

Remark 2.8 *The same argument used above to establish the non-negativity of $\hat{\psi}$ can be used to derive a weak maximum principle in the case that $\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}(t)\,\underset{\sim}{q} \leq 0$ for a.e. $t \in [0,T]$ and $\underset{\sim}{q} \in D$.*

Let

$$L = ess.sup_{\underset{\sim}{q}\in D}\ \hat{\psi}_0(\underset{\sim}{q})/\sqrt{M(\underset{\sim}{q})},$$

where it assumed that the essential supremum above is finite. Suppose that $\hat{\psi}^n \le L\sqrt{M}$*; this is certainly true for* $n = 0$*. Then, following the argument above:*

$$\begin{aligned}
&\mathfrak{B}([\hat{\psi}^{n+1} - L\sqrt{M}\,]_+\,,\,[\hat{\psi}^{n+1} - L\sqrt{M}\,]_+) = \mathfrak{B}(\hat{\psi}^{n+1} - L\sqrt{M}\,,\,[\hat{\psi}^{n+1} - L\sqrt{M}\,]_+)\\
&\quad = \mathfrak{B}(\hat{\psi}^{n+1}\,,\,[\hat{\psi}^{n+1} - L\sqrt{M}\,]_+) - L\mathfrak{B}(\sqrt{M}\,,\,[\hat{\psi}^{n+1} - L\sqrt{M}\,]_+)\\
&\quad = \ell(\hat{\psi}^n; [\hat{\psi}^{n+1} - L\sqrt{M}\,]_+) - L\mathfrak{B}(\sqrt{M}\,,\,[\hat{\psi}^{n+1} - L\sqrt{M}\,]_+)\\
&\quad = \frac{1}{\Delta t}\int_D (\hat{\psi}^n(\underset{\sim}{q}) - L\sqrt{M})[\hat{\psi}^{n+1} - L\sqrt{M}\,]_+\,\mathrm{d}\underset{\sim}{q}\\
&\qquad + L\,\mathrm{Wi}\int_D (\underset{\approx}{\kappa}\underset{\sim}{q}\sqrt{M})\cdot\underset{\sim}{\nabla}_M[\hat{\psi}^{n+1} - L\sqrt{M}]_+\,\mathrm{d}\underset{\sim}{q},
\end{aligned}$$

where the diffusion term in $\mathfrak{B}(\cdot,\cdot)$ *vanishes because* $\sqrt{M} \in \ker(\underset{\sim}{\nabla}_M)$*. The term on the second-last line above is non-positive by the inductive hypothesis and, after integrating by parts, we deduce that the term on the last line is also non-positive when* $\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}\,\underset{\sim}{q} \le 0$.[1] *Therefore,* $[\hat{\psi}^{n+1} - L\sqrt{M}\,]_+ = 0$*;* i.e.*,* $\hat{\psi}^{n+1} \le L\sqrt{M}$*. Then, in the same way as in Lemma 2.7, on passage to the limit* $\Delta t \to 0_+$*, this implies that*

$$ess.sup_{(\underset{\sim}{q},t)\in D\times[0,T]}\ \psi(\underset{\sim}{q},t)/M(\underset{\sim}{q}) \le ess.sup_{\underset{\sim}{q}\in D}\ \psi_0(\underset{\sim}{q})/M(\underset{\sim}{q}),$$

which can be thought of as a maximum principle for the initial-boundary value problem in the case that $\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}\,\underset{\sim}{q} \le 0$*.* $\diamond$

By the next lemma, if $\underset{\approx}{\kappa} \in (\mathrm{H}^1(0,T))^{d\times d}$ and $\hat{\psi}_0 \in \mathrm{H}^1_0(D;M)$, then stability can be established in stronger norms than in Lemma 2.4.

Lemma 2.9 (The second stability inequality) *Let* $\Delta t = T/N_T$*,* $N_T \ge 1$*,* $\underset{\approx}{\kappa} \in (\mathrm{H}^1(0,T))^{d\times d}$*,* $\hat{\psi}^0 \in \mathrm{H}^1_0(D;M)$*, and define* $c_0 := 1 + 4\mathrm{Wi}b\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)}$*. If* Δt *is such that* $0 < c_0\Delta t \le 1/2$*, then, for all* m *such that* $1 \le m \le N_T$*,*

$$\begin{aligned}
&\Delta t\sum_{n=0}^{m-1}\left\|\frac{\hat{\psi}^{n+1} - \hat{\psi}^n}{\Delta t}\right\|^2 + \frac{1}{4\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^m\|^2 + \frac{1}{2\mathrm{Wi}}\sum_{n=0}^{m-1}\Delta t\left\|\underset{\sim}{\nabla}_M\frac{\hat{\psi}^{n+1} - \hat{\psi}^n}{\sqrt{\Delta t}}\right\|^2\\
&\qquad \le \mathrm{e}^{2c_1 m\Delta t}\left\{2\Delta t\sum_{n=0}^{m-1}\|\mu^{n+1}\|^2 + 12\mathrm{Wi}\max_{1\le n\le m}\|\underset{\sim}{\nu}^n\|^2 + \Delta t\sum_{n=1}^{m-1}\left\|\frac{\underset{\sim}{\nu}^{n+1} - \underset{\sim}{\nu}^n}{\Delta t}\right\|^2\right.\\
&\qquad\qquad \left. + \frac{1}{\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\hat{\psi}^0\|^2 + \left(b\,\|\underset{\approx}{\kappa}t\|^2_{\mathrm{L}^2(0,T)} + 12\mathrm{Wi}b\,\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)}\right)\mathfrak{S}(\hat{\psi}^0,\mu,\underset{\sim}{\nu},\mathrm{Wi},m\Delta t)\right\},
\end{aligned}$$

[1] In fact, if $\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}(t)\,\underset{\sim}{q} \le 0$ for all $\underset{\sim}{q} \in \mathbb{R}^d$ and $t \in [0,T]$, and $\mathrm{tr}(\underset{\approx}{\kappa}(t)) = 0$ for all $t \in [0,T]$, then it must be the case that $\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}(t)\underset{\sim}{q} = 0$ for all $\underset{\sim}{q} \in \mathbb{R}^d$ and $t \in [0,T]$.

where $\mathfrak{S}(\hat{\psi}^0, \mu, \underset{\sim}{\nu}, \mathrm{Wi}, m\Delta t)$ *is the right-hand side of the inequality from Lemma 2.4 and* $c_1 = 4\mathrm{Wi}(1 + b\,\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)})$.

Proof. The proof is similar to that of Lemma 2.4, except one uses the test function $\hat{\varphi} = (\hat{\psi}^{n+1} - \hat{\psi}^n)/\Delta t$. $\square$

It follows from Lemma 2.9, by an identical argument as in the proof of Theorem 2.5, that the weak solution $\hat{\psi}$ of (2.17) belongs to $\mathrm{H}^1(0,T;\mathrm{L}^2(D)) \cap \mathrm{L}^\infty(0,T;\mathrm{H}^1_0(D;M))$, provided that $\underset{\approx}{\kappa} \in (\mathrm{H}^1(0,T))^{d\times d}$ and $\hat{\psi}_0 \in \mathrm{H}^1_0(D;M)$.

The stability result in Lemma 2.4 will be useful in Section 2.3, but for now, note that setting $\mu = 0$ and $\underset{\sim}{\nu} = \underset{\sim}{0}$ in Lemmas 2.4 and 2.9 demonstrates the unconditional stability of the time semidiscretisation in various norms. Also note that, evidently, any fully-discrete method based on the semidiscrete scheme (2.11) and conforming Galerkin discretisation in $\underset{\sim}{q}$ using a finite-dimensional subspace $\mathcal{P}_N(D)$ of $\mathrm{H}^1_0(D;M)$ will be unconditionally stable in the norms appearing on the left-hand sides of the bounds in Lemmas 2.4 and 2.9.

2.2.1 Well-posedness of a Chauvière–Lozinski type transformed FENE model

In this section we show that, in the case of the FENE model, the weak formulation resulting from the substitution $\hat{\psi} := \psi/M^{2s/b}$ with $b \geq 4s^2/(2s-1)$ and $s > 1/2$ also leads to a well-posed problem and a stable semidiscretisation in any number of space dimensions. The minimum value of the function $s \in (0,\infty) \mapsto 4s^2/(2s-1)$ is attained at $s = 1$, yielding the maximum range of b values, $b \geq 4$. This transformation was proposed by Chauvière & Lozinski [59,24,23,60] in the special cases $s = 2$ and $s = 2.5$, where these values were chosen on the basis of numerical experiments in two and three space dimensions, respectively. For the sake of brevity, we shall confine ourselves to establishing an energy estimate analogous to our first stability inequality in Lemma 2.4, and the discussion in this section is restricted to the FENE model.

Inserting $\psi(\underset{\sim}{q}) = [M(\underset{\sim}{q})]^{2s/b}\hat{\psi}(\underset{\sim}{q})$ into our model problem (2.1), where now M is the

FENE Maxwellian, yields, on noting that $\mathrm{tr}(\underset{\approx}{\kappa})(t) = 0$ for all $t \in [0,T]$,

$$\frac{\partial\hat{\psi}}{\partial t} - \frac{1}{2\mathrm{Wi}}\Delta_q\hat{\psi} = \frac{1}{2\mathrm{Wi}}\left[\left(1-\frac{4s}{b}\right)\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-1}\underset{\sim}{q} - 2\mathrm{Wi}(\underset{\approx}{\kappa}\,\underset{\sim}{q})\right]\cdot\underset{\sim}{\nabla}_q\hat{\psi}$$
$$+\frac{1}{2\mathrm{Wi}}\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-2}\left[d\left(1-\frac{2s}{b}\right)\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)+\right.$$
$$\left.\frac{2(s-1)(2s-b)}{b^2}|\underset{\sim}{q}|^2 + \frac{4s\mathrm{Wi}}{b}(\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}\,\underset{\sim}{q})\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)\right]\hat{\psi}. \tag{2.20}$$

Denoting by $\underset{\sim}{A}(\underset{\sim}{q},t)$ the expression in the first square bracket on the right-hand side of (2.20) and by $B(\underset{\sim}{q},t)$ the expression in the second square bracket, multiplying (2.20) by any $\hat{\varphi} \in \mathrm{H}^1_0(D)$, integrating the resulting expression over D, and integrating by parts in the second term on the left-hand side, yields the following weak formulation.

Given $\hat{\psi}_0 = \psi_0/M^{2s/b} \in \mathrm{L}^2(D)$, find $\hat{\psi} \in \mathrm{L}^\infty(0,T;\mathrm{L}^2(D)) \cap \mathrm{L}^2(0,T;\mathrm{H}^1_0(D))$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_D \hat{\psi}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_q\hat{\psi}\cdot\underset{\sim}{\nabla}_q\hat{\varphi}\,\mathrm{d}\underset{\sim}{q}$$
$$= \frac{1}{2\mathrm{Wi}}\int_D (\underset{\sim}{A}(\underset{\sim}{q},t)\cdot\underset{\sim}{\nabla}_q\hat{\psi})\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-2} B(\underset{\sim}{q},t)\,\hat{\psi}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q}, \tag{2.21}$$

for all $\hat{\varphi} \in \mathrm{H}^1_0(D)$, in the sense of distributions on $(0,T)$, and with $\hat{\psi}(\cdot,0) = \hat{\psi}_0$.

The backward Euler semidiscretisation of this weak formulation is as follows.

Given $\hat{\psi}^0 := \hat{\psi}_0 = \psi_0/M^{2s/b} \in \mathrm{L}^2(D)$, find $\hat{\psi}^{n+1} \in \mathrm{H}^1_0(D)$, $n = 0,1,\ldots,N_T-1$, such that

$$\int_D \frac{\hat{\psi}^{n+1}-\hat{\psi}^n}{\Delta t}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_q\hat{\psi}^{n+1}\cdot\underset{\sim}{\nabla}_q\hat{\varphi}\,\mathrm{d}\underset{\sim}{q}$$
$$= \frac{1}{2\mathrm{Wi}}\int_D (\underset{\sim}{A}(\underset{\sim}{q},t^{n+1})\cdot\underset{\sim}{\nabla}_q\hat{\psi}^{n+1})\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q}$$
$$+\frac{1}{2\mathrm{Wi}}\int_D \left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-2} B(\underset{\sim}{q},t^{n+1})\,\hat{\psi}^{n+1}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q}, \tag{2.22}$$

for all $\hat{\varphi} \in \mathrm{H}^1_0(D)$.

We begin by showing that, for Δt sufficiently small and all $b \geq 4s^2/(2s-1)$ and $s > 1/2$, this problem has a unique solution. To this end, for $t \in [0,T]$ fixed, we

consider the bilinear form defined on $\mathrm{H}^1_0(D)\times\mathrm{H}^1_0(D)$ by

$$\mathfrak{C}(\hat\psi,\hat\varphi) := \frac{1}{\Delta t}\int_D \hat\psi\,\hat\varphi\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_q\hat\psi\cdot\underset{\sim}{\nabla}_q\hat\varphi\,\mathrm{d}\underset{\sim}{q} \\ -\frac{1}{2\mathrm{Wi}}\int_D (\underset{\sim}{A}(\underset{\sim}{q},t)\cdot\underset{\sim}{\nabla}_q\hat\psi)\,\hat\varphi\,\mathrm{d}\underset{\sim}{q} - \frac{1}{2\mathrm{Wi}}\int_D \left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-2} B(\underset{\sim}{q},t)\hat\psi\,\hat\varphi\,\mathrm{d}\underset{\sim}{q}.$$

Now, taking $\hat\varphi=\hat\psi\in\mathrm{C}^\infty_0(D)$, integration by parts in the third integral in the definition of $\mathfrak{C}$, and then merging the resulting integral with the fourth integral in the definition of $\mathfrak{C}$, yields

$$\begin{aligned}\mathfrak{C}(\hat\psi,\hat\psi) &= \frac{1}{\Delta t}\|\hat\psi\|^2+\frac{1}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_q\hat\psi\|^2+\frac{1}{2\mathrm{Wi}}\left(2s-1-\frac{4s^2}{b}\right)\int_D\frac{|\underset{\sim}{q}|^2}{b}\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-2}|\hat\psi|^2\,\mathrm{d}q\\ &\quad -\frac{1}{4\mathrm{Wi}}\int_D\left[d+\frac{8s\mathrm{Wi}}{b}(\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}\,\underset{\sim}{q})\right]\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-1}|\hat\psi|^2\,\mathrm{d}\underset{\sim}{q}.\end{aligned}$$

Assuming that $b\geq 4s^2/(2s-1)$ with $s>1/2$, and recalling that $|\underset{\sim}{q}|<\sqrt{b}$ for $\underset{\sim}{q}\in D$, we then have that

$$\mathfrak{C}(\hat\psi,\hat\psi)\geq\frac{1}{\Delta t}\|\hat\psi\|^2+\frac{1}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_q\hat\psi\|^2-\frac{1}{4\mathrm{Wi}}(d+8s\mathrm{Wi}\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)})\int_D\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-1}|\hat\psi|^2\,\mathrm{d}\underset{\sim}{q}.$$

Let us note that for, any $\beta>0$,

$$\int_D\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-1}|\hat\psi|^2\,\mathrm{d}\underset{\sim}{q}\leq\frac{1}{4\beta}\int_D|\hat\psi|^2\,\mathrm{d}\underset{\sim}{q}+\beta\int_D\left(1-\frac{|\underset{\sim}{q}|^2}{b}\right)^{-2}|\hat\psi|^2\,\mathrm{d}\underset{\sim}{q}. \qquad (2.23)$$

Hence, by (2.7) and fixing β as the unique solution of the equation $4b\left(d+8s\mathrm{Wi}\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)}\right)\beta=1$, we have that

$$\mathfrak{C}(\hat\psi,\hat\psi)\geq\frac{1}{\Delta t}\left(1-\frac{b\Delta t}{4\mathrm{Wi}}(d+8s\mathrm{Wi}\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)})^2\right)\|\psi\|^2+\frac{1}{4\mathrm{Wi}}\|\underset{\sim}{\nabla}_q\hat\psi\|^2\qquad\forall\hat\psi\in\mathrm{C}^\infty_0(D).$$

Recalling that $\mathrm{C}^\infty_0(D)$ is dense in $\mathrm{H}^1_0(D)$ and, by [13] and [14], also in the $(1-|\underset{\sim}{q}|^2/b)^{-2}$-weighted L^2 space, $\mathrm{L}^2_{M^{-4/b}}(D)$, we thus deduce that, for any $\Delta t<4\mathrm{Wi}/(b(d+8s\mathrm{Wi}\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)})^2)$, the bilinear form $\mathfrak{C}$ is coercive on $\mathrm{H}^1_0(D)\times\mathrm{H}^1_0(D)$. The existence of a unique solution $\{\hat\psi^n\}_{n=0}^{N_T}$ to the semidiscretisation (2.22) then follows from the Lax–Milgram theorem, as in the previous section. Using the above coercivity argument, the proof of stability of (2.22), stated in Lemma 2.10 below, is completely analogous to the proof of Lemma 2.4 and is therefore omitted.

Lemma 2.10 (Stability inequality) *Let* $\Delta t = T/N_T$, $N_T \geq 1$, $\underset{\approx}{\kappa} \in (\mathrm{C}[0,T])^{d\times d}$, $\hat{\psi}^0 \in \mathrm{L}^2(D)$, $b \geq 4s^2/(2s-1)$ *with* $s > 1/2$, *and define* $c_0 := b(d+8s\mathrm{Wi}\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)})^2/(2\mathrm{Wi})$. *If* Δt *is such that* $0 < c_0\Delta t \leq 1/2$, *then we have, for all* m *such that* $1 \leq m \leq N_T$,

$$\|\hat{\psi}^m\|^2 + \sum_{n=0}^{m-1} \Delta t \left\| \frac{\hat{\psi}^{n+1} - \hat{\psi}^n}{\sqrt{\Delta t}} \right\|^2 + \sum_{n=0}^{m-1} \frac{\Delta t}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_q \hat{\psi}^{n+1}\|^2 \leq \mathrm{e}^{2c_0 m\Delta t}\|\hat{\psi}^0\|^2.$$

Using Lemma 2.10, the existence of a unique weak solution to (2.21) can be established in the same way as for the symmetrised formulation.

2.3 The fully-discrete method

We now return to the semidiscrete method (2.11) based on the symmetrised version of the Fokker–Planck equation and describe the construction of a fully-discrete numerical method that stems from this semidiscretisation. At the end of the section we shall comment on the extension of our results to a fully-discrete method based on the semidiscretisation (2.22) of the Chauvière–Lozinski-transformed Fokker–Planck equation (2.20) for the FENE model.

Let $\mathcal{P}_N(D)$ be a finite-dimensional subspace of $\mathrm{H}^1_0(D;M)$, to be chosen below, and let $\hat{\psi}^n_N \in \mathcal{P}_N(D)$ be the solution at time level n of our fully-discrete Galerkin method:

$$\int_D \frac{\hat{\psi}^{n+1}_N - \hat{\psi}^n_N}{\Delta t}\,\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} - \int_D (\underset{\approx}{\kappa}^{n+1}\,\underset{\sim}{q}\,\hat{\psi}^{n+1}_N)\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_M\hat{\psi}^{n+1}_N\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\,\mathrm{d}\underset{\sim}{q} = 0$$
$$\forall\hat{\varphi}\in\mathcal{P}_N(D),\quad n=0,\ldots,N_T-1, \qquad (2.24)$$

$$\hat{\psi}^0_N(\cdot) := \text{the } \mathrm{L}^2(D) \text{ orthogonal projection of } \hat{\psi}_0(\cdot) = \hat{\psi}(\cdot,0) \text{ onto } \mathcal{P}_N(D). \qquad (2.25)$$

Remark 2.11 *If the linear space* $\mathcal{P}_N(D)$ *is selected so that* $\sqrt{M} \in \mathcal{P}_N(D)$, *then, since* $\sqrt{M} \in Ker(\underset{\sim}{\nabla}_M)$, *it follows on taking* $\hat{\varphi} = \sqrt{M}$ *in* (2.24) *that*

$$\int_D \sqrt{M(\underset{\sim}{q})}\,\hat{\psi}^n_N(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q} = \int_D \sqrt{M(\underset{\sim}{q})}\,\hat{\psi}^0_N(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q}, \qquad n=1,\ldots,N_T,$$

whereby, on letting $\psi^n_N := \sqrt{M}\hat{\psi}^n_N$, *we have that*

$$\int_D \psi^n_N(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q} = \int_D \psi^0_N(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q}, \qquad n=1,\ldots,N_T.$$

The function ψ^n_N *represents an approximation to the probability density function* $\psi = \sqrt{M}\hat{\psi}$ *at* $t = t^n$. *Since, by Lemma 2.7,* $\int_D \psi(\underset{\sim}{q},t)\,\mathrm{d}\underset{\sim}{q} = \int_D \psi_0(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q} = 1$ *for all* $t \geq 0$, *we deduce, by choosing* $\mathcal{P}_N(D)$ *so that* $\sqrt{M} \in \mathcal{P}_N(D)$, *that this integral identity is preserved under discretisation. The integral* $\int_D \psi(\underset{\sim}{q},t)\,\mathrm{d}\underset{\sim}{q}$ *will sometimes be referred to as the* volume *of* ψ. $\diamond$

Our objective is to derive a bound on the global error $e_N^n := \hat{\psi}(\cdot,t^n) - \hat{\psi}_N^n$. Clearly,

$$e_N^n = (\hat{\psi}(\cdot,t^n) - \hat{\Pi}_N\hat{\psi}(\cdot,t^n)) + (\hat{\Pi}_N\hat{\psi}(\cdot,t^n) - \hat{\psi}_N^n) =: \eta^n + \xi^n,$$

where $\hat{\Pi}_N\hat{\psi}(\cdot,t^n) \in \mathcal{P}_N(D)$ is a certain projection of $\hat{\psi}(\cdot,t^n)$ onto $\mathcal{P}_N(D)$ that will be defined below. For the moment, the specific choices of $\mathcal{P}_N \subset \mathrm{H}_0^1(D;M)$ and $\hat{\Pi}_N$ are irrelevant. Note also that η is defined for a.e. $t \in (0,T)$, *i.e.* not only at the discrete time-levels.

We begin by bounding norms of ξ in terms of suitable norms of η. Substituting ξ into (2.24), setting $\hat{\varphi} = \xi^{n+1}$, and noting that $\xi^n = \hat{\psi}(\cdot,t^n) - \hat{\psi}_N^n - \eta^n$, we have

$$\begin{aligned}\int_D \frac{\xi^{n+1}-\xi^n}{\Delta t}\,\xi^{n+1}\,\mathrm{d}\underset{\sim}{q} - \int_D (\underset{\approx}{\kappa}^{n+1}\,\underset{\sim}{q}\,\xi^{n+1})\cdot\underset{\sim}{\nabla}_M\xi^{n+1}\,\mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_M\xi^{n+1}\cdot\underset{\sim}{\nabla}_M\xi^{n+1}\,\mathrm{d}\underset{\sim}{q}\\ = \int_D \mu^{n+1}\,\xi^{n+1}\,\mathrm{d}\underset{\sim}{q} + \int_D \underset{\sim}{\nu}^{n+1}\cdot\underset{\sim}{\nabla}_M\xi^{n+1}\,\mathrm{d}\underset{\sim}{q},\end{aligned} \tag{2.26}$$

for $n = 0,\ldots,N_T-1$, where

$$\mu^{n+1} := \left(\frac{\hat{\psi}(\cdot,t^{n+1})-\hat{\psi}(\cdot,t^n)}{\Delta t} - \frac{\partial\hat{\psi}}{\partial t}(\cdot,t^{n+1})\right) - \frac{\eta^{n+1}-\eta^n}{\Delta t}, \tag{2.27}$$

$$\underset{\sim}{\nu}^{n+1} := \underset{\approx}{\kappa}^{n+1}\underset{\sim}{q}\eta^{n+1} - \frac{1}{2\mathrm{Wi}}\underset{\sim}{\nabla}_M\eta^{n+1}. \tag{2.28}$$

Since $\mathcal{P}_N(D) \subset \mathrm{H}_0^1(D;M)$, (2.26) is in the form of (2.14); hence, applying Lemma 2.4, we obtain

$$\|\xi^m\|^2 + \frac{1}{2\mathrm{Wi}}\sum_{n=0}^{m-1}\Delta t\|\underset{\sim}{\nabla}_M\xi^{n+1}\|^2 \le \mathrm{e}^{2c_0m\Delta t}\left\{\|\xi^0\|^2 + \sum_{n=0}^{m-1}2\Delta t\left(\|\mu^{n+1}\|^2 + 4\mathrm{Wi}\|\underset{\sim}{\nu}^{n+1}\|^2\right)\right\}, \tag{2.29}$$

for $m = 1,\ldots,N_T$. Let us first consider the term $\|\xi^0\|$ on the right-hand side of (2.29). Since $\hat{\psi}_N^0$ is the $\mathrm{L}^2(D)$ orthogonal projection of $\hat{\psi}(\cdot,0) = \hat{\psi}_0$ onto $\mathcal{P}_N(D)$, we have $(\xi^0,\hat{\varphi}_N) = -(\eta^0,\hat{\varphi}_N)$ for all $\hat{\varphi}_N \in \mathcal{P}_N(D)$. Setting $\hat{\varphi}_N = \xi^0$ here and applying the Cauchy–Schwarz inequality on the right-hand side yields $\|\xi^0\| \le \|\eta^0\|$.

By the triangle inequality we have the following bound on $\|\underset{\sim}{\nu}^{n+1}\|$:

$$\|\underset{\sim}{\nu}^{n+1}\| \le \sqrt{b}\,|\underset{\approx}{\kappa}^{n+1}|\,\|\eta^{n+1}\| + \frac{1}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\eta^{n+1}\|, \qquad n = 0,\ldots,N_T-1.$$

Hence for the third term on the right-hand-side of (2.29), we have

$$\begin{aligned}\sum_{n=0}^{m-1}8\mathrm{Wi}\Delta t\|\underset{\sim}{\nu}^{n+1}\|^2 &\le \sum_{n=0}^{m-1}\Delta t\left(16\mathrm{Wi}b|\underset{\approx}{\kappa}^{n+1}|^2\|\eta^{n+1}\|^2 + \frac{4}{\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\eta^{n+1}\|^2\right)\\ &\le 4c_2\sum_{n=0}^{m-1}\Delta t\|\eta^{n+1}\|^2_{\mathrm{H}_0^1(D;M)} = 4c_2\|\eta\|^2_{\ell^2(0,t^m;\mathrm{H}_0^1(D;M))},\end{aligned}$$

for $m = 1, \ldots, N_T$, where $c_2 := \max\left(1/\mathrm{Wi}\,,\, 4\mathrm{Wi}b\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T)}\right)$.

It remains to bound $\|\mu^{m+1}\|$. We begin by observing that

$$\|\mu^{m+1}\| \leq \left\| \frac{\hat{\psi}(\cdot,t^{n+1}) - \hat{\psi}(\cdot,t^n)}{\Delta t} - \frac{\partial \hat{\psi}}{\partial t}(\cdot,t^{n+1}) \right\| + \left\| \frac{\eta^{n+1}-\eta^n}{\Delta t} \right\| =: I + II.$$

Bounding both I and II by Taylor's theorem with integral remainder yields

$$\begin{aligned} I^2 &\leq \Delta t \int_{t^n}^{t^{n+1}} \left\| \frac{\partial^2 \hat{\psi}}{\partial t^2}(\cdot,t) \right\|^2 \mathrm{d}t, \\ II^2 &\leq \int_D \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \left| \frac{\partial \eta}{\partial t}(\underset{\sim}{q},t) \right|^2 \mathrm{d}t \; \mathrm{d}\underset{\sim}{q} = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \left\| \frac{\partial \eta}{\partial t}(\cdot,t) \right\|^2 \mathrm{d}t. \end{aligned}$$

Therefore, we now have that

$$\begin{aligned} \sum_{n=0}^{m-1} 2\Delta t \|\mu^{n+1}\|^2 &\leq 4\sum_{n=0}^{m-1} \Delta t^2 \int_{t^n}^{t^{n+1}} \left\| \frac{\partial^2 \hat{\psi}}{\partial t^2}(\cdot,t) \right\|^2 \mathrm{d}t + 4\sum_{n=0}^{m-1} \int_{t^n}^{t^{n+1}} \left\| \frac{\partial \eta}{\partial t}(\cdot,t) \right\|^2 \mathrm{d}t \\ &= 4\Delta t^2 \left\| \frac{\partial^2 \hat{\psi}}{\partial t^2} \right\|^2_{\mathrm{L}^2(0,t^m;\mathrm{L}^2(D))} + 4 \left\| \frac{\partial \eta}{\partial t} \right\|^2_{\mathrm{L}^2(0,t^m;\mathrm{L}^2(D))}. \end{aligned}$$

Combining the bounds on the three terms on the right-hand side of (2.29) we deduce that

$$\begin{aligned} &\|\xi^m\|^2 + \frac{1}{2\mathrm{Wi}} \sum_{n=0}^{m-1} \Delta t \|\underset{\sim}{\nabla}_M \xi^{n+1}\|^2 \\ &\quad \leq \mathrm{e}^{2c_0 m\Delta t} \left(\|\eta^0\|^2 + 4c_2 \|\eta\|^2_{\ell^2(0,t^m;\mathrm{H}^1_0(D;M))} \right. \\ &\qquad \left. + 4 \left\| \frac{\partial \eta}{\partial t} \right\|^2_{\mathrm{L}^2(0,t^m;\mathrm{L}^2(D))} + 4\Delta t^2 \left\| \frac{\partial^2 \hat{\psi}}{\partial t^2} \right\|^2_{\mathrm{L}^2(0,t^m;\mathrm{L}^2(D))} \right). \end{aligned} \tag{2.30}$$

It remains to bound the first three terms in the bracket on the right-hand side of (2.30). To do so we need to make a specific choice of the finite-dimensional space $\mathcal{P}_N(D)$ from which approximations to $\hat{\psi} \in \mathrm{H}^1_0(D;M)$ are sought, and we also need to specify the projector $\hat{\Pi}_N$. These issues will be discussed in the next section. We shall then return, in Section 2.5, to (2.30) and complete the convergence analysis of the numerical method.

Remark 2.12 *In the case of the FENE model with $b \geq 4s^2/(2s-1)$ and $s > 1/2$ a bound analogous to* (2.30) *can be shown to hold for the fully-discrete version of the semidiscretisation* (2.22) *based on a Chauvière–Lozinski-type transformation, with*

suitable fixed positive constants c_0 and c_2, except that $\mathcal{P}_N(D)$ is then taken to be a finite-dimensional subspace of $\mathrm{H}^1_0(D)$, $\underset{\sim}{\nabla}_M \xi^{n+1}$ on the left-hand side of the bound (2.30) *is replaced by $\underset{\sim}{\nabla}_q \xi^{n+1}$, and the norm $\|\cdot\|_{\ell^2(0,t^m;\mathrm{H}^1_0(D;M))}$ on the right-hand side of* (2.30) *is replaced by $\|\cdot\|_{\ell^2(0,t^m;\mathrm{H}^1_0(D))}$. The main steps of the proof are identical to those above: the Cauchy–Schwarz inequality and inequalities* (2.7) *and* (2.23) *are used in the course of bounding the terms on the right-hand side of an error identity analogous to* (2.26) *relating the sequence $\{\xi^m\}_{m=0}^{N_T}$ to the sequence $\{\eta^m\}_{m=0}^{N_T}$, while the terms on the left-hand side of the error identity are bounded below as in the proof the stability inequality stated in Lemma 2.10.*

We note in particular that the fully-discrete version of the semidiscretisation (2.22) *based on a Chauvière–Lozinski type transformation $\hat{\psi} = \psi/M^{2s/b}$ and the finite-dimensional Galerkin subspace $\mathcal{P}_N(D) \subset \mathrm{H}^1_0(D)$ is unconditionally stable in the sense that the sequence of numerical solutions $\{\hat{\psi}_N^n\}_{n=0}^{N_T}$ generated by the fully-discrete scheme satisfies the stability inequality stated in Lemma 2.10, with $\Delta t = T/N_T$, $N_T \geq 1$, $\underset{\approx}{\kappa} \in (\mathrm{C}[0,T])^{d\times d}$, $\hat{\psi}_N^0 \in \mathcal{P}_N(D)$, $b \geq 4s^2/(2s-1)$, $s > 1/2$, $c_0 := b(d+8s\mathrm{Wi}\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)})^2/(2\mathrm{Wi})$, $0 < c_0\Delta t \leq 1/2$, and ψ^m, ψ^{m-1} and ψ^0 replaced by ψ_N^m, ψ_N^{m-1} and ψ_N^0, respectively, without any conditions relating Δt to N. The proof of this is identical to that of Lemma 2.10,* mutatis mutandis. *We thus deduce that for $b \gg 1$ a time-step limitation of the form $\Delta t = \mathcal{O}(b^{-1})$ is needed in order to ensure that $0 < c_0\Delta t \leq 1/2$, and thereby the stability of the method. In this respect the scheme behaves identically to the fully-discrete numerical method* (2.24), (2.25), *based on the symmetrised form of the Fokker–Planck equation (*cf. *the conditions of Lemma 2.4, for example).* ◇

2.4 Approximation results

It was shown in Section 2.1(b) that, under Hypotheses A and B, $\mathrm{H}^1_0(D) \subset \mathrm{H}^1(D;M) = \mathrm{H}^1_0(D;M)$. Therefore, any finite-dimensional space $\mathcal{P}_N(D) \subset \mathrm{H}^1_0(D)$ is, trivially, also contained in $\mathrm{H}^1_0(D;M)$. The aim now is to make a specific choice of $\mathcal{P}_N(D)$ and to explore the approximation properties of the chosen space.

Remark 2.13 *As in Remark 2.11, if, in addition, $\sqrt{M} \in \mathcal{P}_N(D)$, then*

$$\int_D \psi_N^n(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q} = \int_D \psi_N^0(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q}.$$

In the notation of Lemma 1.3, this can be written as $\varrho_N^n = \varrho_N^0$. Since, by Hypothesis B, $\sqrt{M} \in \mathrm{H}^1_0(D)$, one can ensure that this integral identity holds by including $\sqrt{M}$ in the finite-dimensional space $\mathcal{P}_N(D)$. ◇

The definition of $\mathcal{P}_N(D)$ and the choice of the projector $\hat{\Pi}_N : \mathrm{H}^1_0(D;M) \to \mathcal{P}_N(D)$ will depend on the number d of space dimensions. Since the case of $d = 2$ is sufficiently representative, for the sake of brevity and ease of presentation we shall confine ourselves to two space dimensions in this section, that is, when D is a disc of radius $\sqrt{b}$ in $\mathbb{R}^2$.

Let D_0 denote the slit disc $D_0 := D \setminus \{(q_1, 0) : 0 \le q_1 < \sqrt{b}\}$. It is natural to transform D_0 into the rectangle $(r,\theta) \in R := (0,1) \times (0, 2\pi)$ in a polar co-ordinate system, using the (bijective) change of variables $\underset{\sim}{q} = (q_1, q_2) = (\sqrt{b}\, r\cos\theta, \sqrt{b}\, r \sin\theta) \in D_0$ where $(r,\theta) \in R$. Given $f \in \mathrm{H}^1(D)$, define $\tilde{f}$ on R by

$$\tilde{f}(r,\theta) = f(q_1,q_2), \quad \underset{\sim}{q} = (q_1,q_2) \in D_0, \quad (r,\theta) \in R, \quad q_1 = \sqrt{b}\, r\cos\theta, \quad q_2 = \sqrt{b}\, r \sin\theta. \tag{2.31}$$

Thus,

$$\|f\|^2_{\mathrm{H}^1(D)} = \|f\|^2_{\mathrm{H}^1(D_0)} = \int_0^1 r \int_0^{2\pi} \left(b|\tilde{f}|^2 + |\mathrm{D}_r \tilde{f}|^2 + \left| \frac{\mathrm{D}_\theta \tilde{f}}{r} \right|^2 \right) \mathrm{d}\theta \; \mathrm{d}r.$$

where D_r denotes differentiation with respect to r. Motivated by this identity and writing, here and henceforth, $\tilde{w}(r) := r$ for the weight-function on the interval $(0,1)$, the space $\tilde{\mathrm{H}}^1_{\tilde{w}}(R)$ is defined as:

$$\tilde{\mathrm{H}}^1_{\tilde{w}}(R) := \{\tilde{f} \in \mathrm{L}^2_{\mathrm{loc}}(0,1;\mathrm{H}^1_p(0,2\pi)) : \tilde{f} \in \mathrm{L}^2_{\tilde{w}}(R), \quad \mathrm{D}_r\tilde{f} \in \mathrm{L}^2_{\tilde{w}}(R) \quad \text{and} \quad \frac{1}{r}\mathrm{D}_\theta \tilde{f} \in \mathrm{L}^2_{\tilde{w}}(R)\}, \tag{2.32}$$

equipped with the norm $\|\cdot\|_{\tilde{\mathrm{H}}^1_{\tilde{w}}(R)}$ defined by

$$\|\tilde{f}\|^2_{\tilde{\mathrm{H}}^1_{\tilde{w}}(R)} := \int_0^1 \tilde{w}(r) \int_0^{2\pi} \left(|\tilde{f}|^2 + |\mathrm{D}_r\tilde{f}|^2 + \left| \frac{\mathrm{D}_\theta \tilde{f}}{r} \right|^2 \right) \mathrm{d}\theta \; \mathrm{d}r, \tag{2.33}$$

where $\mathrm{L}^2_{\tilde{w}}(R)$ is the $\tilde{w}$-weighted space of square-integrable functions on R, with norm $\|\cdot\|_{\mathrm{L}^2_{\tilde{w}}(R)}$ defined by

$$\|\tilde{f}\|^2_{\mathrm{L}^2_{\tilde{w}}(R)} := \int_0^1 \tilde{w}(r) \int_0^{2\pi} |\tilde{f}(r,\theta)|^2 \,\mathrm{d}\theta \,\mathrm{d}r = \int_R |\tilde{f}(r,\theta)|^2 \, r \,\mathrm{d}r \,\mathrm{d}\theta,$$

and, for a non-negative integer t, the periodic Sobolev space $\mathrm{H}^t_p(0,2\pi)$ is given by

$$\mathrm{H}^t_p(0,2\pi) := \{\tilde{f} \in \mathrm{H}^t_{\mathrm{loc}}(\mathbb{R}) : \tilde{f}(\theta + 2\pi) = \tilde{f}(\theta) \quad \forall \theta \in \mathbb{R}\}.$$

$\tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$ denotes the subspace of $\tilde{\mathrm{H}}^1_{\tilde{w}}(R)$ consisting of all functions $\tilde{f}$ such that the trace $\tilde{f}(1,\cdot) = 0$.

We shall also require weighted Sobolev spaces of the form $\mathrm{H}^{s,t}_{\tilde{w}}(R) := \mathrm{H}^s_{\tilde{w}}(0,1;\mathrm{H}^t_p(0,2\pi))$, equipped (for non-negative integers s and t) equipped with the norm:

$$\|\tilde{f}\|^2_{\mathrm{H}^{s,t}_{\tilde{w}}(R)} := \sum_{0\le i\le s,\, 0\le j\le t} \int_0^1 \tilde{w}(r) \int_0^{2\pi} |\mathrm{D}_r^i \mathrm{D}_\theta^j \tilde{f}(r,\theta)|^2 \,\mathrm{d}\theta\,\mathrm{d}r.$$

Similarly, for integers $s \ge 1$ and $t \ge 0$, we define $\mathrm{H}^{s,t}_{\tilde{w},0}(R) := \mathrm{H}^s_{\tilde{w},0}(0,1;\mathrm{H}^t_p(0,2\pi))$, where $\mathrm{H}^s_{\tilde{w},0}(0,1) := \mathrm{H}^s_{\tilde{w}}(0,1) \cap \mathrm{H}^1_{\tilde{w},0}(0,1)$, and $\mathrm{H}^1_{\tilde{w},0}(0,1)$ denotes the set of all $\tilde{u} \in \mathrm{H}^1_{\tilde{w}}(0,1)$ such that $\tilde{u}(1) = 0$. $\mathrm{H}^1_{\tilde{w},0}(0,1)$ is endowed with the following inner product and norm:

$$(\tilde{u},\tilde{v})_{\mathrm{H}^1_{\tilde{w},0}(0,1)} := \int_0^1 \tilde{w}(r)\,\mathrm{D}_r\tilde{u}\,\mathrm{D}_r\tilde{v}\;\mathrm{d}r \qquad \text{and} \qquad \|\tilde{u}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} := \{(\tilde{u},\tilde{u})_{\mathrm{H}^1_{\tilde{w},0}(0,1)}\}^{\frac{1}{2}}.$$

Note that $\tilde{w}$ is a Jacobi weight function when transformed to $s \in (-1,1)$, since $\tilde{w}(r(s)) = \frac{1}{2}(1+s)$.[2] This fact will be important later in this section.

Next, the projection operators are introduced. Due to the cartesian product structure of the set R it is natural to define distinct projection operators in the r and θ co-ordinate directions. In the θ-direction, the orthogonal projection in the $\mathrm{L}^2(0,2\pi)$ inner product is used (*i.e.*, truncation of the Fourier series). This is denoted by $P^F_N : \mathrm{L}^2(0,2\pi) \to \mathbb{S}_N(0,2\pi)$, for $N \ge 1$, where $\mathbb{S}_N(0,2\pi)$ is the space of all trigonometric polynomials in $\theta \in [0,2\pi]$ of degree N or less.[3] Also, let $\mathbb{S}_{N_\theta,0}(0,2\pi)$ be the orthogonal complement in $\mathbb{S}_{N_\theta}(0,2\pi)$, with respect to the $\mathrm{L}^2(0,2\pi)$ inner product, of the one-dimensional subspace spanned by constant functions.

The appropriate choice of projector in the r-direction is less immediate. First of all, for $N \ge 1$, let the operator $P^J_N : \mathrm{H}^1_{\tilde{w},0}(0,1) \to \mathbb{P}_{N,0}(0,1)$ be the orthogonal projection in the $\mathrm{H}^1_{\tilde{w},0}(0,1)$ inner product,[4] where $\mathbb{P}_{N,0}(0,1)$ is the space of all algebraic polynomials in $r \in [0,1]$, of degree N or less, that vanish at $r = 1$.

It is tempting to define a two-dimensional projector onto $\mathbb{S}_N(0,2\pi)\otimes\mathbb{P}_{N,0}(0,1)$ as the tensor product of the projectors P^F_N and P^J_N. Unfortunately, this choice is inadequate due to the presence of the singular factor $1/r$ in the weighted Sobolev norm $\|\cdot\|_{\tilde{\mathrm{H}}^1_{\tilde{w}}(R)}$, and a different definition is required. The lemma below motivates the choice of the two-dimensional projector.

Lemma 2.14 (Decomposition Lemma) *Let $\tilde{g} \in \tilde{\mathrm{H}}^1_{\tilde{w}}(R)$ and, for $\varepsilon \in (0,1)$, define $R_\varepsilon := (\varepsilon,1)\times(0,2\pi)$. There exist $\tilde{g}_1 \in \mathrm{H}^1_{\tilde{w}}(0,1)$ and $\tilde{g}_2 \in \mathrm{H}^{0,1}_{\tilde{w}}(R)$, with $\tilde{g}_2 \in \mathrm{H}^1(R_\varepsilon)$*

[2] Jacobi weight functions are of the form $(1-s)^\alpha(1+s)^\beta, s \in (-1,1)$ with $\alpha,\beta > -1$.
[3] The superscript F indicates Fourier projection.
[4] The J superscript indicates projection in a Jacobi-weighted inner-product.

for each $\varepsilon \in (0,1)$ *and* $r\tilde{g}_2 \in \tilde{\mathrm{H}}^1_{\tilde{w}}(R)$, *such that*

$$\tilde{g}(r,\theta) = \tilde{g}_1(r) + r\tilde{g}_2(r,\theta) \quad \textit{for a.e. } (r,\theta) \in R \qquad \textit{and} \qquad \tilde{g}_1(r) := \frac{1}{2\pi}(g(r,\cdot),1)_{\mathrm{L}^2(0,2\pi)}.$$

This is the unique such decomposition of $\tilde{g}$. *If* $\tilde{g} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$, *then* $\tilde{g}_1 \in \mathrm{H}^1_{\tilde{w},0}(0,1)$ *and* $r\tilde{g}_2 \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$, *with* $\tilde{g}_2(1,\cdot) = 0$ *in the sense of the trace theorem on* $\mathrm{H}^1(R_\varepsilon)$, $\varepsilon \in (0,1)$.

Proof. Let $\tilde{g} \in \tilde{\mathrm{H}}^1_{\tilde{w}}(R)$; then, by virtue of Fubini's theorem, $\tilde{g}(r,\cdot) \in \mathrm{H}^1_p(0,2\pi)$ for a.e. $r \in (0,1)$. Let us define, for $r \in (0,1)$, the Fourier coefficients of $\tilde{g}(r,\cdot)$ by

$$\tilde{\gamma}_n(r) := \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \tilde{g}(r,\theta) \exp(-\mathrm{i}n\theta)\,\mathrm{d}\theta, \qquad n = 0, 1, \ldots.$$

According to Parseval's identity,

$$\|\tilde{g}\|^2_{\tilde{\mathrm{H}}^1_{\tilde{w}}(R)} = \sum_{n\in\mathbb{Z}} \int_0^1 \left(|\tilde{\gamma}_n(r)|^2 + |\tilde{\gamma}_n'(r)|^2 + n^2 \left| \frac{\tilde{\gamma}_n(r)}{r} \right|^2 \right) r\,\mathrm{d}r < \infty,$$

whereby, in particular, $\tilde{\gamma}_0 \in \mathrm{H}^1_{\tilde{w}}(0,1)$ and

$$\tilde{\gamma}_n \in \mathrm{H}^1(0,1;r^{-1},r) := \left\{ \tilde{f} \in \mathrm{H}^1_{\mathrm{loc}}(0,1) \,:\, \int_0^1 \left(r^{-1}|\tilde{f}(r)|^2 + r|\tilde{f}'(r)|^2 \right) \mathrm{d}r < \infty \right\},$$

for all $n \in \mathbb{Z} \setminus \{0\}$.

For any $\varepsilon \in (0,1)$ and $n \in \mathbb{Z} \setminus \{0\}$, $\tilde{\gamma}_n \in \mathrm{H}^1(\varepsilon,1)$, and hence by a standard Sobolev embedding, $\tilde{\gamma}_n \in \mathrm{C}(0,1]$. Also, for $0 < r_1 < r_2 < 1$,

$$\begin{aligned} \tilde{\gamma}_n(r_2)^2 - \tilde{\gamma}_n(r_1)^2 &= \int_{r_1}^{r_2} \frac{\mathrm{d}}{\mathrm{d}s}(\tilde{\gamma}_n(s)^2)\,\mathrm{d}s = 2\int_{r_1}^{r_2} \frac{\tilde{\gamma}_n(s)}{\sqrt{s}} \sqrt{s}\,\tilde{\gamma}_n'(s)\,\mathrm{d}s \\ &\le 2\left(\int_{r_1}^{r_2} s^{-1}|\tilde{\gamma}_n(s)|^2\,\mathrm{d}s \right)^{\frac{1}{2}} \left(\int_{r_1}^{r_2} s|\tilde{\gamma}_n'(s)|^2\,\mathrm{d}s \right)^{\frac{1}{2}}, \end{aligned}$$

which is finite by the definition of $\mathrm{H}^1(0,1;r^{-1},r)$, and hence the left-most integral above is finite also. Since the integral is a continous function of its limits, it follows that $\tilde{\gamma}_n^2 \in \mathrm{C}[0,1]$, and hence that $|\tilde{\gamma}_n| = \sqrt{\tilde{\gamma}_n^2} \in \mathrm{C}[0,1]$. We now show that $\tilde{\gamma}_n \in \mathrm{C}(0,1]$ and $|\tilde{\gamma}_n| \in \mathrm{C}[0,1]$ implies that $\tilde{\gamma}_n \in \mathrm{C}[0,1]$.

There are two cases to consider; (i) $|\tilde{\gamma}_n(0)| = 0$, and (ii) $|\tilde{\gamma}_n(0)| > 0$. In case (i), we set $\tilde{\gamma}_n(0) := 0$. Then $|\tilde{\gamma}_n(r) - \tilde{\gamma}_n(0)| = |\tilde{\gamma}_n(r)| = |\,|\tilde{\gamma}_n(r)| - |\tilde{\gamma}_n(0)|\,| \to 0_+$ as $r \to 0_+$, by the continuity of $|\tilde{\gamma}_n|$ on $[0,1]$. In case (ii), there exists $\delta > 0$ such that $|\tilde{\gamma}_n(r)| > 0$ for $r \in [0,\delta]$. Hence the sign of $\tilde{\gamma}_n$ does not change on $(0,\delta]$, so that $\tilde{\gamma}_n$ is either $|\tilde{\gamma}_n|$ or $-|\tilde{\gamma}_n|$ on the interval $(0,\delta]$. Since $|\tilde{\gamma}_n|, -|\tilde{\gamma}_n| \in \mathrm{C}[0,1]$, we can define $\tilde{\gamma}_n(0)$ to be one of $|\tilde{\gamma}_n(0)|$ or $-|\tilde{\gamma}_n(0)|$ so that $\tilde{\gamma}_n \in \mathrm{C}[0,1]$ also.

Now, since $\tilde{\gamma}_n \in \mathrm{C}[0,1]$, Parseval's identity above then implies that, necessarily, $\tilde{\gamma}_n(0) = 0$ for all $n \in \mathbb{Z} \setminus \{0\}$.

Let $\tilde{G}_n(r) := \tilde{\gamma}_n(r)/r$ for $n \in \mathbb{Z} \setminus \{0\}$, $r \in (0,1]$ and $\tilde{E}_n(\theta) := (\exp(\mathrm{i}n\theta))/\sqrt{2\pi}$, $n \in \mathbb{Z}$, $\theta \in [0, 2\pi]$. By Parseval's identity, again, $\sqrt{r^2+n^2}\,\tilde{G}_n \in \mathrm{L}^2_{\tilde{w}}(0,1)$, $n \in \mathbb{Z} \setminus \{0\}$. The following Fourier series expansion of $\tilde{g}$ can be written as follows:

$$\tilde{g} = \frac{1}{\sqrt{2\pi}}\,\tilde{\gamma}_0 + r \sum_{n \in \mathbb{Z}\setminus\{0\}} \tilde{G}_n \tilde{E}_n,$$

with equality in the sense of $\tilde{\mathrm{H}}^1_{\tilde{w}}(R)$. We define $\tilde{g}_1 := \tilde{\gamma}_0/\sqrt{2\pi}$ and $\tilde{g}_2 = \sum_{n\in\mathbb{Z}\setminus\{0\}} \tilde{G}_n\tilde{E}_n$ to deduce the stated decomposition $\tilde{g}(r,\theta) = \tilde{g}_1(r) + r\tilde{g}_2(r,\theta)$, and we note that $\tilde{g}_1 = \frac{1}{2\pi}(\tilde{g},1)_{\mathrm{L}^2(0,2\pi)} \in \mathrm{H}^1_{\tilde{w}}(0,1)$ and $\tilde{g}_2 \in \mathrm{H}^{0,1}_{\tilde{w}}(R)$; moreover, trivially, $r\tilde{g}_2 = \tilde{g} - \tilde{g}_1 \in \tilde{\mathrm{H}}^1_{\tilde{w}}(R)$. Also, since $\tilde{g} \in \tilde{\mathrm{H}}^1_{\tilde{w}}(R)$ it follows that $\tilde{g} \in \mathrm{H}^1(R_\varepsilon)$ and $\tilde{g}_1 \in \mathrm{H}^1(\varepsilon,1)$ for any $\varepsilon \in (0,1)$. Hence, $\tilde{g}_2 = (\tilde{g} - \tilde{g}_1)/r \in \mathrm{H}^1(R_\varepsilon)$ for any $\varepsilon \in (0,1)$.

For $\tilde{g}_1 = \tilde{\gamma}_0/\sqrt{2\pi}$ fixed, as in the statement of the lemma, the uniqueness of $\tilde{g}_2$ follows easily by *reductio ad absurdum*: suppose that $\tilde{h}_2$ is another function, with the same regularity properties as $\tilde{g}_2$, and such that $\tilde{g} = \tilde{g}_1 + r\tilde{h}_2$. Then, $r(\tilde{h}_2 - \tilde{g}_2) = 0$ a.e. on R, and therefore $\tilde{h}_2 = \tilde{g}_2$ a.e. on R.

The final statement of the lemma follows directly from the definitions of $\tilde{\gamma}_n$, $n \in \mathbb{Z}$ and the definitions of $\tilde{g}_1$ and $\tilde{g}_2$ via the $\tilde{\gamma}_n$, $n \in \mathbb{Z}$. $\square$

Suppose that $\tilde{g} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$. On applying Lemma 2.14 we deduce that $\tilde{g}$ has the (unique) decomposition

$$\tilde{g}(r,\theta) = \tilde{g}_1(r) + r\tilde{g}_2(r,\theta), \tag{2.34}$$

where $\tilde{g}_1 := \frac{1}{2\pi}(\tilde{g},1)_{\mathrm{L}^2(0,2\pi)} \in \mathrm{H}^1_{\tilde{w},0}(0,1)$, $\tilde{g}_2 \in \mathrm{H}^{0,1}_{\tilde{w}}(R)$ and $\tilde{g}_2(1,\cdot) = 0$. Note also that $(g_2(r,\cdot),1))_{\mathrm{L}^2(0,2\pi)} = 0$ for a.e. $r \in (0,1)$. We shall assume in addition that $\tilde{g}_2(\cdot,\theta) \in \mathrm{H}^1_{\tilde{w},0}(0,1)$ for a.e. $\theta \in (0,2\pi)$; by virtue of Fubini's theorem, a convenient sufficient condition for this is that $\tilde{g}_2 \in \mathrm{H}^{1,0}_{\tilde{w},0}(R)$, for example. We then define

$$\tilde{P}^J_N\tilde{g}(\cdot,\theta) := P^J_N\tilde{g}_1(\cdot) + rP^J_N\tilde{g}_2(\cdot,\theta), \qquad \theta \in (0,2\pi),$$

where $P^J_N : \mathrm{H}^1_{\tilde{w},0}(0,1) \to \mathbb{P}_{N,0}(0,1)$ is the orthogonal projector defined above.

There are a number of approximation results available in the literature related to projectors in Jacobi-weighted inner products (see for example [12] or [20]). Since the setting here is specific, we shall establish the required approximation properties of the univariate projector P^J_N from first principles. The approximation properties of $\tilde{P}^J_N$ and of our two-dimensional projector $P^F_N\tilde{P}^J_N$ will then follow. The relevant results are stated in the next two lemmas.

Lemma 2.15 *Suppose that* $\tilde{g} \in \mathrm{H}^k_{\tilde{w},0}(0,1)$ *with* $k \geq 1$*; then,*

$$\|\tilde{g} - P_N^J \tilde{g}\|_{\mathrm{H}^1_{\tilde{w}}(0,1)} \leq c N^{1-k} \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)} \tag{2.35}$$

and

$$\|\tilde{g} - P_N^J \tilde{g}\|_{\mathrm{L}^2_{\tilde{w}}(0,1)} \leq c N^{-k} \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)}. \tag{2.36}$$

Proof. First consider (2.35). Note that by Pythagoras' theorem,

$$\|\tilde{g} - P_N^J \tilde{g}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} = \left(\|\tilde{g}\|^2_{\mathrm{H}^1_{\tilde{w},0}(0,1)} - \|P_N^J \tilde{g}\|^2_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \right)^{\frac{1}{2}} \leq \|\tilde{g}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \leq \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)}.$$

If $k = 1$, the right-most term in this chain is equal to $1 \cdot N^{1-k} \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)}$, while if $k \geq 2$ and $1 \leq N < k-1$, then it is bounded by $(k-1)^{k-1} N^{1-k} \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)}$.

Finally, if $k \geq 2$ and $N \geq \max(2, k-1)$, then recall that, by the definition of P_N^J,

$$\|\tilde{g} - P_N^J \tilde{g}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \leq \|\tilde{g} - \tilde{v}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \qquad \forall \tilde{v} \in \mathbb{P}_{N,0}(0,1).$$

Select, in particular,

$$\tilde{v}(r) = -\int_r^1 Q_{N-1}^J \mathrm{D}_s \tilde{g}(s)\, \mathrm{d}s, \qquad r \in [0,1],$$

where Q_{N-1}^J is the orthogonal projector in $\mathrm{L}^2_{\tilde{w}}(0,1)$ onto $\mathbb{P}_{N-1}(0,1)$, the set of all algebraic polynomials of degree $N-1$ or less on the interval $[0,1]$. Thus,

$$\|\tilde{g} - P_N^J \tilde{g}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \leq \|\mathrm{D}_r \tilde{g} - \mathrm{D}_r \tilde{v}\|_{\mathrm{L}^2_{\tilde{w}}(0,1)} = \|\mathrm{D}_r \tilde{g} - Q_{N-1}^J (\mathrm{D}_r \tilde{g})\|_{\mathrm{L}^2_{\tilde{w}}(0,1)} \leq c\,(N-1)^{1-k} \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)},$$

where the last bound (scaled from the standard interval $(-1,1)$ to $(0,1)$) comes from Sec. 5.7.1 of Canuto *et al.* [20], and is valid for $N \geq \max(2, k-1)$, $k \geq 2$. Hence, after bounding $(N-1)^{1-k}$ by $2^{k-1} N^{1-k}$ (recall that $N \geq 2$ by hypothesis), it follows that

$$\|\tilde{g} - P_N^J \tilde{g}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \leq c\, 2^{k-1} N^{1-k} \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)}.$$

Now choosing $\hat{c} = \max\{(k-1)^{k-1}, c\, 2^{k-1}\}$ for $k \geq 1$, with the convention that $0^0 := 1$,

$$\|\tilde{g} - P_N^J \tilde{v}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \leq \hat{c} N^{1-k} \|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)}$$

for all $N \geq 1$ (regardless of whether or not $N \geq k-1$).

For any $\tilde{v} \in \mathrm{H}^1_{\tilde{w},0}(0,1)$, we have:

$$\begin{aligned}
\|\tilde{v}\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} &= \int_0^1 \tilde{v}^2(r) r\, \mathrm{d}r = \int_0^1 \left(\int_r^1 (\sqrt{s}\, \mathrm{D}_s \tilde{v}(s) \frac{1}{\sqrt{s}}\, \mathrm{d}s \right)^2 r\, \mathrm{d}r \\
&\leq \int_0^1 r \left(\int_r^1 |\mathrm{D}_s \tilde{v}(s)|^2 s\, \mathrm{d}s \right) \left(\int_r^1 \frac{1}{s}\, \mathrm{d}s \right) \mathrm{d}r \\
&\leq \left(\int_0^1 r |\log r|\, \mathrm{d}r \right) \|\tilde{v}\|^2_{\mathrm{H}^1_{\tilde{w},0}(0,1)} = \frac{1}{4} \|\tilde{v}\|^2_{\mathrm{H}^1_{\tilde{w},0}(0,1)},
\end{aligned} \tag{2.37}$$

where we make the substitution $r = e^t$ to evaluate $\int_0^1 r|\log r|\,\mathrm{d}r$. It follows from the Friedrichs inequality above that $\|\cdot\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)}$ and $\|\cdot\|_{\mathrm{H}^1_{\tilde{w}}(0,1)}$ are equivalent norms on $\mathrm{H}^1_{\tilde{w},0}(0,1)$, and therefore (2.35) holds for any $N \geq 1$.

The proof of (2.36) is based on a duality argument. Let $e := \tilde{g} - P_N^J \tilde{g}$ and note that, by the hypotheses of the lemma on $\tilde{g}$, we have $e \in \mathrm{L}^2_{\tilde{w}}(0,1)$. Consider the mixed Neumann–Dirichlet boundary-value problem:

$$-\mathrm{D}_r(r\mathrm{D}_r\, z_e(r)) = r\, e(r), \quad r \in (0,1), \qquad \lim_{r\to 0_+} r\mathrm{D}_r z_e(r) = 0, \quad z_e(1) = 0. \tag{2.38}$$

By (2.37) and the Lax–Milgram theorem, this has a unique weak solution $z_e \in \mathrm{H}^1_{\tilde{w},0}(0,1)$ satisfying

$$(z_e, v)_{\mathrm{H}^1_{\tilde{w},0}(0,1)} = (e, v)_{\mathrm{L}^2_{\tilde{w}}(0,1)} \qquad \forall v \in \mathrm{H}^1_{\tilde{w},0}(0,1). \tag{2.39}$$

Also, by (2.37),

$$\|z_e\|^2_{\mathrm{H}^1_{\tilde{w}}(0,1)} \leq \frac{5}{16}\|e\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}.$$

We shall show that in fact $\mathrm{D}_r^2 z_e \in \mathrm{L}^2_{\tilde{w}}(0,1)$, and thereby $z_e \in \mathrm{H}^2_{\tilde{w},0}(0,1)$. To this end, note that

$$\mathrm{D}_r z_e(r) = -\frac{1}{r}\int_0^r s\, e(s)\,\mathrm{d}s, \qquad r \in (0,1].$$

Hence, $\mathrm{D}_r z_e \in \mathrm{C}(0,1]$ and, on recalling that $e \in \mathrm{L}^2_{\tilde{w}}(0,1)$, the Cauchy–Schwarz inequality yields

$$|\mathrm{D}_r z_e(r)|^2 \leq \frac{1}{2}\int_0^r s|e(s)|^2\,\mathrm{d}s, \qquad r \in (0,1]. \tag{2.40}$$

This inequality implies that $\lim_{r\to 0+} \mathrm{D}_r z_e(r) = 0$ and that, for any $\varepsilon \in (0,1)$,

$$\int_\varepsilon^1 \frac{1}{r}\,|\mathrm{D}_r z_e(r)|^2\,\mathrm{d}r \leq \frac{1}{2\varepsilon}\int_0^1 s|e(s)|^2\,\mathrm{d}s.$$

Thus, $\sqrt{r}(r^{-1}\mathrm{D}_r z_e) \in \mathrm{L}^2(\varepsilon,1)$; hence, by (2.38), $\sqrt{r}\,\mathrm{D}_r^2 z_e = -\sqrt{r}\,(e + r^{-1}\mathrm{D}_r z_e) \in \mathrm{L}^2(\varepsilon,1)$. Multiplying this equality by $\sqrt{r}\,\mathrm{D}_r^2 z_e$ and integrating over the interval $(\varepsilon,1)$ gives

$$\int_\varepsilon^1 r\,|\mathrm{D}_r^2 z_e(r)|^2\,\mathrm{d}r + \int_\varepsilon^1 \mathrm{D}_r z_e(r)\,\mathrm{D}_r^2 z_e(r)\,\mathrm{d}r = -\int_\varepsilon^1 r\, e(r)\,\mathrm{D}_r^2 z_e(r)\,\mathrm{d}r.$$

Hence, by computing explicitly the second integral on the left-hand side and applying Cauchy's inequality $|\alpha\beta| \leq \frac{1}{2}(\alpha^2 + \beta^2)$ on the right-hand side, we obtain

$$\int_\varepsilon^1 r\,|\mathrm{D}_r^2 z_e(r)|^2\,\mathrm{d}r + |\mathrm{D}_r z_e(1)|^2 \leq \int_\varepsilon^1 r\,|e(r)|^2\,\mathrm{d}r + |\mathrm{D}_r z_e(\varepsilon)|^2.$$

Passing to the limit $\varepsilon \to 0_+$ and omitting the second term on the left-hand side gives that $\mathrm{D}_r^2 z_e \in \mathrm{L}^2_{\tilde{w}}(0,1)$ and

$$\int_0^1 r\,|\mathrm{D}_r^2 z_e(r)|^2\,\mathrm{d}r \le \int_0^1 r\,|e(r)|^2\,\mathrm{d}r.$$

Combining this with our earlier bound from (2.39), we have that $\|z_e\|^2_{\mathrm{H}^2_{\tilde{w}}(0,1)} \le \frac{21}{16}\|e\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}$.

We are now ready to embark on the analysis of the projection error in the $\mathrm{L}^2_{\tilde{w}}(0,1)$ norm. Recalling that $e = \tilde{g} - P_N^J\tilde{g} \in \mathrm{H}^1_{\tilde{w},0}(0,1)$, we deduce from the weak formulation (2.39), the definition of the orthogonal projector P_N^J, the Cauchy–Schwarz inequality, (2.35) and the $\mathrm{H}^2_{\tilde{w}}(0,1)$ norm bound just derived that

$$\begin{aligned}
\|\tilde{g} - P_N^J\tilde{g}\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} &= (e, \tilde{g} - P_N^J\tilde{g})_{\mathrm{L}^2_{\tilde{w}}(0,1)} = (z_e, \tilde{g} - P_N^J\tilde{g})_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \\
&= (\tilde{g} - P_N^J\tilde{g}, z_e - P_N^J z_e)_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \\
&\le \|\tilde{g} - P_N^J\tilde{g}\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)}\|z_e - P_N^J z_e\|_{\mathrm{H}^1_{\tilde{w},0}(0,1)} \\
&\le cN^{1-k}\|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)} \cdot N^{-1}\|z_e\|_{\mathrm{H}^2_{\tilde{w}}(0,1)} \\
&\le cN^{-k}\|\tilde{g}\|_{\mathrm{H}^k_{\tilde{w}}(0,1)}\|\tilde{g} - P_N^J\tilde{g}\|_{\mathrm{L}^2_{\tilde{w}}(0,1)}, \qquad k \ge 1.
\end{aligned}$$

Dividing the left-most and the right-most term in this chain by $\|\tilde{g} - P_N^J\tilde{g}\|_{\mathrm{L}^2_{\tilde{w}}(0,1)}$ gives (2.36). $\square$

Next, for $\tilde{g} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$, with decomposition given in (2.34), we define the projection operator $\tilde{\Pi}_N : \tilde{\mathrm{H}}^1_{\tilde{w},0}(R) \to \mathcal{P}_N(R)$ as:

$$(\tilde{\Pi}_N\tilde{g})(r,\theta) := (P^F_{N_\theta}\tilde{P}^J_{N_r}\tilde{g})(r,\theta) = (\tilde{P}^J_{N_r}P^F_{N_\theta}\tilde{g})(r,\theta),$$

where the finite-dimensional space $\mathcal{P}_N(R)$ is defined as

$$\mathcal{P}_N(R) := \mathbb{P}_{N_r,0}(0,1) \oplus (r\mathbb{P}_{N_r,0}(0,1) \otimes \mathbb{S}_{N_\theta,0}(0,2\pi)).$$

The structure of this space reflects the decomposition (2.34). Note that the constant functions have been factored out of the space $\mathbb{S}_{N_\theta}(0,2\pi)$ in the definition of $\mathcal{P}_N(R)$; this is appropriate because, as observed above, $(g_2(r,\cdot),1)_{\mathrm{L}^2(0,2\pi)} = 0$. The lemma below establishes optimal order approximation results for this projector.

Lemma 2.16 *Let $\tilde{g} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$, with decomposition $\tilde{g}(r,\theta) = \tilde{g}_1(r) + r\tilde{g}_2(r,\theta)$, where $\tilde{g}_1 = \frac{1}{2\pi}(\tilde{g},1)_{\mathrm{L}^2(0,2\pi)} \in \mathrm{H}^1_{\tilde{w},0}(0,1)$, $\tilde{g}_2 \in \mathrm{H}^{0,1}_{\tilde{w}}(R)$, $\tilde{g}_2(1,\cdot) = 0$, and assume, in addition, that $\tilde{g}_2(\cdot,\theta) \in \mathrm{H}^1_{\tilde{w},0}(0,1)$ for a.e. $\theta \in (0,2\pi)$. If $\tilde{g}_1 \in \mathrm{H}^{k+1}_{\tilde{w}}(0,1)$ and $\tilde{g}_2 \in \mathrm{H}^{k+1,0}_{\tilde{w}}(R) \cap \mathrm{H}^{k,1}_{\tilde{w}}(R) \cap \mathrm{H}^{0,l+1}_{\tilde{w}}(R) \cap \mathrm{H}^{1,l}_{\tilde{w}}(R)$ for some $k,l \ge 1$, then*

$$\begin{aligned}
\|\tilde{g} - \tilde{\Pi}_N\tilde{g}\|_{\tilde{\mathrm{H}}^1_{\tilde{w}}(R)} \le C_1 N_r^{-k}\left(\|\tilde{g}_1\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)} + \|\tilde{g}_2\|^2_{\mathrm{H}^{k+1,0}_{\tilde{w}}(R)} + \|\tilde{g}_2\|^2_{\mathrm{H}^{k,1}_{\tilde{w}}(R)}\right)^{\frac{1}{2}} \\
+ C_2 N_\theta^{-l}\left(\|\tilde{g}_2\|^2_{\mathrm{H}^{0,l+1}_{\tilde{w}}(R)} + \|\tilde{g}_2\|^2_{\mathrm{H}^{1,l}_{\tilde{w}}(R)}\right)^{\frac{1}{2}}. \qquad (2.41)
\end{aligned}$$

If $\tilde{g}_1 \in \mathrm{H}^k_{\tilde{w}}(0,1)$ *and* $\tilde{g}_2 \in \mathrm{H}^{k,0}_{\tilde{w}}(R) \cap \mathrm{H}^{0,l}_{\tilde{w}}(R)$ *for some* $k, l \geq 1$, *then*

$$\|\tilde{g} - \tilde{\Pi}_N \tilde{g}\|_{\mathrm{L}^2_{\tilde{w}}(R)} \leq C_1 N_r^{-k} \left(\|\tilde{g}_1\|^2_{\mathrm{H}^k_{\tilde{w}}(0,1)} + \|\tilde{g}_2\|^2_{\mathrm{H}^{k,0}_{\tilde{w}}(R)} \right)^{\frac{1}{2}} + C_2 N_\theta^{-l} \|\tilde{g}_2\|_{\mathrm{H}^{0,l}_{\tilde{w}}(R)}. \quad (2.42)$$

Proof. The left-hand side in (2.41) is given by:

$$\begin{aligned} \|\tilde{g} - \tilde{\Pi}_N \tilde{g}\|^2_{\tilde{\mathrm{H}}^1_{\tilde{w}}(R)} &= \int_0^1 \tilde{w}(r) \int_0^{2\pi} \left\{ (\tilde{g} - \tilde{\Pi}_N \tilde{g})^2 + (\mathrm{D}_r \tilde{g} - \mathrm{D}_r(\tilde{\Pi}_N \tilde{g}))^2 \right\} \mathrm{d}\theta \, \mathrm{d}r \\ &\quad + \int_0^1 r^{-1} \int_0^{2\pi} (\mathrm{D}_\theta \tilde{g} - \mathrm{D}_\theta(\tilde{\Pi}_N \tilde{g}))^2 \, \mathrm{d}\theta \, \mathrm{d}r \;=:\; I + II. \end{aligned}$$

First consider term I. The two terms in the, inner, θ-integral in I will be treated separately. Using the L^2-error bound for Fourier projection, as well as the fact that $\|P^F_{N_\theta}\|_{\mathcal{L}(\mathrm{L}^2_{\tilde{p}}(0,2\pi), \mathrm{L}^2_{\tilde{p}}(0,2\pi))} \leq 1$, it follows that

$$\begin{aligned} \|\tilde{g}(r,\cdot) - \tilde{\Pi}_N \tilde{g}(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)} &\leq \left(\|\tilde{g}(r,\cdot) - P^F_{N_\theta} \tilde{g}(r,\cdot)\|_{\mathrm{L}^2(0,2\pi)} + \|P^F_{N_\theta}(\tilde{g}(r,\cdot) - \tilde{P}^J_{N_r} \tilde{g}(r,\cdot))\|_{\mathrm{L}^2(0,2\pi)} \right)^2 \\ &\leq \left(C_3 N_\theta^{-l} \|\mathrm{D}^l_\theta \tilde{g}(r,\cdot)\|_{\mathrm{L}^2(0,2\pi)} + \|\tilde{g}(r,\cdot) - \tilde{P}^J_{N_r} \tilde{g}(r,\cdot)\|_{\mathrm{L}^2(0,2\pi)} \right)^2 \\ &\leq 2C_3^2 N_\theta^{-2l} \|\mathrm{D}^l_\theta \tilde{g}_2(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)} + 2\|\tilde{g}(r,\cdot) - \tilde{P}^J_{N_r} \tilde{g}(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)}, \end{aligned}$$

where $\mathrm{D}^l_\theta \tilde{g} = r \mathrm{D}^l_\theta \tilde{g}_2$ and $0 \leq r \leq 1$ have been used in the last line. Similarly,

$$\begin{aligned} \|\mathrm{D}_r \tilde{g}(r,\cdot) - \mathrm{D}_r(\tilde{\Pi}_N \tilde{g}(r,\cdot))\|^2_{\mathrm{L}^2(0,2\pi)} &\leq 2\|\mathrm{D}_r \tilde{g} - P^F_{N_\theta} \mathrm{D}_r \tilde{g}\|^2_{\mathrm{L}^2(0,2\pi)} \\ &\quad + 2\|\mathrm{D}_r P^F_{N_\theta} \tilde{g} - \mathrm{D}_r P^F_{N_\theta} \tilde{P}^J_{N_r} \tilde{g}(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)} \\ &\leq 2C_3^2 N_\theta^{-2l} \|\mathrm{D}^l_\theta \mathrm{D}_r \tilde{g}(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)} \\ &\quad + 2\|\mathrm{D}_r \tilde{g} - \mathrm{D}_r \tilde{P}^J_{N_r} \tilde{g}(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)} \\ &\leq 4C_3^2 N_\theta^{-2l} \left(\|\mathrm{D}^l_\theta \tilde{g}_2(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)} + \|\mathrm{D}_r \mathrm{D}^l_\theta \tilde{g}_2(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)} \right) \\ &\quad + 2\|\mathrm{D}_r \tilde{g}(r,\cdot) - \mathrm{D}_r \tilde{P}^J_{N_r} \tilde{g}(r,\cdot)\|^2_{\mathrm{L}^2(0,2\pi)}. \end{aligned}$$

Therefore,

$$\begin{aligned} I \leq 6\, C_3^2 N_\theta^{-2l} \int_0^{2\pi} \left(\|\mathrm{D}^l_\theta \tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} + \|\mathrm{D}_r \mathrm{D}^l_\theta \tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} \right) \mathrm{d}\theta \\ + 2 \int_0^{2\pi} \|\tilde{g}(\cdot,\theta) - \tilde{P}^J_{N_r} \tilde{g}(\cdot,\theta)\|^2_{\mathrm{H}^1_{\tilde{w}}(0,1)} \, \mathrm{d}\theta. \end{aligned}$$

The final term on the right-hand side of the last inequality can then be bounded using the univariate estimate (2.35):

$$\begin{aligned}
\|\tilde{g}(\cdot,\theta) - \tilde{P}^J_{N_r}\tilde{g}(\cdot,\theta)\|^2_{\mathrm{H}^1_{\tilde{w}}(0,1)} &\leq 2\|\tilde{g}_1 - P^J_{N_r}\tilde{g}_1\|^2_{\mathrm{H}^1_{\tilde{w}}(0,1)} + 2\|r(\tilde{g}_2(\cdot,\theta) - P^J_{N_r}\tilde{g}_2(\cdot,\theta))\|^2_{\mathrm{H}^1_{\tilde{w}}(0,1)} \\
&\leq C^2 N_r^{-2k}\|\tilde{g}_1\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)} \\
&\quad + 2\int_0^1 \tilde{w}(r)\left\{(2+r^2)(\tilde{g}_2(r,\theta) - P^J_{N_r}\tilde{g}_2(r,\theta))^2 + 2r^2(\mathrm{D}_r(\tilde{g}_2(r,\theta) - P^J_{N_r}\tilde{g}_2(r,\theta)))^2\right\}\,\mathrm{d}r \\
&\leq C^2 N_r^{-2k}\|\tilde{g}_1\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)} + 6\|\tilde{g}_2(\cdot,\theta) - P^J_{N_r}\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{H}^1_{\tilde{w}}(0,1)} \\
&\leq C_4^2 N_r^{-2k}\left(\|\tilde{g}_1\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)} + \|\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)}\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
I \leq 6\,C_3^2 N_\theta^{-2l}\int_0^{2\pi}\left(\|\mathrm{D}^l_\theta\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} + \|\mathrm{D}_r\mathrm{D}^l_\theta\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\right)\mathrm{d}\theta \\
+ 2\,C_4^2 N_r^{-2k}\int_0^{2\pi}\left(\|\tilde{g}_1\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)} + \|\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)}\right)\mathrm{d}\theta, \qquad (2.43)
\end{aligned}$$

which is an optimal-order bound on I.

Next, consider II. Since θ-differentiation commutes with the projectors $P^J_{N_r}$ and $P^F_{N_\theta}$, it follows that

$$\begin{aligned}
II \;\leq\; & 2\int_0^1 r^{-1}\int_0^{2\pi}|\mathrm{D}_\theta\tilde{g}(r,\theta) - P^F_{N_\theta}\mathrm{D}_\theta\tilde{g}(r,\theta)|^2\,\mathrm{d}\theta\;\mathrm{d}r \\
& + 2\int_0^1 r^{-1}\int_0^{2\pi}|P^F_{N_\theta}\mathrm{D}_\theta\tilde{g}(r,\theta) - \tilde{P}^J_{N_r}(P^F_{N_\theta}\mathrm{D}_\theta\tilde{g}(r,\theta))|^2\,\mathrm{d}\theta\;\mathrm{d}r.
\end{aligned}$$

Therefore,

$$\begin{aligned}
II \;\leq\; & 2\int_0^1 r^{-1}\int_0^{2\pi}\left|r\mathrm{D}_\theta\tilde{g}_2(r,\theta) - rP^F_{N_\theta}\mathrm{D}_\theta\tilde{g}_2(r,\theta)\right|^2\,\mathrm{d}\theta\;\mathrm{d}r \\
& + 2\int_0^{2\pi}\int_0^1 r^{-1}|rP^F_{N_\theta}\mathrm{D}_\theta\tilde{g}_2(r,\theta) - \tilde{P}^J_{N_r}(rP^F_{N_\theta}\mathrm{D}_\theta\tilde{g}_2(r,\theta))|^2\;\mathrm{d}r\;\mathrm{d}\theta \\
\leq\; & C_5^2 N_\theta^{-2l}\int_0^1\tilde{w}(r)\int_0^{2\pi}|\mathrm{D}^{l+1}_\theta\tilde{g}_2(r,\theta)|^2\;\mathrm{d}\theta\;\mathrm{d}r \\
& + 2\int_0^{2\pi}\int_0^1\tilde{w}(r)|P^F_{N_\theta}\mathrm{D}_\theta\tilde{g}_2(r,\theta) - P^J_{N_r}(P^F_{N_\theta}\mathrm{D}_\theta\tilde{g}_2(r,\theta))|^2\;\mathrm{d}r\;\mathrm{d}\theta \\
\leq\; & C_5^2 N_\theta^{-2l}\int_0^{2\pi}\|\mathrm{D}^{l+1}_\theta\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\,\mathrm{d}\theta + C_6^2 N_r^{-2k}\int_0^{2\pi}\|P^F_{N_\theta}\mathrm{D}_\theta\tilde{g}_2(r,\theta)\|^2_{\mathrm{H}^k_{\tilde{w}}(0,1)}\,\mathrm{d}\theta.
\end{aligned}$$

Where the $\mathrm{L}^2_w(0,1)$ norm error bound for $P^J_{N_r}$ stated in (2.36), as well as the fact that $\tilde{P}^J_{N_r}(r\tilde{g}_2) = rP^J_{N_r}(\tilde{g}_2)$ have been used in the argument above. For the second integral

in the last line in the bound on II,

$$\sum_{j=0}^{k}\int_0^1 \tilde{w}(r)\|P^F_{N_\theta}\mathrm{D}_r^j\mathrm{D}_\theta\tilde{g}_2(\cdot,r)\|^2_{\mathrm{L}^2(0,2\pi)}\,\mathrm{d}r \le \sum_{j=0}^{k}\int_0^1 \tilde{w}(r)\|\mathrm{D}_r^j\mathrm{D}_\theta\tilde{g}_2(\cdot,r)\|^2_{\mathrm{L}^2(0,2\pi)}\,\mathrm{d}r.$$

Therefore,

$$II \le C_5^2 N_\theta^{-2l}\int_0^{2\pi}\|\mathrm{D}_\theta^{l+1}\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\,\mathrm{d}\theta + C_6^2 N_r^{-2k}\int_0^{2\pi}\|\mathrm{D}_\theta\tilde{g}_2(\cdot,\theta)\|^2_{\mathrm{H}^k_{\tilde{w}}(0,1)}\,\mathrm{d}\theta.$$

Combining the bounds for I and II with suitable constants C_1 and C_2, gives

$$\begin{aligned}\|\tilde{g}-P^F_{N_\theta}\tilde{P}^J_{N_r}\tilde{g}\|_{\tilde{\mathrm{H}}^1_{\tilde{w}}(R)} \;\le\; & C_1N_r^{-k}\left\{\int_0^{2\pi}(\|\tilde{g}_1\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)}+\|\tilde{g}_2\|^2_{\mathrm{H}^{k+1}_{\tilde{w}}(0,1)}+\|\mathrm{D}_\theta\tilde{g}_2\|^2_{\mathrm{H}^k_{\tilde{w}}(0,1)})\,\mathrm{d}\theta\right\}^{\frac12}\\ & +C_2N_\theta^{-l}\left\{\int_0^{2\pi}(\|\mathrm{D}_\theta^{l+1}\tilde{g}_2\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}+\|\mathrm{D}_\theta^{l}\tilde{g}_2\|^2_{\mathrm{H}^1_{\tilde{w}}(0,1)})\,\mathrm{d}\theta\right\}^{\frac12}, \qquad (2.44)\end{aligned}$$

which is (2.41). The proof of the $\mathrm{L}^2_{\tilde{w}}(R)$ norm bound (2.42) is very similar: its main ingredients are, in fact, contained in the argument above. Therefore, for the sake of brevity, the details are omitted here. □

The bounds (2.41) and (2.42) can now be straightforwardly mapped from R to D_0. We define $\mathcal{P}_N(D)$ as $\mathcal{P}_N(R)$ mapped from R to D_0 using the polar coordinate transformation (2.31), and we suppose that $\hat{\psi}\in\mathcal{H}^{k+1,l+1}(D)$, with $k,l\ge1$, where

$$\begin{aligned}\mathcal{H}^{k,l}(D) \;:=\; & \{g\in\mathrm{H}^1_0(D)\,:\,\tilde{g}\in\tilde{\mathrm{H}}^1_{\tilde{w},0}(R)\text{ has a decomposition }\tilde{g}(r,\theta)=\tilde{g}_1(r)+r\tilde{g}_2(r,\theta),\\ & \text{with }\tilde{g}_1=\tfrac{1}{2\pi}(\tilde{g},1)_{\mathrm{L}^2(0,2\pi)}\in\mathrm{H}^k_{\tilde{w},0}(0,1)\\ & \text{and }\tilde{g}_2\in\mathrm{H}^{k,0}_{\tilde{w},0}(R)\cap\mathrm{H}^{k-1,1}_{\tilde{w}}(R)\cap\mathrm{H}^{0,l}_{\tilde{w}}(R)\cap\mathrm{H}^{1,l-1}_{\tilde{w}}(R)\},\end{aligned}$$

equipped with the norm $\|g\|_{\mathcal{H}^{k,l}(D)}:=\left(\|g\|^2_{\mathcal{H}^k_r(D)}+\|g\|^2_{\mathcal{H}^l_\theta(D)}\right)^{\frac12}$ where, for $\tilde{g}=\tilde{g}_1+r\tilde{g}_2\in\mathcal{H}^{k,l}(D)$,

$$\begin{aligned}\|g\|_{\mathcal{H}^k_r(D)} &:= \left(\|\tilde{g}_1\|^2_{\mathrm{H}^k_{\tilde{w}}(0,1)}+\|\tilde{g}_2\|^2_{\mathrm{H}^{k,0}_{\tilde{w}}(R)}+\|\tilde{g}_2\|^2_{\mathrm{H}^{k-1,1}_{\tilde{w}}(R)}\right)^{\frac12},\\ \|g\|_{\mathcal{H}^l_\theta(D)} &:= \left(\|\tilde{g}_2\|^2_{\mathrm{H}^{0,l}_{\tilde{w}}(R)}+\|\tilde{g}_2\|^2_{\mathrm{H}^{1,l-1}_{\tilde{w}}(R)}\right)^{\frac12}.\end{aligned}$$

We define

$$\hat{\Pi}_N:\mathcal{H}^{1,1}(D)\to\mathcal{P}_N(D)\quad\text{by}\quad(\hat{\Pi}_Ng)(q_1,q_2)=(\tilde{\Pi}_N\tilde{g})(r,\theta),\qquad g\in\mathcal{H}^{1,1}(D).$$

Thus, recalling (2.8) and noting that $\mathcal{H}^{k,l}(D)\subset\mathrm{H}^1_0(D)\subset\mathrm{H}^1_0(D;M)$, $k,l\ge1$, we deduce from (2.41) that

$$\|\hat{\psi}-\hat{\Pi}_N\hat{\psi}\|_{\mathrm{H}^1_0(D;M)}\le C_1N_r^{-k}\|\hat{\psi}\|_{\mathcal{H}^{k+1}_r(D)}\;+\;C_2N_\theta^{-l}\|\hat{\psi}\|_{\mathcal{H}^{l+1}_\theta(D)} \qquad (2.45)$$

for all $\hat{\psi} \in \mathcal{H}^{k+1,l+1}(D)$, with $k, l \geq 1$. Similarly, we obtain from (2.42) that

$$\|\hat{\psi} - \hat{\Pi}_N \hat{\psi}\|_{\mathrm{L}^2(D)} \leq C_1 N_r^{-k} \|\hat{\psi}\|_{\mathcal{H}_r^k(D)} \ + \ C_2 N_\theta^{-l} \|\hat{\psi}\|_{\mathcal{H}_\theta^l(D)} \tag{2.46}$$

for all $\hat{\psi} \in \mathcal{H}^{k,l}(D)$, with $k, l \geq 1$.

2.5 Convergence analysis of the numerical method

In this section we use the two-dimensional approximation results derived in Section 2.4 to complete the convergence analysis of the fully-discrete numerical method (2.24), (2.25), based on the symmetrised form of the Fokker–Planck equation. At the end of the section we shall comment on the extension of our results to a fully-discrete method that stems from the alternative semidiscretisation (2.22) in the case of the FENE model.

We see from (2.30) that in order to obtain bounds on the norms of ξ appearing on the left-hand side of (2.30) we need to bound the following terms:

$$\|\eta^0\|, \quad \|\eta\|_{\ell^2(0,T;\mathrm{H}_0^1(D;M))} \quad \text{and} \quad \left\| \frac{\partial \eta}{\partial t} \right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(D))}.$$

It follows from (2.45), (2.46) and the definition of $\eta := \hat{\psi} - \hat{\Pi}_N \hat{\psi}$ that

$$\|\eta^0\| \leq \|\hat{\psi}_0 - \hat{\Pi}_N \hat{\psi}_0\| \leq C_1 N_r^{-k} \|\hat{\psi}_0\|_{\mathcal{H}_r^k(D)} \ + \ C_2 N_\theta^{-l} \|\hat{\psi}_0\|_{\mathcal{H}_\theta^l(D)},$$

$$\|\eta\|_{\ell^2(0,T;\mathrm{H}_0^1(D;M))} \leq C_1 N_r^{-k} \|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_r^{k+1}(D))} \ + \ C_2 N_\theta^{-l} \|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_\theta^{l+1}(D))},$$

$$\left\| \frac{\partial \eta}{\partial t} \right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(D))} \leq C_1 N_r^{-k} \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|_{\mathrm{L}^2(0,T;\mathcal{H}_r^k(D))} \ + \ C_2 N_\theta^{-l} \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|_{\mathrm{L}^2(0,T;\mathcal{H}_\theta^l(D))},$$

with $k, l \geq 1$, provided that $\hat{\psi}$ is such that the right-hand sides of these inequalities are finite. Substituting these three bounds into the right-hand side of (2.30) we deduce, with $m\Delta t \leq T$, $m = 0, 1, \ldots, N_T$, that

$$\begin{aligned}
&\|\xi\|_{\ell^\infty(0,T;\mathrm{L}^2(D))} + \|\underset{\sim}{\nabla}_M \xi\|_{\ell^2(0,T;\mathrm{L}^2(D))} \\
&\quad \leq C_1 N_r^{-k} \left(\|\hat{\psi}_0\|_{\mathcal{H}_r^k(D)} + \|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_r^{k+1}(D))} + \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|_{\mathrm{L}^2(0,T;\mathcal{H}_r^k(D))} \right) \\
&\quad + \ C_2 N_\theta^{-l} \left(\|\hat{\psi}_0\|_{\mathcal{H}_\theta^l(D)} + \|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_\theta^{l+1}(D))} + \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|_{\mathrm{L}^2(0,T;\mathcal{H}_\theta^l(D))} \right) \\
&\qquad + C_3 \Delta t \left\| \frac{\partial^2 \hat{\psi}}{\partial t^2} \right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(D))}.
\end{aligned} \tag{2.47}$$

Note, also, that

$$\begin{aligned}\|\eta\|_{\ell^\infty(0,T;\mathrm{L}^2(D))} &\leq C_1 N_r^{-k}\|\hat{\psi}\|_{\ell^\infty(0,T;\mathcal{H}_r^k(D))} + C_2 N_\theta^{-l}\|\hat{\psi}\|_{\ell^\infty(0,T;\mathcal{H}_\theta^l(D))}, \quad (2.48)\\ \|\nabla_M\eta\|_{\ell^2(0,T;\mathrm{L}^2(D))} &\leq C_1 N_r^{-k}\|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_r^{k+1}(D))} + C_2 N_\theta^{-l}\|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_\theta^{l+1}(D))}. \quad (2.49)\end{aligned}$$

Now, by the triangle inequality,

$$\begin{aligned}&\|\hat{\psi}-\hat{\psi}_N\|_{\ell^\infty(0,T;\mathrm{L}^2(D))} + \|\nabla_M(\hat{\psi}-\hat{\psi}_N)\|_{\ell^2(0,T;\mathrm{L}^2(D))}\\ &\quad\leq \|\xi\|_{\ell^\infty(0,T;\mathrm{L}^2(D))} + \|\nabla_M\xi\|_{\ell^2(0,T;\mathrm{L}^2(D))}\\ &\qquad\quad +\|\eta\|_{\ell^\infty(0,T;\mathrm{L}^2(D))} + \|\nabla_M\eta\|_{\ell^2(0,T;\mathrm{L}^2(D))},\end{aligned}$$

whereby (2.47), (2.48) and (2.49) give

$$\begin{aligned}&\|\hat{\psi}-\hat{\psi}_N\|_{\ell^\infty(0,T;\mathrm{L}^2(D))} + \|\nabla_M(\hat{\psi}-\hat{\psi}_N)\|_{\ell^2(0,T;\mathrm{L}^2(D))}\\ &\quad\leq C_1 N_r^{-k}\left(\|\hat{\psi}\|_{\ell^\infty(0,T;\mathcal{H}_r^k(D))} + \|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_r^{k+1}(D))} + \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathcal{H}_r^k(D))}\right)\\ &\qquad + C_2 N_\theta^{-l}\left(\|\hat{\psi}\|_{\ell^\infty(0,T;\mathcal{H}_\theta^l(D))} + \|\hat{\psi}\|_{\ell^2(0,T;\mathcal{H}_\theta^{l+1}(D))} + \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathcal{H}_\theta^l(D))}\right)\\ &\qquad\quad + C_3\Delta t\left\|\frac{\partial^2\hat{\psi}}{\partial t^2}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(D))}.\end{aligned}$$

We recall that $\psi = \sqrt{M}\hat{\psi}$, and we define $\psi_N^n := \sqrt{M}\hat{\psi}_N^n$. Consequently,

$$\begin{aligned}&\|\psi-\psi_N\|_{\ell^\infty(0,T;\mathfrak{H})} + \|\psi-\psi_N\|_{\ell^2(0,T;\mathfrak{K})}\\ &\quad\leq C_1 N_r^{-k}\left(\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^\infty(0,T;\mathcal{H}_r^k(D))} + \left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathcal{H}_r^{k+1}(D))} + \left\|\frac{1}{\sqrt{M}}\frac{\partial\psi}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathcal{H}_r^k(D))}\right)\\ &\qquad + C_2 N_\theta^{-l}\left(\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^\infty(0,T;\mathcal{H}_\theta^l(D))} + \left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathcal{H}_\theta^{l+1}(D))} + \left\|\frac{1}{\sqrt{M}}\frac{\partial\psi}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathcal{H}_\theta^l(D))}\right)\\ &\qquad\quad + C_3\Delta t\left\|\frac{1}{\sqrt{M}}\frac{\partial^2\psi}{\partial t^2}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(D))}, \quad (2.50)\end{aligned}$$

with $k, l \geq 1$, provided that ψ is such that right-hand side is finite.

That completes the convergence analysis of the method in the case of $d = 2$. For $d = 3$ the argument is identical, and rests on a three dimensional analogue of Lemma 2.14, this is discussed further in Section 2.6.3.

Starting from the second stability inequality stated in Lemma 2.9 and proceeding in an identical manner as above, one can derive analogous error bounds in the $h^1(0,T;\mathfrak{H})$ and $\ell^\infty(0,T;\mathfrak{K})$ norms.

Remark 2.17 *In the case of the FENE Maxwellian, $\sqrt{M} \in \mathcal{P}_N(D)$ if, and only if, there exists a positive integer m such that $b = 4m$ and $N_r \geq 2m$. In order to ensure that, more generally, $\sqrt{M} \in \mathcal{P}_N(D)$ regardless of the specific choice of b and the value of N_r, one can simply enrich $\mathcal{P}_N(D)$ by adding $\sqrt{M}$ as an extra basis function. However, in general the polynomials in $\mathcal{P}_N(D)$ approximate $\sqrt{M}$ very closely, so this leads to a highly ill-conditioned basis. A better solution is to add the component of $\sqrt{M}$ orthogonal to $\mathcal{P}_N(D)$ (in the $\mathrm{L}^2(D)$ inner product, for example,) to the basis, rather than $\sqrt{M}$ itself. This is implemented in Section 2.6 for a numerical example in which b is not divisible by 4 and is shown to work well in that case.* $\diamond$

Remark 2.18 *We make a second comment regarding the FENE model. Starting from the variant of the inequality* (2.30) *alluded to in Remark 2.10 in connection with the fully-discrete spectral method based on the semidiscretisation* (2.22) *with $b \geq 4s^2/(2s-1)$ and $s > 1/2$, one can derive an optimal-order error bound analogous to* (2.50)*. The core of the argument is identical to the one above, and is therefore omitted.* $\diamond$

2.6 Numerical results

Numerical methods for solving the Fokker–Planck equation arising from the FENE dumbbell model for dilute polymeric fluids have been the focus of some attention recently; Du *et al.* [29] developed a finite difference scheme that preserved the unit integral property and the positivity of ψ, Chauvière & Lozinski [23,24,59,60] developed a spectral method for this problem and Ammar *et al.* [2, 3] proposed a reduced-basis method for solving the Fokker–Planck equation for FENE dumbbell chains. For a survey of, alternative, stochastic techniques for the numerical simulation of polymeric liquids we refer to the monograph of Öttinger [68] and the article of Jourdain, Lelièvre, and Le Bris [42], for example. The computational results we present in this section are for the FENE potential only, although it would be straightforward to modify the numerical methods to apply to more general potentials that satisfy Hyptheses A and B.

In Section 2.6.1, we discuss the implementation of two spectral Galerkin methods for the $d = 2$ case based on the formulation (2.24), (2.25). We then present computational results for these schemes in order to illustrate their behaviour in practice, as well as to provide experimental support for the convergence theory developed in Section 2.5. Next, we compare the two spectral Galerkin methods based on the formulation (2.24), (2.25) with the method of Chauvière & Lozinski based on the 'original' form (2.4) of the Fokker–Planck equation (or, more precisely, its transformed version (2.20) resulting

from the substitution (2.60), with $s = 2$). Section 2.6.1 is concluded with a discussion of the convergence rate of the extra-stress tensor, $\underset{\approx}{\tau}$.

In Section 2.6.2, we present some numerical results for a semi-implicit temporal discretisation of the Fokker–Planck equation in order to compare its performance with the backward Euler scheme that has been emphasised in this chapter. Finally, we consider the implementation of spectral Galerkin method in three spatial dimensions in Section 2.6.3, and we show some computational results to demonstrate that the 3-dimensional scheme exhibits essentially the same behaviour as the schemes considered in the $d = 2$ case in Section 2.6.1.

2.6.1 Numerical methods in the two dimensional case

With $D \subset \mathbb{R}^2$, we suppose that $\hat{\psi} \in \mathrm{H}^1_0(D)$ and hence, $\tilde{\psi} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$, where $\tilde{\psi}(r,\theta) := \hat{\psi}(q_1, q_2)$ with $q_1 = \sqrt{b}\, r\cos\theta$, $q_2 = \sqrt{b}\, r\sin\theta$. Using the decomposition (2.34), $\tilde{\psi}$ can be written in polar coordinates as follows:

$$\tilde{\psi}(r,\theta) = \tilde{\psi}_1(r) + r\tilde{\psi}_2(r,\theta), \qquad (r,\theta) \in R = (0,1)\times(0,2\pi), \tag{2.51}$$

where, as in Section 2.4, r has been scaled from $(0,\sqrt{b})$ to $(0,1)$, and $\tilde{\psi}_1 := \frac{1}{2\pi}(\tilde{\psi},1)_{\mathrm{L}^2(0,2\pi)}$. In the context of spectral methods in polar coordinates, (2.51) is referred to by Shen as the *essential pole condition* [75]. This condition is a 'first-order' form of the following full pole-condition [30]: in order that a function

$$\tilde{\psi}(r,\theta) = \sum_{n\in\mathbb{Z}} \tilde{\gamma}_n(r)\tilde{E}_n(\theta), \qquad \text{where} \quad \tilde{E}_n(\theta) := \frac{1}{\sqrt{2\pi}}\exp(\mathrm{i}n\theta),$$

is infinitely differentiable when transformed from polar to cartesian coordinates, it is necessary that, for each $n \in \mathbb{Z}\setminus\{0\}$,

$$\tilde{\gamma}_n(r) = \mathcal{O}(r^{|n|}) \qquad \text{as } r \to 0_+. \tag{2.52}$$

That (2.51) is a 'first-order' form of the full pole condition is easily seen by writing $\tilde{\gamma}_n(r) = r^{|n|}\tilde{G}_n(r)$, with $\tilde{G}_n(r) = \mathcal{O}(1)$ as $r \to 0_+$; hence,

$$\tilde{\psi}(r,\theta) = \frac{1}{\sqrt{2\pi}}\,\tilde{\gamma}_0(r) + r\sum_{n\in\mathbb{Z}\setminus\{0\}}^{\infty} r^{|n|-1}\tilde{G}_n(r)\tilde{E}_n(\theta) =: \tilde{\psi}_1(r) + r\tilde{\psi}_2(r,\theta),$$

with $\tilde{\psi}_1(r) = \tilde{\gamma}_0(r)/\sqrt{2\pi} = \frac{1}{2\pi}(\tilde{\psi},1)_{\mathrm{L}^2(0,2\pi)}$, as required.

The full pole condition (2.52) is consistent with the result established in the proof of Lemma 2.14 stating that the expansion coefficients $\tilde{\gamma}_n$, $n \in \mathbb{Z}\setminus\{0\}$, of a function

in $\tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$ satisfy $\tilde{\gamma}_n(r) = o(1)$ as $r \to 0_+$, although the conditions (2.52) are clearly much more restrictive.

In order to fit into the framework of the numerical analysis in Sections 2.4 and 2.5, each element of $\mathcal{P}_N(R)$ should satisfy (2.51) to ensure that $\mathcal{P}_N(D)$ is contained in $\mathrm{H}^1_0(D)$. The discrete space $\mathcal{P}_N(R)$, introduced in Section 2.4, satisfies this property. In this section we define a spectral Galerkin method for the Fokker–Planck equation based on a particular basis (denoted $\mathcal{A}$) for $\mathcal{P}_N(R)$ that satisfies the same decomposition.

For the purpose of comparison, we also introduce a second basis, $\mathcal{B}$, in which each function satisfies the full pole condition, (2.52). Thus, on mapping $\mathcal{B}$ from R to D we obtain a basis for a finite-dimensional subspace of $\mathrm{C}^\infty(\overline{D}) \cap \mathrm{C}_0(\overline{D}) \subset \mathrm{H}^1_0(D)$. The reason for considering this second basis is that typical solutions of the FENE Fokker–Planck equation are smooth on D, and therefore it is likely that in practice a Galerkin method based on $\mathcal{B}$ will be more accurate than a method based on $\mathcal{A}$: mapping the basis $\mathcal{A}$ from R to D yields a finite-dimensional subspace of $\mathrm{H}^1_0(D)$ only, which contains functions that are not smooth at the origin in D. We note, however, that the span of $\mathcal{B}$ does not coincide with $\mathcal{P}_N(R)$, and therefore the approximation properties of $\mathcal{B}$ are not covered by the results in Section 2.4 that led to the error bounds in Section 2.5. Hence, the numerical results for basis $\mathcal{A}$ are intended to verify the analysis developed in the previous sections, while basis $\mathcal{B}$ is introduced to indicate the gain in performance that can be obtained by satisfying (2.52). By requiring more regularity from the basis than it being a finite-dimensional subspace of $\mathrm{H}^1_0(D)$ one could modify the arguments in Section 2.4 to derive convergence estimates based on a pole condition of higher order than (2.34), but this would make the derivation of the approximation results more laborious (*e.g.*, the projector $\tilde{P}^J_N$ would have to obey (2.52) rather than (2.51)). Before introducing bases $\mathcal{A}$ and $\mathcal{B}$, we make the following observation.

Remark 2.19 *Let $\hat{\psi}$ be the weak solution of* (2.6) *corresponding to a given initial condition $\hat{\psi}_0$, define $\hat{\psi}^*(\underset{\sim}{q},t) := \hat{\psi}(-\underset{\sim}{q},t)$ and suppose that $\hat{\psi}_0$ is invariant under the change of independent variable $\underset{\sim}{q} \mapsto -\underset{\sim}{q}$,* i.e., *$\hat{\psi}_0(\underset{\sim}{q}) = \hat{\psi}_0(-\underset{\sim}{q})$ for a.e. $\underset{\sim}{q} \in D$. On noting that $M(\underset{\sim}{q}) = M(-\underset{\sim}{q})$, $\underset{\sim}{q} \in D$, it follows that the weak formulation* (2.6) *is also invariant under this change of variable; hence $\hat{\psi}$ and $\hat{\psi}^*$ are weak solutions to the same initial boundary-value problem. It follows by uniqueness of the weak solution established in Section 2.2 that $\hat{\psi}(\underset{\sim}{q},t) \equiv \hat{\psi}^*(\underset{\sim}{q},t)$,* i.e., *$\hat{\psi}(\underset{\sim}{q},t) = \hat{\psi}(-\underset{\sim}{q},t)$ for a.e. $\underset{\sim}{q} \in D$ and a.e. $t \in [0,T]$. This evenness of $\hat{\psi}$ in the D domain with respect to $\underset{\sim}{q}$ translates into π-periodicity of $\tilde{\psi}$ in the R domain with respect to θ. An identical statement applies to the numerical solution $(\hat{\psi}^n_N)_{n=0}^{N_T}$ defined by* (2.24), (2.25), *provided $\mathcal{P}_N(D) \subset \mathrm{H}^1_0(D)$ is such that whenever a function $\underset{\sim}{q} \mapsto v(\underset{\sim}{q})$ belongs to $\mathcal{P}_N(D)$ its even reflection $\underset{\sim}{q} \mapsto$*

$v(-\underset{\sim}{q})$ also belongs to $\mathcal{P}_N(D)$: if $\hat{\psi}_0(\underset{\sim}{q}) = \hat{\psi}_0(-\underset{\sim}{q})$ for a.e. $\underset{\sim}{q} \in D$, uniqueness of the $\mathrm{L}^2(D)$ projection of $\hat{\psi}^0$ onto $\mathcal{P}_N(D)$ implies that $\hat{\psi}^0_N(\underset{\sim}{q}) = \hat{\psi}^0_N(-q)$ for a.e. $\underset{\sim}{q} \in D$. Uniqueness of the numerical solution then yields $\hat{\psi}^n_N(\underset{\sim}{q}) = \hat{\psi}^n_N(-\underset{\sim}{q})$ for a.e. $\underset{\sim}{q} \in D$ and all $n = 0, \ldots, N_T$. $\diamond$

The above remark demonstrates that (2.6) captures an important symmetry property of the dumbbell model for polymeric fluids: the configuration probability density function ψ is required to be symmetric about the origin in D because the beads of a dumbbell are indistinguishable. As long as $\hat{\psi}_0$ and $\mathcal{P}_N(D)$ are invariant under the change of independent variable $\underset{\sim}{q} \mapsto -\underset{\sim}{q}$ described in Remark 2.19, the numerical solution will inherit the symmetry of the analytical solution implied by the symmetry of the initial condition. A consequence of this observation is that we should require the basis functions in $\mathcal{A}$ and $\mathcal{B}$ to obey the same symmetry condition; following [61], this is achieved in the definitions below by only including even trigonometric modes in θ. Strictly speaking therefore $\mathcal{A}$ is chosen to be a basis for the linear subspace of $\mathcal{P}_N(R)$ consisting of all π-periodic functions. Note, however, that if the solution were 2π-periodic, then one could simply include odd trigonometric modes as well. We are now ready to define the bases $\mathcal{A}$ and $\mathcal{B}$.

Basis $\mathcal{A}$: Let $\mathcal{A} := \mathcal{A}_1 \cup \mathcal{A}_2$ where:

$$\mathcal{A}_1 := \{(1-r)P_k(r) : k = 0, \ldots, N_r - 1\},$$
$$\mathcal{A}_2 := \{r(1-r)P_k(r)\Phi_{il}(\theta) : k = 0, \ldots, N_r - 1; \quad i = 0, 1; \quad l = 1, \ldots, N_\theta\}.$$

P_k is a polynomial of degree k in $r \in [0, 1]$ and $\Phi_{il}(\theta) = (1 - i)\cos(2l\theta) + i\sin(2l\theta)$, $\theta \in [0, \pi]$. We denote by P_k the kth Chebyshev polynomial scaled from $[-1, 1]$ to $[0, 1]$. The numerical method is not particularly sensitive to this choice of polynomial, however, and other choices work well also. Notice that the polynomials in $\mathcal{A}_1$ and $\mathcal{A}_2$ both contain the factor $(1-r)$ in order to impose the homogeneous Dirichlet boundary condition on ∂D, and functions in $\mathcal{A}_2$ contain an extra factor of r to enforce the essential pole condition. Basis $\mathcal{A}$ is chosen so as to mimic the decomposition (2.51) of the analytical solution $\tilde{\psi} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$ in polar coordinates: the role of $\mathrm{span}(\mathcal{A}_1)$ is to approximate $\tilde{\psi}_1$ while $\mathrm{span}(\mathcal{A}_2)$ is meant to approximate $r\tilde{\psi}_2$.

Basis $\mathcal{B}$: This is, effectively, the basis proposed by Matsushima and Marcus [64] and Verkley [79], except that, as above, we ensure that the functions are zero at $r = 1$ and that they are π-periodic in θ:

$$\mathcal{B} = \{W_{lk}(r)\Phi_{il}(\theta) : k = 0, \ldots, N_r - 1; \quad i = 0, 1; \quad l = i, \ldots, N_\theta\}, \tag{2.53}$$

where $W_{lk}(r) = r^{2l}(1-r^2)J_k^{(0,2l)}(2r^2-1)$ and $J_k^{(\alpha,\beta)}(x)$ is the Jacobi polynomial on $[-1,1]$ of degree k with respect to the weight $(1-x)^\alpha(1+x)^\beta$ (Φ_{il} is the same as in $\mathcal{A}$). Each element of $\mathcal{B}$ satisfies (2.52).

$\mathcal{A}$ and $\mathcal{B}$ both have cardinality $N := N_r\,(2N_\theta+1)$. Expressing trial and test functions in terms of either $\mathcal{A}$ or $\mathcal{B}$, it is now straightforward to determine the discretisation matrices corresponding to the integrals

$$\int_D \hat{\psi}_N^{n+1}\hat{\varphi}\;\mathrm{d}\underset{\sim}{q}, \qquad \int_D \underset{\sim}{\nabla}_M\hat{\psi}_N^{n+1}\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\;\mathrm{d}\underset{\sim}{q}, \qquad \int_D (\underset{\approx}{\kappa}^{n+1}\,\underset{\sim}{q}\,\hat{\psi}_N^{n+1})\cdot\underset{\sim}{\nabla}_M\hat{\varphi}\;\mathrm{d}\underset{\sim}{q} \tag{2.54}$$

from (2.24). We label these matrices $\mathbf{M}$, $\mathbf{S}$ and $\mathbf{C}^{n+1}$ for mass, stiffness and convection respectively.

Using the ansatz $\tilde{\psi}_N^{n+1}(r,\theta) = \sum_{v=1}^N \tilde{\Psi}_v^{n+1}Y_v(r,\theta)$ for trial functions, where Y_v is a basis function (from either $\mathcal{A}$ or $\mathcal{B}$) for $1\le v\le N$, denoting test functions as Y_u for $1\le u\le N$ and mapping (2.54) from D to R yields:

$$\mathbf{M}_{uv} = \int_0^1\int_0^\pi b\,r\,Y_v(r,\theta)Y_u(r,\theta)\;\mathrm{d}r\;\mathrm{d}\theta, \tag{2.55}$$

$$\begin{aligned}\mathbf{S}_{uv} = \int_0^1\int_0^\pi \Bigg\{ & r\frac{\partial Y_v}{\partial r}\frac{\partial Y_u}{\partial r} + \frac{1}{r}\frac{\partial Y_v}{\partial\theta}\frac{\partial Y_u}{\partial\theta} \\ & + \frac{b}{2}\frac{r^2}{1-r^2}\frac{\partial}{\partial r}(Y_uY_v) + \frac{b^2}{4}\frac{r^3}{(1-r^2)^2}Y_vY_u\Bigg\}\;\mathrm{d}r\;\mathrm{d}\theta,\end{aligned} \tag{2.56}$$

$$\begin{aligned}\mathbf{C}_{uv}^{n+1} = & \int_0^1\int_0^\pi br\,Y_v\frac{\partial Y_u}{\partial\theta}(-\kappa_{11}^{n+1}\sin 2\theta - \kappa_{12}^{n+1}\sin^2\theta + \kappa_{21}\cos^2\theta)\;\mathrm{d}r\;\mathrm{d}\theta \\ & + \int_0^1\int_0^\pi \left(b\,r^2\,Y_v\frac{\partial Y_u}{\partial r} + \frac{b^2}{2}\frac{r^3}{1-r^2}Y_vY_u\right)\times \\ & \qquad \left(\kappa_{11}^{n+1}\cos 2\theta + \frac{1}{2}(\kappa_{12}^{n+1}+\kappa_{21}^{n+1})\sin 2\theta\right)\;\mathrm{d}r\;\mathrm{d}\theta.\end{aligned} \tag{2.57}$$

Note that if the Y_u, Y_v do not satisfy (2.51), then the entries of $\mathbf{S}$ may be undefined.

With these discretisation matrices in hand, the numerical solution is computed by solving the following linear system for the coefficient vector $\tilde{\mathbf{\Psi}}^{n+1} := (\tilde{\Psi}_1^{n+1},\ldots,\tilde{\Psi}_N^{n+1})^{\mathrm{T}} \in \mathbb{R}^N$, $n = 0,1,\ldots,N_T-1$:

$$\left(\mathbf{M} + \Delta t\left(\frac{1}{2\mathrm{Wi}}\mathbf{S} - \mathbf{C}^{n+1}\right)\right)\tilde{\mathbf{\Psi}}^{n+1} = \mathbf{M}\tilde{\mathbf{\Psi}}^n, \tag{2.58}$$

with $\tilde{\Psi}^0$ defined by the initial datum. Then, the numerical approximation to the probability density function itself is obtained as $\psi_N^{n+1}(\underset{\sim}{q}) = \sqrt{M(\underset{\sim}{q})}\,\tilde{\psi}_N^{n+1}(r,\theta)$, where $r = |\underset{\sim}{q}|/\sqrt{b}$ and $\tilde{\psi}_N^{n+1}(r,\theta) = \sum_{v=1}^N \tilde{\Psi}_v^{n+1}Y_v(r,\theta)$.

For ease of evaluation, the integrals in (2.55), (2.56) and (2.57) can be factorised into products of 1-dimensional integrals over r and θ. We evaluate the θ-integrals exactly

using trigonometric identities, and, noting that the r-integrands are all polynomials, we use Gauss quadrature to evaluate the r-integrals to machine precision. $\mathbf{M}$ and $\mathbf{S}$ are constant matrices, which can be pre-computed and reused, but if $\underset{\approx}{\kappa}$ is time-varying, we must reassemble $\mathbf{C}^{n+1}$ at every time-step. However, it is straightforward to factor out the dependence of $\mathbf{C}^{n+1}$ on $\underset{\approx}{\kappa}$ so that the integrals that determine $\mathbf{C}^{n+1}$ need not be evaluated more than once. We use LU-decomposition to solve (2.58), which is appropriate because the spectral discretisation matrices are generally of moderate size.

We now present some numerical results. For simplicity, in the computations considered below we always use the normalised Maxwellian (which satisfies the symmetry property required in Remark 2.19 and also has unit volume) as the initial condition, so that $\hat{\psi}_0(\underset{\sim}{q}) = \sqrt{M(\underset{\sim}{q})}$. Also, most of the results presented in this section are for computations in which b was chosen to be divisible by 4 so that the spaces span($\mathcal{A}$) and span($\mathcal{B}$) naturally contain $\sqrt{M}$, as in Remark 2.17. However, the basis enrichment technique described in Remark 2.17 was implemented to obtain the results in Table 2.3 (in which $b = 10$) and, as discussed below, it worked well for that problem.

Henceforth, the two numerical methods that use basis $\mathcal{A}$ and basis $\mathcal{B}$, respectively, will be referred to as method $\mathcal{A}$ and method $\mathcal{B}$.

First of all we present results from solving the Fokker–Planck equation with parameters $b = 16$, $\mathrm{Wi} = 1.2$ and $\kappa_{11} = -\kappa_{22} = 1.1,\ \kappa_{12} = 0.9,\ \kappa_{21} = -0.6$ and with $\Delta t = 0.05$. These parameters were chosen somewhat arbitrarily, but the intention here is to visualise a typical evolution of ψ_N towards steady state, and to provide an initial qualitative comparison of methods $\mathcal{A}$ and $\mathcal{B}$ (quantitative convergence results will be presented below). By taking $(N_r, N_\theta) = (26, 20)$ with basis $\mathcal{A}$ and $(N_r, N_\theta) = (21, 15)$ with basis $\mathcal{B}$, the solutions from the two methods were indistinguishable to the eye and appear to be fully resolved. As foreshadowed above, $\mathcal{A}$ required more degrees-of-freedom than $\mathcal{B}$ to resolve the solution to comparable accuracy in this case because, as can be seen in Figure 2.1, ψ_N is smooth at the origin in cartesian coordinates whereas the basis functions in $\mathcal{A}$ are not necessarily smooth there. Nevertheless, a clear advantage of basis $\mathcal{A}$ over basis $\mathcal{B}$ is that it is built by relying on the essential pole condition only, as manifested by the decomposition in Lemma 2.14, which only requires the most basic smoothness hypothesis, that $\tilde{\psi} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$ (implied by the assumption that the weak solution $\hat{\psi} \in \mathrm{H}^1_0(D; M)$ belongs to $\mathrm{H}^1_0(D)$).

Figure 2.1 shows snapshots of ψ_N at $t = 0,\ t = 1,\ t = 2$ and $t = 3$, and ψ_N is close to steady state at $t = 3$.

To provide a quantitative study of the spatial accuracy of the numerical methods defined in this section, we use the fact that when $\underset{\approx}{\kappa}$ is a symmetric tensor the exact

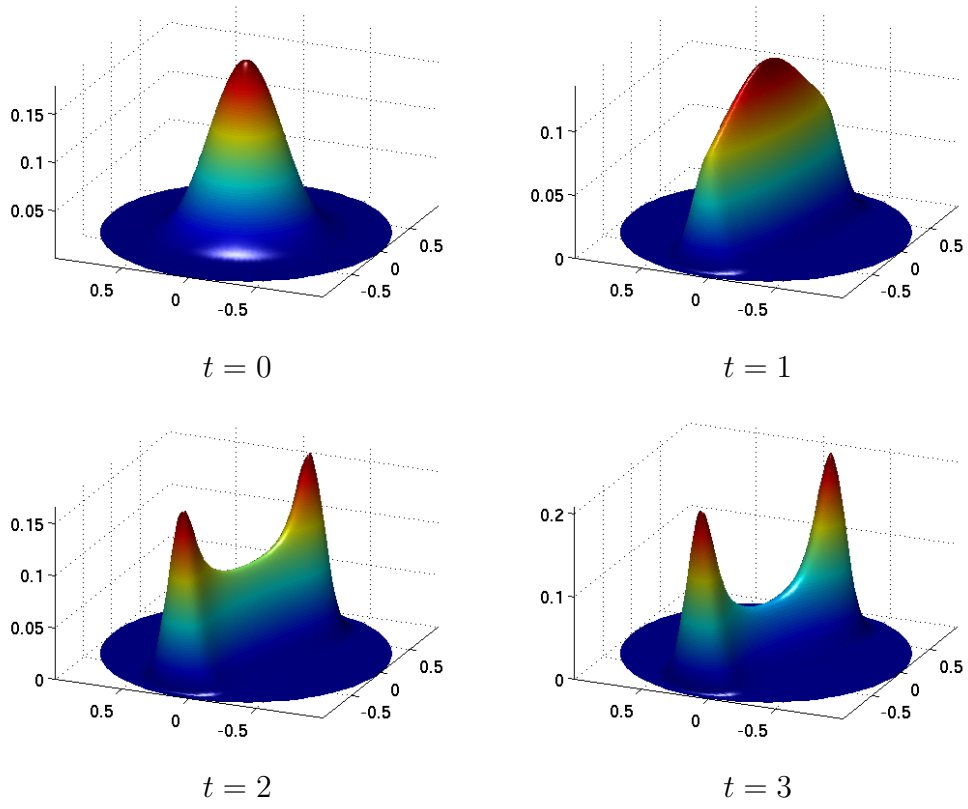


Figure 2.1: Snapshots of ψ_N at $t = 0$, $t = 1$, $t = 2$ and $t = 3$ illustrating evolution towards steady state. In this case, we have $\Delta t = 0.05$, $b = 16$, $\mathrm{Wi} = 1.2$ and $\kappa_{11} = -\kappa_{22} = 1.1$, $\kappa_{12} = 0.9$, $\kappa_{21} = -0.6$. This computation was performed using basis $\mathcal{A}$ and basis $\mathcal{B}$ with $(N_r, N_\theta) = (26, 20)$ and $(N_r, N_\theta) = (21, 15)$, respectively. The solutions were fully resolved in each of these two cases.

steady-state solution of the Fokker–Planck equation is given by

$$\psi_{\mathrm{exact}}(\underset{\sim}{q}) := M(\underset{\sim}{q}) \exp(\mathrm{Wi}\, \underset{\sim}{q}^{\mathrm{T}} \underset{\approx}{\kappa} \underset{\sim}{q}), \tag{2.59}$$

where C is a normalization constant chosen so that $\int_D \psi_{\mathrm{exact}}(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} = 1$; see, [18]. We now consider a particular case, referred to as *extensional flow*, in which $\underset{\approx}{\kappa} = \mathrm{diag}(\delta, -\delta)$. This generally provides a good test case for numerical methods for the Fokker–Planck equation because it yields particularly sharp solution profiles that are challenging to resolve, and also the exact steady-state solution is available for comparison. In order to compare the convergence rates of methods $\mathcal{A}$ and $\mathcal{B}$, we solved two distinct extensional flow problems for: (i) $(b, \mathrm{Wi}, \delta) = (12, 1, 1)$ and (ii) $(b, \mathrm{Wi}, \delta) = (20, 1, 2)$, with a range of choices of (N_r, N_θ). In order to compare to the known exact steady-state solution, we took 2000 time-steps (with $\Delta t = 0.05$ and $T = 100$) in each case so that the final numerical solution is a very close approximation to the steady-state

solution. This allows us to compare the spatial convergence rates of the two numerical methods without worrying about temporal discretisation error. Tables 2.1 and 2.2 show the relative errors (in the $\mathrm{L}^2(D)$ and $\mathrm{H}^1(D;M)$ norms) between the exact and the computed steady-state solutions for extensional flows (i) and (ii), respectively.

We can see from the data in the tables that methods $\mathcal{A}$ and $\mathcal{B}$ converge rapidly for both problem (i) and problem (ii) and that for each choice of (N_r, N_θ), basis $\mathcal{B}$ outperforms basis $\mathcal{A}$ – again this is because the solution profiles are smooth at the origin in cartesian coordinates, see Figure 2.2. Nevertheless, the rapid convergence of method $\mathcal{A}$ is consistent with the spectral error estimates established in Section 2.5 (recall that these error estimates do not apply to method $\mathcal{B}$ because span($\mathcal{B}$) is not the same as $\mathcal{P}_N(R)$ analysed in Section 2.4). It is also clear that problem (ii) is more challenging to resolve than problem (i); with both $\mathcal{A}$ and $\mathcal{B}$, more basis functions are required to attain a given accuracy for problem (ii) than for problem (i). Note that the greater difficulty of resolving extensional flow (ii) is encoded in the convergence estimates in Section 2.5 because the constants in these estimates depend exponentially on b, δ (via $\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)}$) and T. Moreover, the factor $\mathrm{e}^{2c_0 m\Delta t}$ on the right-hand side in Lemma 2.4 permits exponential growth in time of the norm of $\hat{\psi}_N$, and this is reflected in the first row of Table 2.2 in which the solutions computed with $(N_r, N_\theta) = (10, 10)$ for extensional flow (ii) resulted in numerical overflow.[5] Note that this overflow behaviour was only observed in the case of under-resolved computations that led to numerical solutions containing numerical oscillations *i.e.* it was not observed in rows 2, 3 and 4 of Table 2.2; note also that Chauvière & Lozinski's method behaves in the same way for under-resolved solutions, as shown in Table 2.3.

The (fully resolved) solutions corresponding to extensional flow problems (i) and (ii) are shown in Figure 2.2, and in each case both ψ_N and $\tilde{\psi}_N$ are plotted. It is clear that the solution profiles corresponding to (ii) are much more severe, and therefore it is not surprising that more modes were required in this case. The quantity of interest in these computations is ψ_N, but $\tilde{\psi}_N$ is also plotted to emphasise the numerical difficulties that are encountered as b and δ are increased. In the plots corresponding to (i), the peaks in $\tilde{\psi}_N$ are higher than in ψ_N, but only by a factor of about 20. For (ii) on the other hand, the peaks in $\tilde{\psi}_N$ are higher by a factor of roughly 5000. The causes of this behaviour are two-fold: with $\delta = 2$ the flow has stronger extensional character and therefore the solution peaks are expected to be more concentrated and also, the larger

[5] When $\underset{\sim}{q}^{\mathrm{T}}\underset{\approx}{\kappa}(t)\underset{\sim}{q} = 0$ for all $t \in [0,T]$, Lemma 2.4, with $\mu = 0$ and $\underset{\sim}{\nu} = \underset{\sim}{0}$, can be sharpened. The inequality holds with $c_0 = 0$, showing that the expression on the left-hand side of the inequality is bounded by $\|\hat{\psi}^0\|^2$, uniformly in T, b and $\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T)}$.

	Relative $\mathrm{L}^2(D)$ error		Relative $\mathrm{H}^1(D;M)$ error	
(N_r, N_θ)	Basis $\mathcal{A}$	Basis $\mathcal{B}$	Basis $\mathcal{A}$	Basis $\mathcal{B}$
(10,10)	3.63×10^{-2}	4.61×10^{-3}	7.90×10^{-2}	8.82×10^{-3}
(15,15)	3.36×10^{-3}	9.19×10^{-6}	8.58×10^{-3}	2.33×10^{-5}
(20,20)	5.13×10^{-5}	4.63×10^{-9}	1.64×10^{-4}	1.52×10^{-8}
(25,25)	2.94×10^{-7}	1.74×10^{-12}	1.13×10^{-6}	6.94×10^{-12}
(30,30)	8.31×10^{-10}	1.70×10^{-13}	3.77×10^{-9}	1.70×10^{-13}

Table 2.1: Relative errors in the $\mathrm{L}^2(D)$ and $\mathrm{H}^1(D;M)$ norms (*i.e.* $\|\hat{\psi}_N - \hat{\psi}_{\mathrm{exact}}\|/\|\hat{\psi}_{\mathrm{exact}}\|$ and $\|\hat{\psi}_N - \hat{\psi}_{\mathrm{exact}}\|_{\mathrm{H}^1(D;M)}/\|\hat{\psi}_{\mathrm{exact}}\|_{\mathrm{H}^1(D;M)}$, respectively) for extensional flow (i) at steady-state, *i.e.* $b = 12$, $\mathrm{Wi} = 1$ and $\delta = 1$. $\hat{\psi}_N$ is an approximation to the steady-state solution obtained by taking 2000 time-steps with $\Delta t = 0.05$, and $\hat{\psi}_{\mathrm{exact}}$ is the exact steady-state solution which is known in this case because $\underset{\approx}{\kappa}$ is symmetric.

	Relative $\mathrm{L}^2(D)$ error		Relative $\mathrm{H}^1(D;M)$ error	
(N_r, N_θ)	Basis $\mathcal{A}$	Basis $\mathcal{B}$	Basis $\mathcal{A}$	Basis $\mathcal{B}$
(10,10)	–	–	–	–
(15,15)	2.47×10^{-1}	9.57×10^{-2}	1.79×10^{-1}	9.53×10^{-2}
(20,20)	3.91×10^{-2}	1.72×10^{-3}	4.88×10^{-2}	2.54×10^{-3}
(25,25)	9.07×10^{-3}	1.71×10^{-4}	9.77×10^{-3}	2.37×10^{-4}
(30,30)	1.50×10^{-3}	2.97×10^{-6}	2.61×10^{-3}	4.49×10^{-6}
(35,35)	3.37×10^{-4}	2.14×10^{-8}	5.60×10^{-4}	3.66×10^{-8}
(40,40)	2.54×10^{-5}	5.97×10^{-9}	4.55×10^{-5}	5.94×10^{-9}

Table 2.2: Relative errors in the $\mathrm{L}^2(D)$ and $\mathrm{H}^1(D;M)$ norms for extensional flow (ii) at steady-state, *i.e.* $(b, \mathrm{Wi}, \delta) = (20, 1, 2)$. The time-stepping strategy to compute the approximate steady-state solution was the same as in Table 2.1. The hyphens in the first row indicate that we obtained numerical overflow in those computations.

value of b means that $\sqrt{M}$ is more strongly degenerate near ∂D so that $\hat{\psi}_N = \psi_N/\sqrt{M}$ takes larger values near the boundary. This second point can be seen as a drawback, for $b \gg 1$, of the fully-discrete numerical method (2.24), (2.25), based on the symmetrised form of the Fokker–Planck equation. Presumably Chauviére & Lozinski [24] fixed their value of s ($s = 2$ for $d = 2$ and $s = 2.5$ for $d = 3$) in the transformation

$$\hat{\psi}(\underset{\sim}{q}) := \psi(\underset{\sim}{q})/[M(\underset{\sim}{q})]^{2s/b} = \psi(\underset{\sim}{q})/(1 - |\underset{\sim}{q}|^2/b)^s \tag{2.60}$$

so as to avoid a similar effect; indeed, they presented some numerical results for $b = 200$. Values of b this large do not appear to be feasible with the fully-discrete method (2.24), (2.25), based on the substitution $\hat{\psi}_N = \psi_N/\sqrt{M}$.

As has been noted in Remark 2.12, there is in fact no difference between the stability properties of the method based on (2.24), (2.25) and of a Chauvière–Lozinski type

method. However, if $b \gg 1$, for a typical ψ we have that $\|\psi/\sqrt{M}\|_{\mathrm{L}^\infty(D)} = \|\psi/(1 - |\underset{\sim}{q}|^2/b)^{b/4}\|_{\mathrm{L}^\infty(D)} \gg \|\psi/(1 - |\underset{\sim}{q}|^2/b)^2\|_{\mathrm{L}^\infty(D)}$. Hence, compared to a Chauvière–Lozinski type method with the recommended choice of $s = 2$ for $d = 2$, the maximum value of the numerical approximation $\hat{\psi}_N$ to the function $\hat{\psi}$ defined by the scheme (2.24), (2.25) can be much larger when $b \gg 1$, and can thereby require greater computational effort to resolve to a given accuracy. The computational results that we consider in this section are therefore restricted to moderate values of b.

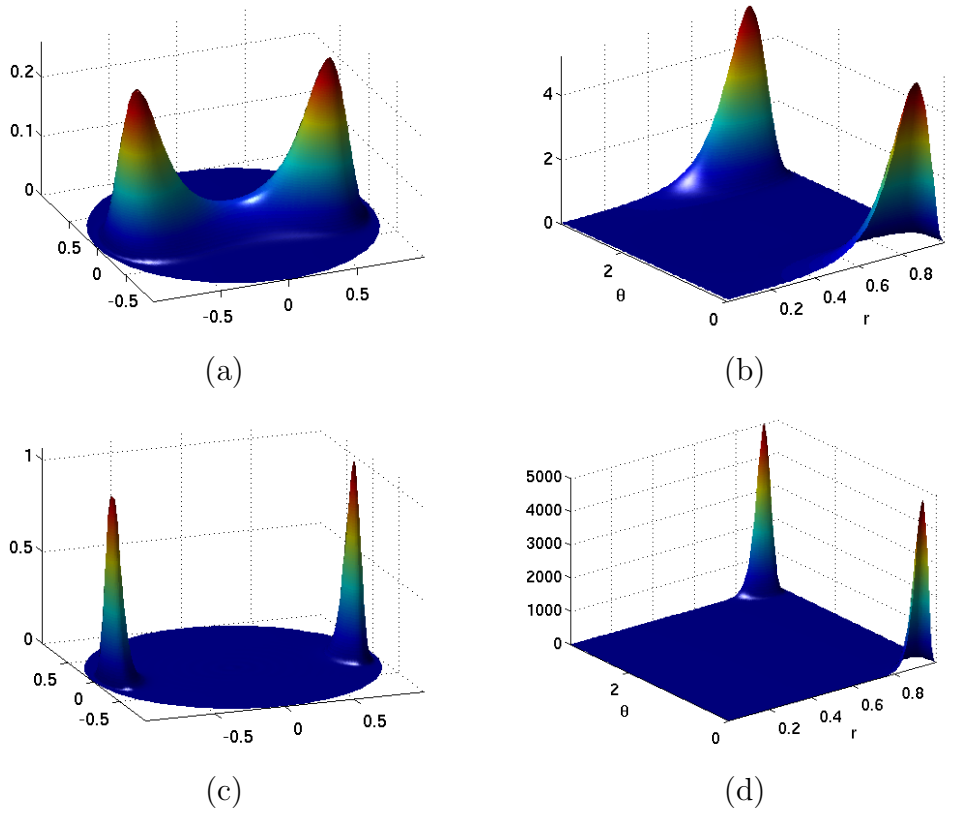


Figure 2.2: Numerical approximations to the steady state solution for extensional flow problems (i) and (ii) using $(N_r, N_\theta) = (30, 30)$ and $(N_r, N_\theta) = (40, 40)$, respectively. Plots (a) and (b) show ψ_N and $\tilde{\psi}_N$ respectively, at steady state for problem (i) and (c), (d) show ψ_N and $\tilde{\psi}_N$ for (ii). The purpose of plots (b) and (d) is to demonstrate that $\tilde{\psi}_N$ usually has a much steeper solution profile than ψ_N and this effect is amplified if either δ or b (or both) are increased.

With these precursors, we now compare the accuracy of methods $\mathcal{A}$ and $\mathcal{B}$ to that of the spectral method of Chauviére & Lozinski discussed in [24]. In Table 2 of that paper, the authors presented convergence data for the $(1, 1)$-component of the *polymeric extra-stress* tensor, $\underset{\approx}{\tau} = (\tau_{ij})$, computed for an extensional flow at steady state for the

parameters $(b, \lambda, \delta) = (10, 1, 5)$. Note that when ψ is a function of $\underset{\sim}{q}$ and t only, $\underset{\approx}{\tau}$ is defined as:

$$\underset{\approx}{\tau}(t) := \int_D \underset{\sim}{F} \otimes \underset{\sim}{q}\, \psi(\underset{\sim}{q}, t)\, \mathrm{d}\underset{\sim}{q} = \int_D \underset{\sim}{F} \otimes \underset{\sim}{q}\, \sqrt{M}\, \hat{\psi}(\underset{\sim}{q}, t)\, \mathrm{d}\underset{\sim}{q}, \tag{2.61}$$

where $\underset{\sim}{F}$ is taken to be the FENE spring force here. Table 2.3 reproduces Chauviére & Lozinski's results and compares them to the corresponding results for methods $\mathcal{A}$ and $\mathcal{B}$. Note that in this problem b is not divisible by 4. Therefore, in order to ensure that the volume of ψ_N is conserved with methods $\mathcal{A}$ and $\mathcal{B}$, we added the component of $\sqrt{M}$ orthogonal to span($\mathcal{A}$) (resp. span($\mathcal{B}$)) to the bases to obtain an enriched discrete space that contains $\sqrt{M}$ (*cf.* Remark 2.17).[6] This ensured that the volume of ψ_N was conserved to machine precision (except in the cases that rounding error polluted the results, these are indicated by hyphens in the table).

The data in Table 2.3 show that for this problem method $\mathcal{B}$ converges at a comparable rate to the method of Chauviére & Lozinski, whereas $\mathcal{A}$ appears to converge more slowly. Note that the reason why method $\mathcal{B}$ and Chauviére & Lozinski's method converge at a similar rate (at least in this case where b is relatively low) is that both methods involve ansatzes that impose extra regularity at the origin in cartesian coordinates; basis $\mathcal{B}$ satisfies the pole condition (2.52), and Chauviére & Lozinski use a transformation that enforces $\frac{\partial \psi}{\partial r}\big|_{r=0} = 0$, which, when combined with π-periodicity in θ, has a similar effect.

	Relative error of τ_{11}		
(N_r, N_θ)	Basis $\mathcal{A}$	Basis $\mathcal{B}$	Chauviére & Lozinski
(11,5)	–	–	–
(13,6)	–	4.8×10^{-2}	0.35
(21,10)	1.8×10^{-3}	2.0×10^{-2}	2.0×10^{-2}
(31,15)	2.1×10^{-4}	1.4×10^{-4}	1.4×10^{-4}
(41,20)	1.3×10^{-5}	8.7×10^{-7}	2.1×10^{-7}

Table 2.3: Comparison of the relative errors in τ_{11} for extensional flow with $(b, \mathrm{Wi}, \delta) = (10, 1, 5)$. The three schemes compared are methods $\mathcal{A}$ and $\mathcal{B}$ and the spectral method of Chauviére & Lozinski. The data for the method of Chauviére & Lozinski is taken from Table 2 in [24].

In fact, as discussed in Section 1.3.3, in the context of deterministic multiscale computations for the micro-macro model, the primary reason for solving the Fokker–Planck equation is to obtain an approximation of $\underset{\approx}{\tau}$. Therefore, the computational results in Table 2.3 are of great interest, and to shed further light on these results we now consider the convergence of $\underset{\approx}{\tau}$ from a theoretical point of view.

[6] Orthogonalisation was performed in the $\mathrm{L}^2(D)$ inner product.

Let $\tilde{\psi} \in \tilde{\mathrm{H}}^1_{\tilde{w},0}(R)$ be the weak solution of (2.6) (transformed to polar coordinates). As in the proof of Lemma 2.14, we have

$$\tilde{\psi}(r,\theta,t) = \tilde{\psi}_1(r,t) + r\sum_{l=1}^{\infty}\Big(\tilde{A}_l(r,t)\cos(2l\theta) + \tilde{B}_l(r,t)\sin(2l\theta)\Big), \tag{2.62}$$

where we have only taken even modes in the sum (*cf.* Remark 2.19) and we use sin and cos functions in (2.62) rather than complex exponentials to match the structure of bases $\mathcal{A}$ and $\mathcal{B}$. For simplicity, we will restrict our attention to the component τ_{11} of $\underset{\approx}{\tau}$, although the other components can be treated in exactly the same way.

We consider τ_{11} to be a functional defined on $\hat{\psi} \in \mathrm{L}^2(D)$ as follows:

$$\tau_{11}(\hat{\psi}) = \int_D F_1(\underset{\sim}{q})\, q_1 \sqrt{M(\underset{\sim}{q})}\, \hat{\psi}(\underset{\sim}{q},t)\, \mathrm{d}\underset{\sim}{q}, \tag{2.63}$$

whereby,

$$\begin{aligned}
|\tau_{11}(\hat{\psi})| &= \left|\int_D q_1^2\, U'(\tfrac{1}{2}|\underset{\sim}{q}|^2)\sqrt{M(\underset{\sim}{q})}\,\hat{\psi}\,\mathrm{d}\underset{\sim}{q}\right| \leq b\left(\int_D U'(\tfrac{1}{2}|\underset{\sim}{q}|^2)^2 M(\underset{\sim}{q})\,\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}}\|\hat{\psi}\| \\
&= \frac{b}{\sqrt{C}}\left(\int_D\left(1-|\underset{\sim}{q}|^2/b\right)^{\frac{b}{2}-2}\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}}\|\hat{\psi}\| = \frac{b}{\sqrt{C}}\left(2\pi b\int_0^1(1-r^2)^{\frac{b}{2}-2}\,r\,\mathrm{d}r\right)^{\frac{1}{2}}\|\hat{\psi}\| \\
&\leq \frac{b}{\sqrt{C}}\left(2^{\frac{b}{2}-1}\pi b\int_0^1(1-r)^{\frac{b}{2}-2}\,\mathrm{d}r\right)^{\frac{1}{2}}\|\hat{\psi}\|,
\end{aligned} \tag{2.64}$$

where C is the normalisation constant from (1.30). Hence, we require $b > 2$ so that $\tau_{11} \in \mathrm{L}^2(D)' = \mathrm{L}^2(D)$; this is the same condition that we assume for b throughout this thesis.

Applying τ_{11} to (2.62) gives:

$$\begin{aligned}
\tau_{11}(\hat{\psi}) &= \frac{b^2}{\sqrt{C}}\int_0^1\int_0^{2\pi}(1-r^2)^{\frac{b}{4}-1}r^3\cos^2(\theta)\,\tilde{\psi}(r,\theta,t)\,\mathrm{d}r\,\mathrm{d}\theta \\
&= \frac{\pi\, b^2}{\sqrt{C}}\int_0^1 r^3(1-r^2)^{\frac{b}{4}-1}\left(\tilde{\psi}_1(r,t)+\frac{r}{2}\left(\tilde{A}_1(r,t)\right)\right)\,\mathrm{d}r.
\end{aligned} \tag{2.65}$$

This shows that, quite remarkably, due to orthogonality with $\cos^2(\theta) = \frac{1}{2} + \frac{1}{2}\cos(2\theta)$ over $\theta \in (0, 2\pi)$, the functional τ_{11} filters out all but two terms of the infinite series in (2.62). The same filtering occurs for Galerkin spectral methods that use trigonometric polynomials in θ, such as method $\mathcal{A}$, method $\mathcal{B}$ or the method of Chauviére & Lozinski. We consider method $\mathcal{A}$ below, but the same approach could be applied to the other methods.

Suppose, using basis $\mathcal{A}$, that our numerical solution is defined as follows:

$$\tilde{\psi}_N(r,\theta) = (1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}_{0,k}P_k(r) + r(1-r)\sum_{i=0}^{1}\sum_{l=1}^{N_\theta}\sum_{k=0}^{N_r-1}\tilde{\Psi}^i_{l,k}P_k(r)\Phi_{il}(\theta).$$

Then, assuming $N_\theta \geq 1$, we have

$$\tau_{11}(\hat{\psi}_N) = \frac{\pi\, b^2}{\sqrt{C}}\int_0^1 r^3(1-r^2)^{\frac{b}{4}-1}\left[\left((1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}_{0,k}P_k(r)\right) + \frac{r}{2}\left((1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^0_{1,k}P_k(r)\right)\right]\,\mathrm{d}r.$$

It follows that

$$\begin{aligned}\tau_{11}(\hat{\psi}(t^n)) - \tau_{11}(\hat{\psi}^n_N) = \frac{\pi\, b^2}{\sqrt{C}}\int_0^1 r^3(1-r^2)^{\frac{b}{4}-1}\Bigg[&\left(\tilde{\psi}_1(r,t^n) - (1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^n_{0,k}P_k(r)\right)\\ &+\frac{1}{2}\left(r\tilde{A}_1(r,t^n) - r(1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^{0,n}_{1,k}P_k(r)\right)\Bigg]\,\mathrm{d}r. \qquad (2.66)\end{aligned}$$

Applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}|\tau_{11}(\hat{\psi}(t^n)) - \tau_{11}(\hat{\psi}^n_N)|^2 \;\leq\; & C_*\left\|\tilde{\psi}_1(r,t^n) - (1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^n_{0,k}P_k(r)\right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\\ &+\frac{C_*}{4}\left\|r\tilde{A}_1(r,t^n) - r(1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^{0,n}_{1,k}P_k(r)\right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}, \qquad (2.67)\end{aligned}$$

where,

$$C_* = \begin{cases}\frac{2\pi^2 b^4}{(b/2-1)\,C}, & 2<b<4\\ \frac{\pi^2\, b^4}{3\,C}, & b\geq 4\end{cases} \qquad (2.68)$$

and, as in Section 2.4, $\mathrm{L}^2_{\tilde{w}}(0,1)$ is the r-weighted L^2 space.

On the other hand, using Parseval's identity, we have

$$\begin{aligned}\|\hat{\psi}(\cdot,t^n) - \hat{\psi}^n_N(\cdot)\|^2_{\mathrm{L}^2(D)} \;=\;& b\int_0^1\int_0^{2\pi}|\tilde{\psi}(r,\theta,t^n) - \tilde{\psi}^n_N(r,\theta)|^2 r\,\mathrm{d}r\,\mathrm{d}\theta\\ =\;& 2\pi b\left\|\tilde{\psi}_1(r,t^n) - (1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^n_{0,k}P_k(r)\right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\\ &+\pi b\sum_{l=1}^{N_\theta}\left\|r\tilde{A}_l(r,t^n) - r(1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^{0,n}_{l,k}P_k(r)\right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\\ &+\pi b\sum_{l=1}^{N_\theta}\left\|r\tilde{B}_l(r,t^n) - r(1-r)\sum_{k=0}^{N_r-1}\tilde{\Psi}^{1,n}_{l,k}P_k(r)\right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\\ &+\pi b\sum_{l=N_\theta+1}^{\infty}\left(\left\|r\tilde{A}_l(r,t^n)\right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} + \left\|r\tilde{B}_l(r,t^n)\right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)}\right). \qquad (2.69)\end{aligned}$$

It follows that

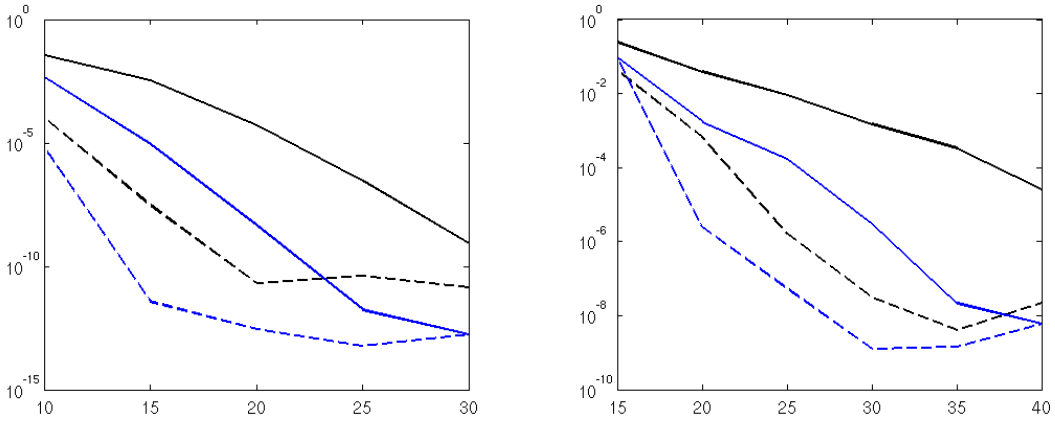
$$\|\tau_{11}(\tilde{\psi}) - \tau_{11}(\tilde{\psi}_N)\|_{\ell^\infty(0,T)} \leq \sqrt{\frac{C_*}{2\pi b}}\, \|\hat{\psi} - \hat{\psi}_N\|_{\ell^\infty(0,T;\mathrm{L}^2(D))}. \tag{2.70}$$

However, more importantly, we can see that the bound in (2.67) contains only two terms from the infinite sum in (2.69) (albeit with different constants) and therefore we expect that the error in τ_{11} will typically be much smaller than the error in $\hat{\psi}$.

In practical computations, this manifests as superconvergence of $\underset{\approx}{\tau}$. We demonstrate this superconvergence here by comparing the $\mathrm{L}^2(D)$ convergence data for $\hat{\psi}$ from Tables 2.1 and 2.2 with the corresponding errors in τ_{11}. These results are plotted in Figure 2.3 and we can clearly see that, prior to stagnation due to rounding error, τ_{11} converges at a faster rate than $\hat{\psi}$, and the error in τ_{11} is typically orders of magnitude smaller than the $\hat{\psi}$ error. This behaviour is extremely advantageous for micro-macro computations in where the accuracy of $\underset{\approx}{\tau}$ (rather than $\hat{\psi}$) is crucial.

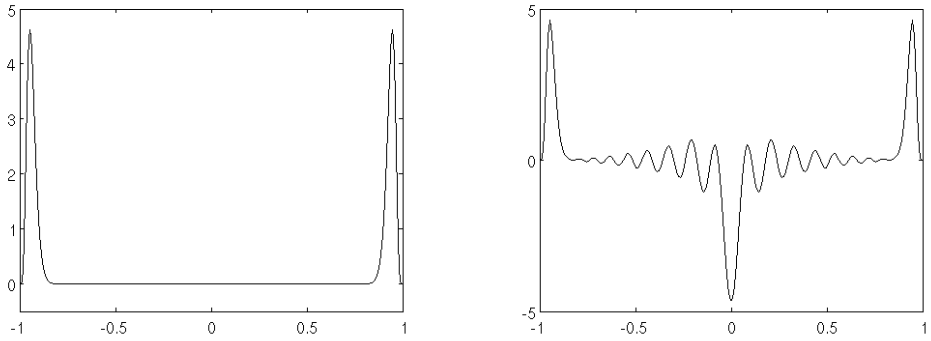
One interesting thing to note from Figure 2.3(b) is that the error in τ_{11} appears to stagnate at around 10^{-10} with both method $\mathcal{A}$ and method $\mathcal{B}$, and in fact, the error increases to some extent when the number of spectral basis functions is increased further (*e.g.* compare $N_r = N_\theta = 35$ to $N_r = N_\theta = 40$ in the plot); this increase in error is due to the fact that the condition number of the linear system (2.58) increases with N_r and N_θ and hence we can lose extra digits of accuracy for larger values of N_r, N_θ. Similarly, the condition number is larger for method $\mathcal{A}$ than for method $\mathcal{B}$ for the computations considered in Figure 2.3(a), which is why the error in τ_{11} for method $\mathcal{A}$ stagnates at around 10^{-11}, whereas the error from method $\mathcal{B}$ stagnates at 10^{-13}.

Remark 2.20 *It was proved in Lemma 2.7 that the weak solution of the initial-boundary-value problem* (2.1), (2.2), (2.3) *is non-negative a.e. on D. This property is not guaranteed to hold for the numerical solution. However, our numerical experiments consistently show that if there are sufficiently many modes in the approximation space to accurately resolve the solution then this non-negativity property is preserved under discretisation. This is illustrated in Figure 2.4 in which two cross-sections of the numerical solution for the $(b, \mathrm{Wi}, \delta) = (12, 1, 5)$ extensional flow are shown: the numerical solution on the left is fully resolved, while the one on the right is under-resolved. In the under-resolved case there are oscillations and clearly $\psi_N \geq 0$ is not satisfied throughout D, whereas the non-negativity property is accurately captured in the fully resolved case.* ◇



(a) (b)

Figure 2.3: Comparison of convergence of $\hat{\psi}$ and $\underset{\sim}{\tau}$. In both plots, the horizontal axis shows the value of N_r and N_θ (chosen to be equal in these computations). Plot (a) shows data for the computations considered in Table 2.1 and plot (b) corresponds to Table 2.2. In both (a) and (b), the solid black line represents the relative $\mathrm{L}^2(D)$ error in $\hat{\psi}$ for method $\mathcal{A}$, and the solid blue line represents the corresponding data for method $\mathcal{B}$. The dashed black line shows τ_{11} errors arising from method $\mathcal{A}$, and the dashed blue line is analogous for method $\mathcal{B}$.



(a) (b)

Figure 2.4: Cross-sections of the solution of the extensional flow problem with $b = 12$, $\mathrm{Wi} = 1$ and $\delta = 5$ at steady state, obtained using method $\mathcal{B}$. The fully-resolved solution in (a) was obtained using $(N_r, N_\theta) = (41, 20)$, and the under-resolved solution in (b) was obtained with $(N_r, N_\theta) = (26, 20)$.

2.6.2 The semi-implicit numerical method

Up until now we have confined our attention to the backward Euler temporal discretisation of the Fokker–Planck equation, as defined in (2.24). However, as we shall see, the semi-implicit discretisation, which is identical to (2.24) except that the term

$\int_D \underset{\approx}{\kappa} \underset{\sim}{q} \hat{\psi}_N \hat{\varphi} \, \mathrm{d}\underset{\sim}{q}$ is treated explicitly in time, is important in Chapter 3. Therefore, as a precursor to the next chapter, we consider this semi-implicit scheme here.

It should be noted that all of the analytical results that we obtained for the backward Euler temporal discretisation (also referred to from now on as the fully-implicit discretisation) in this chapter also carry across to the semi-implicit scheme – we do not consider the details here, but, for example, in the process of proving Lemma 3.7 in the next chapter, we establish a stability result for the semi-implicit scheme that is almost identical to Lemma 2.4. However, although the $\mathrm{L}^2(D)$ stability estimates (and therefore also the asymptotic convergence results) are essentially identical for the fully-implicit and semi-implicit schemes, we show in this section that for practical computations, the fully-implicit discretisation tends to be much more stable in the sense that solutions obtained from the semi-implicit scheme are more likely to exhibit the exponential growth in time in the $\mathrm{L}^2(D)$ norm that is allowed due to the constant $\mathrm{e}^{2c_0 m \Delta t}$ in Lemma 2.4. This is not surprising; it is well-known that fully-implicit schemes are generally more stable than semi-implicit and explicit schemes for parabolic and hyperbolic PDEs.

All the details of the implementation of the semi-implicit method carry over from the discussion of the fully-implicit method above; the only difference is that instead of (2.58), the linear system in this case is:

$$\left(\mathbf{M} + \frac{\Delta t}{2\mathrm{Wi}} \mathbf{S} \right) \tilde{\mathbf{\Psi}}^{n+1} = \left(\mathbf{M} + \Delta t \mathbf{C}^n \right) \tilde{\mathbf{\Psi}}^n. \tag{2.71}$$

We now present some numerical results that compare the fully-implicit and semi-implicit schemes. We only consider method $\mathcal{B}$ here, with the understanding that the behaviour for method $\mathcal{A}$ is essentially the same.

First of all, we repeated the computations in Table 2.1 (for an extensional flow with $(b, \mathrm{Wi}, \delta) = (12, 1, 1)$) using the semi-implicit scheme, and the results were identical to those reported in Table 2.1 for the backward Euler discretisation. However, on increasing the Weissenberg number from 1 to 5 we then observed significant differences between the two schemes. The results for the $\mathrm{Wi} = 5$ computations are summarised in Table 2.4.

We can see from the table that with $(N_r, N_\theta) = (15, 15)$, the semi-implicit scheme led to numerical solutions for which the $\mathrm{L}^2(D)$ norm error grew rapidly in time for all three time-step sizes, $\Delta t = 0.1,\ 0.05$ and 0.01 (indicated by hyphens in the table), whereas the fully-implicit scheme had an $\mathcal{O}(1)$ error in each of these cases. In the computations with $(N_r, N_\theta) = (20, 20)$, the fully-implicit method again performed better; we needed to take $\Delta t = 0.01$ in order to get an accurate solution with the semi-implicit scheme, whereas the fully-implicit scheme was accurate with $\Delta t = 0.1$. Finally, for

	$\Delta t = 0.1,\ N_T = 250$		$\Delta t = 0.05,\ N_T = 500$		$\Delta t = 0.01,\ N_T = 2500$	
(N_r, N_θ)	Imp.	Semi-Imp.	Imp.	Semi-Imp.	Imp.	Semi-Imp.
(15,15)	1.49	–	1.49	–	1.49	–
(20,20)	4.67×10^{-2}	–	4.67×10^{-2}	$1.14 \times 10^{+2}$	4.67×10^{-2}	4.67×10^{-2}
(25,25)	2.96×10^{-3}	–	2.96×10^{-3}	2.96×10^{-3}	2.96×10^{-3}	2.96×10^{-3}
(30,30)	1.44×10^{-4}	–	1.44×10^{-4}	1.44×10^{-4}	1.44×10^{-4}	1.44×10^{-4}

Table 2.4: This table shows the relative $\mathrm{L}^2(D)$ error, with respect to the exact steady-state solution, for the implicit and semi-implicit schemes (using method $\mathcal{B}$) applied to an extensional flow problem with $(b, \mathrm{Wi}, \delta) = (12, 5, 1)$. Three different time-step sizes were tested, and the total number of time-steps, N_T, was varied in order to ensure that $T = N_T \Delta t$ was the same in each case.

$(N_r, N_\theta) = (25, 25)$ and $(N_r, N_\theta) = (30, 30)$, the two schemes behaved identically for $\Delta t = 0.05$ and $\Delta t = 0.01$, but the fully-implicit scheme remained accurate for $\Delta t = 0.1$ whereas the semi-implicit scheme did not.

These observations indicate that the fully-implicit scheme is reliable for coarser spatial discretisations and larger Δt than the semi-implicit scheme. This is especially noticeable when the Weissenberg number is increased (recall that the two methods behaved identically for the extensional flow with $\mathrm{Wi} = 1$). Note also that scaling $\underset{\approx}{\kappa}$ has roughly the same effect as scaling Wi, *e.g.* the steady state solution (assuming it exists) depends on the product $\mathrm{Wi}\,\underset{\approx}{\kappa}$ and not on Wi or $\underset{\approx}{\kappa}$ separately.[7] Hence, based on the results in Table 2.4, we conclude that it is preferable to use the fully-implicit temporal discretisation for problems in which Wi or $|\underset{\approx}{\kappa}|$, or both, are large (compared to, say, 1).

2.6.3 Three dimensional implementation of the spectral method

We now consider the implementation of the spectral method developed in this chapter in the case $d = 3$. This is closely related to the two-dimensional case, the primary differences being that we now use the spherical coordinate change of variables:

$$\underset{\sim}{q} = (\sqrt{b} r \cos\theta \sin\phi\,,\ \sqrt{b} r \sin\theta \sin\phi\,,\ \sqrt{b} r \cos\phi), \quad (r, \theta, \phi) \in R := (0,1)\times(0,2\pi)\times(0,\pi),$$

instead of (2.31) and, following Chauviére & Lozinski [23], we choose each of our basis functions to be a product of a spherical harmonic in (θ, ϕ) and polynomial in r. Discretisations of this type have also been considered in the recent paper by Guo and Huang [39]. Note that in this section, $\tilde{g}(r, \theta, \phi) := g(q_1, q_2, q_3)$.

[7] This can be seen by scaling Wi and $\underset{\approx}{\kappa}$ in (2.1) and noting that $\frac{\partial \psi}{\partial t}$ vanishes at steady state.

First of all, we redefine the space $\tilde{\mathrm{H}}^1(R)$ for the purposes of this section, in order to ensure that if $g \in \mathrm{H}^1(D)$, $D \subset \mathbb{R}^3$ then $\tilde{g} \in \tilde{\mathrm{H}}^1(R)$. Following the approach in the $d = 2$ case, we define $\|\tilde{g}\|^2_{\tilde{\mathrm{H}}^1(R)}$ by transforming $\|g\|^2_{\mathrm{H}^1(D)}$ from cartesian to spherical coordinates, and hence we have

$$\|\tilde{g}\|^2_{\tilde{\mathrm{H}}^1(R)} := \int_R r^2 \sin\phi \left(|\tilde{g}|^2 + \left|\frac{\partial \tilde{g}}{\partial r}\right|^2 + \frac{1}{r^2}\left|\frac{\partial \tilde{g}}{\partial \phi}\right|^2 + \frac{1}{r^2 \sin^2\phi}\left|\frac{\partial \tilde{g}}{\partial \theta}\right|^2 \right) \mathrm{d}r\, \mathrm{d}\theta\, \mathrm{d}\phi,$$

and,

$$\tilde{\mathrm{H}}^1(R) := \{\tilde{f} \in \mathrm{L}^2_{\mathrm{loc}}(R) \ : \ \tilde{f}(r,\cdot,\phi) \in \mathrm{H}^1_p(0,2\pi) \text{ for a.e. } (r,\phi) \in (0,1)\times(0,\pi) \text{ and } \|\tilde{f}\|_{\tilde{\mathrm{H}}^1(R)} < \infty\}.$$

We denote the spherical harmonics by $S_{l,m} : (\theta,\phi) \mapsto S_{l,m}(\theta,\phi) \in \mathbb{R}$. They are the solutions of the equation

$$\frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\,\frac{\partial}{\partial\phi}S_{l,m}(\theta,\phi)\right) + \frac{1}{\sin^2\phi}\frac{\partial^2}{\partial\theta^2}S_{l,m}(\theta,\phi) + l(l+1)S_{l,m}(\theta,\phi) = 0, \quad (2.72)$$

for a.e. $(\theta,\phi) \in (0,2\pi)\times(0,\pi)$, where (2.72) is the angular part of Laplace's equation in spherical coordinates. It can be shown, by separation of variables, that the solutions of (2.72) are of the form,

$$S_{l,m}(\theta,\phi) = C(l,m)\, P_l^m(\cos\phi) e^{im\theta}, \quad (2.73)$$

for $l \in \mathbb{Z}_{\geq 0}$, $|m| \leq l$, where P_l^m denotes an associated Legendre function and $C(l,m)$ is a normalisation constant. Also, the (appropriately normalised) spherical harmonics satisfy the following orthogonality property:

$$\int_0^{2\pi}\int_0^{\pi} S_{l_1,m_1}(\theta,\phi)\,\overline{S}_{l_2,m_2}(\theta,\phi)\sin\phi\,\mathrm{d}\theta\,\mathrm{d}\phi = \delta_{m_1,m_2}\delta_{l_1,l_2}, \quad (2.74)$$

where the overline notation denotes complex conjugation.

The next lemma shall motivate our definition of a spectral basis in the $d = 3$ case.

Lemma 2.21 *Let* $\tilde{g}(r,\theta) = \sum_{l=0}^{N_{\mathrm{sph}}}\sum_{|m|\leq l}\tilde{\gamma}_l^m(r)S_{l,m}(\theta,\phi)$, $N_{\mathrm{sph}} \in \mathbb{Z}_{\geq 0}$, $\tilde{\gamma}_0^0 \in \mathrm{H}^1_{r^2}(0,1)$ *where* $\mathrm{H}^1_{r^2}(0,1)$ *is the* r^2*-weighted* H^1*-space, and*

$$\tilde{\gamma}_l^m \in \mathrm{H}^1(0,1;1,r^2) := \left\{\tilde{f} \in \mathrm{H}^1_{\mathrm{loc}}(0,1) \ : \ \int_0^1 \left(|\tilde{f}(r)|^2 + r^2|\tilde{f}'(r)|^2\right)\mathrm{d}r < \infty\right\},$$

for $l > 0$. *Then* $\tilde{g} \in \tilde{\mathrm{H}}^1(R)$.

Proof. Periodicity of $\tilde{g}$ in θ follows directly from the definition of the spherical harmonics, hence it only remains to verify that $\|\tilde{g}\|_{\tilde{\mathrm{H}}^1(R)} < \infty$.

Integrating by parts in θ and ϕ (which is valid for spherical harmonics), we obtain:

$$\begin{aligned}\|\tilde{g}\|^2_{\tilde{\mathrm{H}}^1(R)} &= \int_R r^2 \sin\phi \left(|\tilde{g}|^2 + \left|\frac{\partial \tilde{g}}{\partial r}\right|^2 \right) \mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi \\ &\quad - \int_R \sin\phi\,\tilde{g} \left(\frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial\tilde{g}}{\partial\phi}\right) + \frac{1}{\sin^2\phi}\frac{\partial^2\tilde{g}}{\partial\theta^2}\right) \mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi, \quad (2.75)\end{aligned}$$

where the boundary conditions vanish due to periodicity. Substituting the series expression of $\tilde{g}$ into (2.75) and using (2.72) and (2.74), we get:

$$\begin{aligned}\|\tilde{g}\|^2_{\tilde{\mathrm{H}}^1(R)} &= \sum_{l=0}^{N_{\mathrm{sph}}} \sum_{|m|\le l} \int_0^1 r^2 \left\{ |\tilde{\gamma}_l^m(r)|^2\,\mathrm{d}r + \left|\frac{\mathrm{d}\tilde{\gamma}_l^m}{\mathrm{d}r}\right|^2 \right\} \mathrm{d}r \\ &+ \int_R \left\{ \sum_{l_1=0}^{N_{\mathrm{sph}}} \sum_{|m_1|\le l_1} \tilde{\gamma}_{l_1}^{m_1}(r) S_{l_1,m_1}(\theta,\phi) \right\} \left\{ \sum_{l_2=0}^{N_{\mathrm{sph}}} \sum_{|m_2|\le l_2} l_2(l_2+1)\,\tilde{\gamma}_{l_2}^{m_2}(r)\overline{S}_{l_2,m_2}(\theta,\phi) \right\} \sin\phi\,\mathrm{d}\phi\,\mathrm{d}\theta\,\mathrm{d}r \\ &= \sum_{l=0}^{N_{\mathrm{sph}}} \sum_{|m|\le l} \int_0^1 \left\{ r^2|\tilde{\gamma}_l^m(r)|^2 + r^2\left|\frac{\mathrm{d}}{\mathrm{d}r}\tilde{\gamma}_l^m(r)\right|^2 + l(l+1)|\tilde{\gamma}_l^m(r)|^2 \right\} \mathrm{d}r. \quad (2.76)\end{aligned}$$

By the hypotheses on the $\tilde{\gamma}_l^m$, it follows that $\|\tilde{g}\|_{\tilde{\mathrm{H}}^1(R)}$ is finite. $\square$

Note that the $\tilde{\gamma}_l^m$ in Lemma 2.21 need not be bounded on $(0,1)$ since, for example, $r^{-1/4} \in \mathrm{H}^1_{r^2}(0,1) \cap \mathrm{H}^1(0,1;1,r^2)$.

It will be convenient from now on to use the real and imaginary parts of the spherical harmonics rather than the complex exponentials in (2.73), *i.e.*:

$$S^i_{l,m}(\theta,\phi) := C(l,m)\,P_l^m(\cos\phi)((1-i)\cos(m\theta) + i\sin(m\theta)), \quad (2.77)$$

where now $0 \le l \le N_{\mathrm{sph}}$, $i \in \{0,1\}$, and $i \le m \le l$. In this section, we consider basis functions of the following form:

$$Y^{ik}_{lm}(r,\theta,\phi) := (1-r)Q_k(r)S^i_{l,m}(\theta,\phi), \quad (2.78)$$

where $(1-r)Q_k \in \mathbb{P}_{N_r,0}(0,1)$ (as in the $d=2$ case, Q_k is taken to be a Chebyshev polynomial of degree k, $0 \le k \le N_r - 1$, mapped from $[-1,1]$ to $[0,1]$, although other polynomial choices could be considered also). Since $\mathbb{P}_{N_r,0}(0,1) \subset \mathrm{H}^1_{r^2}(0,1) \cap \mathrm{H}^1(0,1;1,r^2)$, it follows from Lemma 2.21, that any finite linear combination of basis functions of the form (2.78) is contained in $\tilde{\mathrm{H}}^1_0(R)$. This is a simpler situation than in two dimensions, since now we do not need to impose a specialised decomposition in order to guarantee inclusion in $\tilde{\mathrm{H}}^1(R)$.

Below we will introduce a basis on which our Galerkin spectral method in three dimensions will be based on. Before defining this basis, however, we first consider the symmetry property discussed in Remark 2.19 in the $d = 3$ case. In fact, most of Remark 2.19 carries over to three dimensions unchanged; the only difference is that now the evenness of $\hat{\psi}$ in the D domain with respect to $\underset{\sim}{q}$ translates to requiring that we only use spherical harmonics in R for which l is an even number. This can be seen by the following argument. Suppose, using the change of variables to spherical coordinates, that $\underset{\sim}{q} \mapsto (r, \theta, \phi)$. Then also $-\underset{\sim}{q} \mapsto (r, \theta + \pi, \pi - \phi)$. Now, the symmetry condition we wish to impose is that for any basis function Y_{lm}^{ik} defined in (2.78), we have $Y_{lm}^{ik}(r, \theta, \phi) = Y_{lm}^{ik}(r, \theta + \pi, \pi - \phi)$. This, in turn, requires that $S_{l,m}^{i}(\theta, \phi) = S_{l,m}^{i}(\theta + \pi, \pi - \phi)$. Noting that,

$$S_{l,m}^{i}(\theta, \phi) = P_l^m(\cos\phi)((1 - i)\cos(m\theta) + i\sin(m\theta)),$$

and

$$S_{l,m}^{i}(\theta + \pi, \pi - \phi) = (-1)^m P_l^m(-\cos\phi)((1 - i)\cos(m\theta) + i\sin(m\theta)),$$

it follows that we can only use associated Legendre functions for which $P_l^m(x) = (-1)^m P_l^m(-x)$, for all $x \in [-1, 1]$. Since the associated Legendre functions are defined as,

$$P_l^m(x) = (-1)^m (1 - x^2)^{m/2} \frac{\mathrm{d}^m}{\mathrm{d}x^m}(P_l(x)),$$

where $P_l(x)$ is a Legendre polynomial of degree l (for which $P_l(x) = (-1)^l P_l(-x)$), it follows that the required symmetry condition is satisfied if, and only if, l is an even number (for any $m = 0, \ldots, l$).

Remark 2.22 *In [23], Chauviére & Lozinski restricted their attention to two dimensional macroscopic velocity fields, in which case a more restrictive symmetry condition was appropriate,* i.e. *that $\psi(r, \theta, \phi) = \psi(r, \theta + \pi, \phi)$, and hence they only considered spherical harmonics for which both l and m were even numbers. Compared to the more general symmetry condition considered above, the condition of Chauviére & Lozinski leads to a reduction in computational effort because for a given N_{sph}, fewer basis functions are used since the spherical harmonics with odd m are discarded, and also it is only necessary to consider $\theta \in (0, \pi)$. In this thesis, however, we are interested in treating the case in which the macroscopic velocity field can be three dimensional, and therefore we require the symmetry condition for Y_{lm}^{ik} identified above.*

With the considerations discussed above in mind, we can now define a basis, denoted $\mathcal{C}$, as follows:

$$\mathcal{C} := \{Y_{lm}^{ik} \ : \ 0 \leq k \leq N_r - 1, \ i \in \{0, 1\}, \ l \in \{0, 2, 4, \ldots, N_{\mathrm{sph}}\} \text{ and } i \leq m \leq l\}.$$

From now on, the numerical method that uses basis $\mathcal{C}$ will be referred to as method $\mathcal{C}$.

At this point, we could take a detour to consider three dimensional approximation results for $\text{span}(\mathcal{C}) \subset \tilde{\text{H}}^1_0(R)$, which would then allow us to extend our convergence results from Section 2.5 to the $d = 3$ case. However, given that we have already considered approximation results in detail for $d = 2$, and given that the approach in the $d = 3$ case would be completely analogous, for the sake of brevity, we omit discussion of approximation theory in three dimensions here. Note, however, that Guo & Huang [39] recently derived approximation results for a spectral method on the unit ball in $\mathbb{R}^3$, which could be applied to the convergence analysis of method $\mathcal{C}$ (*e.g.* see Theorem 2.3 in that paper, which is similar to our approximation result (2.46)).

Below we will test the performance of method $\mathcal{C}$ on some model problems. First of all, however, we specify the spherical coordinate form of the discretisation matrices defined in (2.54). Using the same notation that we used for the discretisation matrices in polar coordinates, we let N denote the total number of basis functions, we set $\tilde{\psi}_N^{n+1}(r,\theta,\phi) = \sum_{v=1}^N \tilde{\Psi}_v^{n+1} Y_v(r,\theta,\phi)$, where Y_v is a basis function from $\mathcal{C}$ for $1 \leq v \leq N$, and we denote the test functions by Y_u for $1 \leq u \leq N$. Then,

$$(M_q)_{uv} = \int_0^1 \int_0^{2\pi} \int_0^{\pi} b^{3/2}\, Y_u Y_v\, r^2 \sin\phi \,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi, \tag{2.79}$$

$$\begin{aligned}(S_q)_{uv} = \int_0^1 \int_0^{2\pi} \int_0^{\pi} \Bigg\{ & b^{1/2} \frac{\partial Y_u}{\partial r}\frac{\partial Y_v}{\partial r} r^2 \sin\phi + b^{1/2}\frac{1}{\sin\phi}\frac{\partial Y_u}{\partial \theta}\frac{\partial Y_v}{\partial \theta} + b^{1/2}\frac{\partial Y_u}{\partial \phi}\frac{\partial Y_v}{\partial \phi}\sin\phi \\ & + \frac{b^{3/2}}{2} r^3 \sin\phi \left(1-r^2\right)^{-1}\left[\frac{\partial Y_u}{\partial r} Y_v + Y_u \frac{\partial Y_v}{\partial r}\right] + \frac{b^{5/2}}{4} r^4 \sin\phi \left(1-r^2\right)^{-2} Y_u Y_v \Bigg\} \,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi,\end{aligned} \tag{2.80}$$

$$\begin{aligned}(C_q^m)_{uv} = \int_0^1 \int_0^{2\pi} \int_0^{\pi} Y_v \Bigg\{ & k_r \left[b^{3/2} r^3 \frac{\partial Y_u}{\partial r} + \frac{b^{5/2}}{2} r^4 \left(1-r^2\right)^{-1} Y_u \right] \\ & + k_\theta\, b^{3/2} r^2 \frac{\partial Y_u}{\partial \theta} + k_\phi\, b^{3/2} r^2 \sin\phi \frac{\partial Y_u}{\partial \phi} \Bigg\} \,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\phi,\end{aligned} \tag{2.81}$$

where $k_r = (\underset{\approx}{\kappa}(\underset{\sim}{x}_m)\underset{\sim}{e}_r)\cdot\underset{\sim}{e}_r$, $k_\theta = (\underset{\approx}{\kappa}(\underset{\sim}{x}_m)\underset{\sim}{e}_r)\cdot\underset{\sim}{e}_\theta$ and $k_\phi = (\underset{\approx}{\kappa}(\underset{\sim}{x}_m)\underset{\sim}{e}_r)\cdot\underset{\sim}{e}_\phi$, with $\underset{\sim}{e}_r,\ \underset{\sim}{e}_\theta,\ \underset{\sim}{e}_\phi$ the unit vectors in the r, θ and ϕ directions:

$$\begin{aligned}\underset{\sim}{e}_r &= (\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi),\\ \underset{\sim}{e}_\theta &= (-\sin\theta, \cos\theta, 0),\\ \underset{\sim}{e}_\phi &= (\cos\theta\cos\phi, \sin\theta\cos\phi, -\sin\phi).\end{aligned}$$

Note that $\underset{\approx}{\kappa}\underset{\sim}{q} = \sqrt{b}\, r\ (k_r \underset{\sim}{e}_r + k_\theta \underset{\sim}{e}_\theta + k_\phi \underset{\sim}{e}_\phi)$, and $(\underset{\sim}{e}_r, \underset{\sim}{e}_\theta, \underset{\sim}{e}_\phi)$ is an orthonormal basis for $\mathbb{R}^3$ for any $(\theta,\phi) \in (0,2\pi)\times(0,\pi)$. We refer to Section 2.6.1 for the details of computing

the discretisation matrices and the solution of the resulting linear system; the approach is completely analogous here.

Next, we present some computational results for method $\mathcal{C}$. We consider the backward Euler temporal discretisation of the FENE Fokker–Planck equation here, as opposed to the semi-implicit scheme considered in Section 2.6.2, and we restrict our attention to producing plots of the same type as in Figure 2.3 in order to visualise the convergence rates for $\hat{\psi}$ and $\underset{\approx}{\tau}$, and also to verify that we obtain superconvergence of $\underset{\approx}{\tau}$ in the $d = 3$ case. Note that theoretical underpinning of the superconvergence of $\underset{\approx}{\tau}$ characterised in (2.67) and (2.69) can also be established in 3-dimensions; the reasoning is the same, except that we use Parseval's identity based on spherical harmonics, as in Lemma 2.21.

As in the two dimensional case, we know the exact steady state solution for problems in which $\underset{\approx}{\kappa}$ is a symmetric 3×3 tensor (*cf.* (2.59)). We now consider two distinct problems; for each problem we have $\underset{\approx}{\kappa} = \underset{\approx}{\kappa}^T$ so that we can compare the numerical solution with the exact steady state solution, and as in Tables 2.1 and 2.2 we take 2000 time-steps with $\Delta t = 0.05$ to obtain an accurate approximation to the steady state solution.

The first problem we consider is a three dimensional extensional flow with $b = 12$, $\mathrm{Wi} = 1$ and $\underset{\approx}{\kappa}$ defined as follows:

$$\underset{\approx}{\kappa} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1/2 & 0 \\ 0 & 0 & -1/2 \end{pmatrix}. \tag{2.82}$$

Figure 2.5(a) shows the convergence plots for $\hat{\psi}$ and τ_{11} for this problem. It is clear from the figure that we obtain spectral convergence of $\hat{\psi}$, and also, just as in Figure 2.3, we observe superconvergence of τ_{11}.

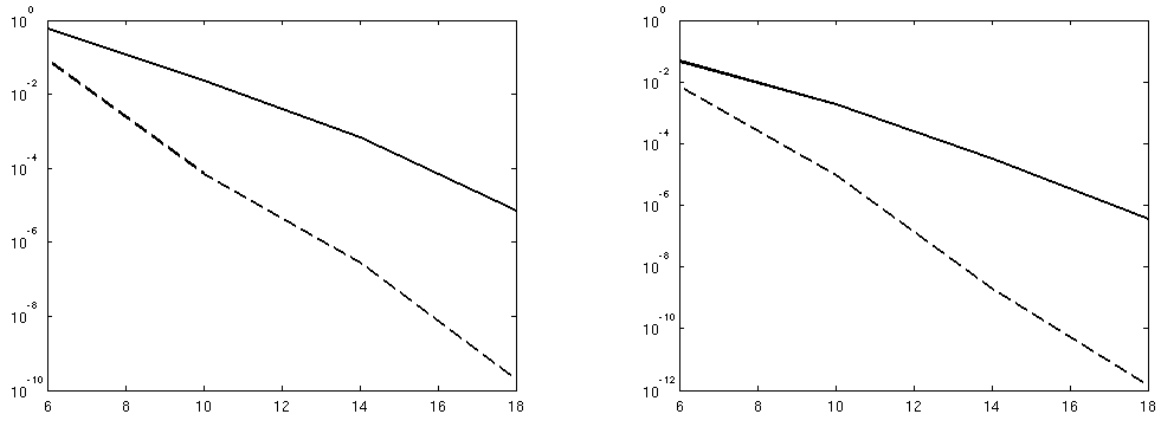
Next, we consider a problem in which $\underset{\approx}{\kappa}$ is a full tensor:

$$\underset{\approx}{\kappa} = \begin{pmatrix} 0.5 & 0.2 & 0.5 \\ 0.2 & -0.25 & -0.4 \\ 0.5 & -0.4 & -0.25 \end{pmatrix}, \tag{2.83}$$

and where $b = 12$ and $\mathrm{Wi} = 1$ again. The convergence plot for this computation is shown in Figure 2.5(b), and the behaviour is much the same as in Figure 2.5(a).

2.7 Conclusions

The purpose of this chapter has been to develop a rigorous foundation for the numerical approximation of Fokker–Planck equations. We restricted our attention to the



(a) (b)

Figure 2.5: Comparison of convergence of $\hat{\psi}$ and $\underset{\approx}{\tau}$ for method $\mathcal{C}$ for two different problems (we compared to the exact steady state solution, (2.59), by taking 2000 time-steps with $\Delta t = 0.05$). Plot (a) corresponds to a three dimensional extensional flow problem with $b = 12$ and $\mathrm{Wi} = 1$ and with $\underset{\approx}{\kappa}$ defined in (2.82). Plot (b) is analogous, except that in this case $\underset{\approx}{\kappa}$ is as in (2.83). In both plots, the horizontal axis represents N_r and N_{sph} (chosen to be equal in these computations), and the solid and dashed lines show the relative $\mathrm{L}^2(D)$ error and relative τ_{11} error, respectively.

configuration space part of (1.44), but the work in this chapter will be built upon in subsequent chapters in order to develop numerical methods on $\Omega \times D$.

We focused on the symmetrised weak formulation of the Maxwellian-transformed equation, and we used the substitution $\hat{\psi} = \psi/\sqrt{M}$. The resulting formulation (2.6) facilitated the development of a number of analytical results in Sections 2.2 and 2.3. Using the approximation results derived in Section 2.4, optimal-order convergence of the fully-discrete spectral Galerkin method (2.24), (2.25) was established for the case of $d = 2$; an analogous procedure could be carried out for $d = 3$. This analysis was performed for spring potentials that satisfy Hypotheses A and B; see Example 2.1.

In the case of the FENE model, we indicated the extension of our analysis to a class of numerical methods based on another change of variable, proposed by Chauvière & Lozinski; here a different transformation, (2.60), is applied to the Fokker–Planck equation. We showed that, at the analytical level at least, the two approaches lead to methods with very similar stability and accuracy properties.

Section 2.6 addressed issues related to the implementation of numerical methods for the FENE Fokker–Planck equation. In Section 2.6.1 we considered two distinct implementations, methods $\mathcal{A}$ and $\mathcal{B}$, for the $d = 2$ case, and these methods were also compared to the spectral method discussed in the paper of Chauviére & Lozinski [24]

on the basis of numerical results reported therein. We showed that methods $\mathcal{A}$ and $\mathcal{B}$ work well for values of b up to about 20, and are comparable to the method formulated in [24] in terms of computational efficiency in this parameter range, with method $\mathcal{B}$ being more accurate than method $\mathcal{A}$, and of a very similar accuracy as the method in [24]. Also, we demonstrated that the convergence of $\underset{\approx}{\tau}$ tends to be much more rapid than the convergence of $\hat{\psi}$ using our Galerkin spectral methods; this is highly advantageous in the context of the micro-macro computations. In Section 2.6.3 we considered the implementation of the Galerkin spectral method, based on the symmetrised formulation, in three spatial dimensions. We constructed a $\tilde{\mathrm{H}}^1(R)$-conforming spectral basis, $\mathcal{C}$, and demonstrated that the convergence properties of the spectral method based on $\mathcal{C}$ are essentially the same as for the two dimensional spectral methods considered in Section 2.6.1.

The numerical methods and analytical results developed in this chapter are built upon in Chapter 3, where we consider the Fokker–Planck equation on $\Omega \times D$.

Chapter 3

Alternating-Direction Methods for the Full Fokker–Planck Equation

3.1 Introduction

In this chapter, we develop numerical methods for the Maxwellian-transformed Fokker–Planck equation posed on $\Omega \times D \times (0,T]$:

$$\frac{\partial \psi}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \psi + \underset{\sim}{\nabla}_q \cdot (\underset{\approx}{\kappa}\, \underset{\sim}{q}\, \psi) = \frac{1}{2\mathrm{Wi}}\, \underset{\sim}{\nabla}_q \cdot \left(M \underset{\sim}{\nabla}_q \frac{\psi}{M} \right), \ (\underset{\sim}{x}, \underset{\sim}{q}, t) \in \Omega \times D \times (0,T], \quad (3.1)$$

$$\psi(\underset{\sim}{x}, \underset{\sim}{q}, 0) = \psi_0(\underset{\sim}{x}, \underset{\sim}{q}), \qquad (\underset{\sim}{x}, \underset{\sim}{q}) \in \Omega \times D. \quad (3.2)$$

Throughout this chapter we assume that $\underset{\sim}{u} : (\underset{\sim}{x}, t) \in \Omega \times (0,T] \mapsto \underset{\sim}{u}(\underset{\sim}{x}, t) \in \mathbb{R}^d$ is an *a priori* defined vector field (hence $\underset{\approx}{\kappa} = \underset{\sim}{\nabla}_x \underset{\sim}{u}$ is known *a priori* also). The precise hypotheses on $\underset{\sim}{u}$ and $\underset{\approx}{\kappa}$ shall be specified below.

The above equation will be referred to as the *full* Fokker–Planck equation, to distinguish it from the equation posed on $D \times (0,T]$ only, that was studied in Chapter 2. From now on, we focus on the Maxwellian-transformed form of the Fokker–Planck equation given above (and its weak formulation in which the prinicpal part of the differential operator is symmetric). However, it should be noted that the numerical methods developed and analysed in the forthcoming sections could just as well be based on the Chauvière–Lozinski-transformed equation that was studied in Section 2.2.1, and was also used to solve the full FENE Fokker–Planck equation in [23, 24, 60].

As discussed in Chapter 1, due to the cartesian product structure of the domain $\Omega \times D$, a natural approach to solving (3.1), (3.2) is to use an operator splitting/alternating-direction approach, *cf.* (1.50), (1.51). This is the approach that we pursue in this chapter. The Galerkin spectral method on D that was developed in Chapter 2 will be used to solve (1.50), and a finite element method for (1.51) will also be introduced.

A finite element method is convenient for the $\underset{\sim}{x}$-direction solver because the physical space domain, Ω, need not have simple geometry. As in Chapter 2, all of the analysis in this chapter is valid for any spring potential that satisfies Hypotheses A and B, but in the computational results section we consider the FENE model only.

We propose a fully-practical alternating-direction Galerkin method for (3.1). The approach is similar in spirit to the alternating-direction method used by Chauvière & Lozinski in [23,24,60]. However, there are some important theoretical questions related to applying alternating-direction methods in this context, which have not previously been addressed in the literature, and we focus on these questions in this chapter. In particular, we consider the stability and convergence analysis of our alternating-direction scheme for (3.1) in Sections 3.4, 3.5, 3.6 and 3.7. It is not obvious *a priori* what effect applying a splitting of the form (1.50), (1.51) will have on a discretisation of (3.1), and therefore it is important to rigorously establish the stability and convergence properties of the alternating-direction numerical methods developed here.

The reader will note that the alternating-direction method under consideration here is nonstandard in the sense we consider d-dimensional cross-sections (rather than one-dimensional cross-sections) of $\Omega \times D$. This poses a formidable computational challenge because, as shall be seen in Section 3.3, we typically need to solve a large number problems posed in d spatial dimensions in each time-step. However, the method is extremely well suited to implementation on a parallel architecture since the $\underset{\sim}{q}$-direction solves are completely independent from one another, and similarly the $\underset{\sim}{x}$-direction solves are decoupled also. We discuss the parallel implementation of our alternating-direction scheme in Section 3.8, and our computational results in Section 3.9 were obtained using this parallel implementation.

The structure of this chapter is as follows. The weak formulation of the full Fokker–Planck equation is discussed in Section 3.2. We then introduce a quadrature-based alternating-direction procedure in Section 3.3 and derive stability results for this scheme in Section 3.4. Using the approximation results in Section 3.6, we then derive convergence estimates in Section 3.7. The implementation of the numerical method is described in Section 3.8, and in Section 3.9, numerical results for the FENE Fokker–Planck equation are presented in the simplified case that the macroscopic velocity, $\underset{\sim}{u}$, is taken to be a constant-in-time vector field.

3.2 Weak formulation and spatial discretisation

The full Fokker–Planck equation considered in this chapter depends on $\underset{\sim}{x} \in \Omega$ as well as $\underset{\sim}{q} \in D$, and therefore we will require the use of slightly different function spaces than in Chapter 2. Let $\mathrm{L}^2(\Omega \times D)$ be defined in the obvious way, and let $(\cdot, \cdot)$ and $\|\cdot\|$ denote the L^2 inner-product and norm over $\Omega \times D$:

$$(f, g) := \int_{\Omega \times D} f(\underset{\sim}{x}, \underset{\sim}{q}) g(\underset{\sim}{x}, \underset{\sim}{q}) \, \mathrm{d}\underset{\sim}{x} \, \mathrm{d}\underset{\sim}{q} \qquad \text{and} \qquad \|f\|^2 := (f, f).$$

We assume throughout this chapter that $\underset{\sim}{u}$ is a divergence-free d-component vector function, *i.e.*

$$\underset{\sim}{\nabla}_x \cdot \underset{\sim}{u}(\underset{\sim}{x}, t) = 0 \qquad \text{for a.e. } (\underset{\sim}{x}, t) \in \Omega \times (0, T]. \tag{3.3}$$

It would be straightforward to adapt the arguments in this chapter to the case where $\underset{\sim}{u}$ is not divergence free, but this would make the analysis more messy and it would shed no further light on the properties of the numerical methods under consideration. Therefore in the interests of clarity and brevity, in this chapter we restrict our attention to the case when (3.3) is satisfied.

Also, we suppose that

$$\underset{\sim}{u} \in \mathrm{L}^\infty(0, T; \mathrm{L}^\infty(\Omega)) \quad \text{and} \quad \underset{\sim}{\nabla}_x \underset{\sim}{u} = \underset{\approx}{\kappa} \in \mathrm{W}^{1,\infty}(0, T; \mathrm{L}^\infty(\Omega)), \tag{3.4}$$

where, to simplify notation, we do not explicitly label the d or $d \times d$ dimensionality of the function spaces for $\underset{\sim}{u}(\underset{\sim}{x}, t) \in \mathbb{R}^d$ and $\underset{\approx}{\kappa}(\underset{\sim}{x}, t) \in \mathbb{R}^{d \times d}$. The assumption in (3.4) for $\underset{\approx}{\kappa}$ is stronger than the assumptions in Chapter 2; recall that in Lemma 2.4 and Theorem 2.5 we required $\underset{\approx}{\kappa} \in \mathrm{C}[0, T]$ and in Lemma 2.10 we required $\underset{\approx}{\kappa} \in \mathrm{H}^1(0, T)$.

We shall also use the following space:

$$\mathcal{X} := \Big\{ \varphi \in \mathrm{L}^2(\Omega \times D) : \varphi \in \mathrm{L}^2(\Omega; \mathrm{H}^1_0(D; M)) \cap \mathrm{H}^1(\Omega; \mathrm{L}^2(D)) \Big\},$$

equipped with the following norm:

$$\|\varphi\|_{\mathcal{X}} := \left\{ \int_{\Omega \times D} \left(|\varphi|^2 + |\underset{\sim}{\nabla}_M \varphi|^2 \right) \mathrm{d}\underset{\sim}{x} \, \mathrm{d}\underset{\sim}{q} \right\}^{\frac{1}{2}}.$$

Employing the substitution $\hat{\psi} = \psi / \sqrt{M}$ that was used in Chapter 2, the weak formulation of (3.1) is as follows: Given $\hat{\psi}_0 \in \mathrm{L}^2(\Omega \times D)$, find $\hat{\psi} \in \mathrm{L}^\infty(0, T; \mathrm{L}^2(\Omega \times D)) \cap \mathrm{L}^2(0, T; \mathcal{X})$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\hat{\psi}, \zeta) + \Big(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi} \, , \, \zeta \Big) - \Big(\underset{\approx}{\kappa} \underset{\sim}{q} \hat{\psi} \, , \, \underset{\sim}{\nabla}_M \zeta \Big) + \frac{1}{2\mathrm{Wi}} \Big(\underset{\sim}{\nabla}_M \hat{\psi} \, , \, \underset{\sim}{\nabla}_M \zeta \Big) = 0 \quad \forall \, \zeta \in \mathcal{X}, \tag{3.5}$$

$$\hat{\psi}(\underset{\sim}{x}, \underset{\sim}{q}, 0) = \hat{\psi}_0(\underset{\sim}{x}, \underset{\sim}{q}), \qquad (\underset{\sim}{x}, \underset{\sim}{q}) \in \Omega \times D, \tag{3.6}$$

in the sense of distributions on $(0,T)$. Following Chapter 2, we impose zero Dirichlet boundary conditions on $\Omega \times \partial D$ for $t \in (0,T]$. For simplicity, we avoid boundary conditions on $\partial\Omega \times D$ by assuming that the macroscopic velocity field is an *enclosed flow*, *i.e.* that

$$\underset{\sim}{u} \cdot \underset{\sim}{n} = 0 \text{ on } \partial\Omega, \tag{3.7}$$

where $\underset{\sim}{n} \in \mathbb{R}^d$ is the unit outward normal for Ω. Also, the initial condition (3.6) is understood to be imposed in a weak sense and, as in Chapter 2, ψ is recovered by multiplying $\hat{\psi}$ by $\sqrt{M}$.

The term containing $\underset{\approx}{\kappa}$ in (3.5) will be of particular interest since, as we shall see, it is the most difficult term to treat using an alternating-direction method. We introduce the following bilinear form notation for this term, which will be convenient later on:

$$C(\underset{\approx}{\kappa}; f, g) := \left(\underset{\approx}{\kappa}\, \underset{\sim}{q}\, f \,,\, \underset{\sim}{\nabla}_M g \right). \tag{3.8}$$

Next, we establish a statement analogous to Lemma 1.3 for the weak solution of (3.5). Recall that

$$\varrho(\underset{\sim}{x}, t) := \int_D \psi(\underset{\sim}{x}, \underset{\sim}{q}, t)\, \mathrm{d}\underset{\sim}{q} = \int_D \sqrt{M(\underset{\sim}{q})}\, \hat{\psi}(\underset{\sim}{x}, \underset{\sim}{q}, t)\, \mathrm{d}\underset{\sim}{q}.$$

Noting from Hypothesis B in Chapter 2 that $\sqrt{M} \in \mathrm{H}^1_0(D) \subset \mathrm{H}^1_0(D; M)$, we set $\zeta = \sqrt{M}$ in (3.5), to obtain

$$\left(\frac{\partial \hat{\psi}}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}, \sqrt{M} \right) = \left(\frac{\partial \psi}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \psi, 1 \right) = \int_\Omega \left(\frac{\partial \varrho}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \varrho \right) \mathrm{d}\underset{\sim}{x} = 0. \tag{3.9}$$

Due to (3.7), the material volume Ω does not change with time and therefore applying the Reynolds transport theorem as in Lemma 1.3, we obtain,

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \varrho(\underset{\sim}{x}, t)\, \mathrm{d}\underset{\sim}{x} = 0, \tag{3.10}$$

or equivalently, $\int_\Omega \varrho(\underset{\sim}{x}, t)\, \mathrm{d}\underset{\sim}{x} = \int_\Omega \varrho_0(\underset{\sim}{x})\, \mathrm{d}\underset{\sim}{x}$ for $t \in (0,T]$.

Remark 3.1 *By taking test functions of the form $\zeta = \chi_S \sqrt{M}$, where χ_S is a mollified characteristic function for $S \subset \Omega$, one could extend the above result to arbitrary subsets of Ω and therefore recover Lemma 1.3 in its full generality for the weak solution.* $\diamond$

We now introduce the spatial discretisation of (3.5), (3.6). Let V_h be a N_Ω-dimensional $\mathrm{H}^1(\Omega)$-conforming finite element space corresponding to a triangulation $\mathcal{T}_h$ of Ω. Also, as in Chapter 2, let $\mathcal{P}_N(D) \subset \mathrm{H}^1_0(D) \subset \mathrm{H}^1_0(D; M)$ be an N_D-dimensional

space spanned by a set of spectral basis functions on D (such as $\mathcal{A}, \mathcal{B}$ or $\mathcal{C}$ from Section 2.6). Noting that $V_h \otimes \mathcal{P}_N(D) \subset \mathcal{X}$, we obtain a spatially discrete formulation of the full Fokker–Planck equation as follows:

Let $\hat{\psi}_{h,N}(\cdot,\cdot,0) \in V_h \otimes \mathcal{P}_N(D)$ be the $\mathrm{L}^2(\Omega \times D)$ projection of $\hat{\psi}_0$ onto $V_h \otimes \mathcal{P}_N(D)$. Find $\hat{\psi}_{h,N}(\cdot,\cdot,t) \in V_h \otimes \mathcal{P}_N(D)$, $t \in (0,T]$ satisfying (3.5) for all $\zeta \in V_h \otimes \mathcal{P}_N(D)$ in the sense of distributions on $(0,T)$.

It would be possible to finite difference in time the spatially discrete formulation defined above in order to obtain a fully-discrete numerical method. However, this would be impractical in the present context because the discrete problem at each time-level would be posed on the domain $\Omega \times D$. As we have indicated, a more reasonable alternative is to use an alternating-direction method to split each $2d$-dimensional solve into a sequence of d-dimensional solves. This idea is considered in detail in the next section.

3.3 The alternating-direction numerical method

We begin this section by presenting a brief general overview of alternating-direction methods and we will then consider how to derive an alternating-direction method for (3.5), (3.6).

We concentrate on schemes that use a Galerkin spatial discretisation since this will allow us to use arguments analogous to those in Sections 2.2 and 2.3 in order to establish stability and convergence properties. The seminal work on alternating-direction methods of this type is by Douglas & Dupont [28]. In the example below, we illustrate the approach of Douglas & Dupont by considering a Galerkin-based alternating-direction method for the constant-coefficient heat equation in two spatial dimensions.

Example 3.2 *Suppose* $(x,y,t) \in (a_1,a_2) \times (b_1,b_2) \times (0,T) \mapsto u(x,y,t) \in \mathbb{R}$, *with* $u(\cdot,\cdot,0) = u_0(\cdot,\cdot)$ *and*

$$\frac{\partial u}{\partial t} - \Delta u = 0, \qquad \text{on } (x,y,t) \in (a_1,a_2) \times (b_1,b_2) \times (0,T),$$

with homogeneous Dirichlet boundary conditions in space. The corresponding weak formulation of this problem is:
Find $u \in \mathrm{L}^\infty(0,T;\mathrm{L}^2((a_1,b_1) \times (a_2,b_2))) \cap \mathrm{L}^2(0,T;\mathrm{H}_0^1((a_1,b_1) \times (a_2,b_2)))$ *satisfying*

$$\int_\Omega \frac{\partial u}{\partial t} v \,\mathrm{d}x\,\mathrm{d}y + \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x\,\mathrm{d}y = 0 \qquad \forall v \in \mathrm{H}_0^1((a_1,b_1) \times (a_2,b_2)), \tag{3.11}$$

$$u(x,y,0) = u_0(x,y), \qquad (x,y) \in (a_1,a_2) \times (b_1,b_2), \tag{3.12}$$

in the sense of distributions on $(0,T)$.

Suppose that X_h *and* Y_h *are* $\mathrm{H}_0^1(a_1,b_1)$- *and* $\mathrm{H}_0^1(a_2,b_2)$*-conforming finite element spaces, respectively, with bases* $\{v_i \in X_h : 1 \leq i \leq N\}$ *and* $\{w_i \in Y_h : 1 \leq i \leq N\}$ *such that* $X_h = span(\{v_i\}_{1\leq i\leq N})$ *and* $Y_h = span(\{w_i\}_{1\leq i\leq N})$. *Let* $X_h \otimes Y_h$ *denote the following tensor product space:*

$$X_h \otimes Y_h := \left\{ z \ : \ z = \sum_{i,j=1}^{N} \alpha_{ij} v_i \, w_j, \ \alpha_{ij} \in \mathbb{R} \text{ for each } 1 \leq i,j \leq N \right\}.$$

It follows that $X_h \otimes Y_h \subset \mathrm{H}_0^1(a_1,b_1;\mathrm{H}_0^1(a_2,b_2)) \subset \mathrm{H}_0^1((a_1,b_1)\times(a_2,b_2))$. *Using this tensor product finite element space we define a finite element scheme for this problem by replacing* $\mathrm{H}_0^1((a_1,b_1)\times(a_2,b_2))$ *with* $X_h \otimes Y_h$ *in the weak formulation above. Also, supposing we employ Crank–Nicolson finite differencing to discretise* (3.11) *in time, then we obtain the following fully discrete problem (written in matrix form) at each time-step: Given* $u_h^n \in X_h \otimes Y_h$, *find* $u_h^{n+1} \in X_h \otimes Y_h$ *satisfying*

$$\begin{aligned} &\left(M_x \otimes M_y + \frac{\Delta t}{2} \left(S_x \otimes M_y + M_x \otimes S_y \right) \right) u_h^{n+1} \\ &\qquad = \left(M_x \otimes M_y - \frac{\Delta t}{2} \left(S_x \otimes M_y + M_x \otimes S_y \right) \right) u_h^{n}, \end{aligned} \tag{3.13}$$

where M_x *and* S_x *(resp.* M_y *and* S_y*) are the* X_h *(resp.* Y_h*) mass and stiffness matrices, and the matrix tensor product*[1] *is defined as follows for matrices* $A \in \mathbb{R}^{m\times n}$ *and* $B \in \mathbb{R}^{p\times q}$:

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mp\times nq}.$$

Since the matrices in (3.13) *are tensor products of the* x- *and* y*-direction discretisation matrices, we can approximate* (3.13) *using the following two stage method:*

$$\left(M_x + \frac{\Delta t}{2} S_x \right) \otimes I \, u_h^{n*} = \left(M_x - \frac{\Delta t}{2} S_x \right) \otimes I \, u_h^{n} \tag{3.14}$$

$$I \otimes \left(M_y + \frac{\Delta t}{2} S_y \right) u_h^{n+1} = I \otimes \left(M_y - \frac{\Delta t}{2} S_y \right) u_h^{n*}. \tag{3.15}$$

These equations define the fully discrete Galerkin alternating-direction method for this problem. We refer to (3.14) *as the* $\underset{\sim}{x}$*-direction stage and to* (3.15) *as the* y*-direction stage.*

[1] Also referred to as the Kronecker product.

By multiplying (3.14) *by* $I \otimes (M_y - \Delta t/2 S_y)$ *and* (3.15) *by* $(M_x + \Delta t/2 S_x) \otimes I$*, we see that the Galerkin alternating-direction method is equivalent to the following:*

$$\left(M_x \otimes M_y + \frac{\Delta t}{2} \left(S_x \otimes M_y + M_x \otimes S_y \right) + \frac{(\Delta t)^2}{4} S_x \otimes S_y \right) u_h^{n+1} = \left(M_x \otimes M_y - \frac{\Delta t}{2} \left(S_x \otimes M_y + M_x \otimes S_y \right) + \frac{(\Delta t)^2}{4} S_x \otimes S_y \right) u_h^n. \quad (3.16)$$

This is referred to as the equivalent one-step method for (3.14), (3.15)*. We can see that the one-step method is identical to the Crank-Nicolson scheme,* (3.13)*, except for the presence of the* $\mathcal{O}((\Delta t)^2)$ *perturbation terms in* (3.16)*.*

Using the approach of Douglas & Dupont, the next step is to rewrite (3.16) *in inner product form as follows: Given* $u_h^n \in X_h \otimes Y_h$*, find* $u_h^{n+1} \in X_h \otimes Y_h$ *satisfying*

$$\int_\Omega \frac{u_h^{n+1} - u_h^n}{\Delta t} v_h \, \mathrm{d}x \, \mathrm{d}y + \frac{1}{2} \int_\Omega \left\{ \nabla u_h^{n+1} \cdot \nabla v_h + \frac{\Delta t}{2} \left(\frac{\partial u_h^{n+1}}{\partial x} \frac{\partial v_h}{\partial y} + \frac{\partial u_h^{n+1}}{\partial y} \frac{\partial v_h}{\partial x} \right) \right\} \mathrm{d}x \, \mathrm{d}y = \frac{1}{2} \int_\Omega \left\{ -\nabla u_h^n \cdot \nabla v_h + \frac{\Delta t}{2} \left(\frac{\partial u_h^n}{\partial x} \frac{\partial v_h}{\partial y} + \frac{\partial u_h^n}{\partial y} \frac{\partial v_h}{\partial x} \right) \right\} \mathrm{d}x \, \mathrm{d}y \quad (3.17)$$

for all $v_h \in X_h \otimes Y_h$*. From here, one can use standard energy analysis to establish stablity and convergence properties of* (3.17)*, and therefore, equivalently, of* (3.14)*,* (3.15)*.*

We now apply the approach described in Example 3.2 to the weak formulation, (3.5). First of all, define the bases

$$\{Y_k \in \mathcal{P}_N(D) : 1 \le k \le N_D\} \quad \text{and} \quad \{X_i \in V_h : 1 \le i \le N_\Omega\}, \quad (3.18)$$

such that $\mathrm{span}(\{Y_k\}_{1 \le k \le N_D}) = \mathcal{P}_N(D)$ and $\mathrm{span}(\{X_i\}_{1 \le i \le N_\Omega}) = V_h$. Recalling (2.54), we define $M_q, S_q \in \mathbb{R}^{N_D \times N_D}$ as

$$(M_q)_{lk} := \int_D Y_k(\underset{\sim}{q}) Y_l(\underset{\sim}{q}) \, \mathrm{d}\underset{\sim}{q}, \quad (3.19)$$

$$(S_q)_{lk} := \int_D \underset{\sim}{\nabla}_M Y_k(\underset{\sim}{q}) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q}) \, \mathrm{d}\underset{\sim}{q}. \quad (3.20)$$

Similarly, $M_x, T_x \in \mathbb{R}^{N_\Omega \times N_\Omega}$ are defined as follows:

$$(M_x)_{ij} := \int_\Omega X_i(\underset{\sim}{x}) X_j(\underset{\sim}{x}) \, \mathrm{d}\underset{\sim}{x}, \quad (3.21)$$

$$(T_x)_{ij} := \int_\Omega (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x X_j(\underset{\sim}{x})) \, X_i(\underset{\sim}{x}) \, \mathrm{d}\underset{\sim}{x}. \quad (3.22)$$

A fully discrete form of (3.5) using a backward-Euler time discretisation can be written as follows: Given $\hat{\psi}_N^n = \sum_{jl} \gamma_{jl}^n X_j Y_l \in V_h \otimes \mathcal{P}_N(D)$, find the vector $\underset{\sim}{\gamma}^{n+1} \in \mathbb{R}^{N_D N_\Omega}$, defining a function $\hat{\psi}_N^{n+1} = \sum_{jl} \gamma_{jl}^{n+1} X_j Y_l \in V_h \otimes \mathcal{P}_N(D)$, such that

$$M_x \otimes M_q \left(\frac{\underset{\sim}{\gamma}^{n+1} - \underset{\sim}{\gamma}^n}{\Delta t} \right) + T_x \otimes M_q \underset{\sim}{\gamma}^{n+1} + \frac{1}{2\mathrm{Wi}} M_x \otimes S_q \underset{\sim}{\gamma}^{n+1} - C(\underset{\approx}{\kappa}^{n+1}; \hat{\psi}_N^{n+1}, \zeta_{ik}) = 0, \tag{3.23}$$

where $\zeta_{ik} = X_i \times Y_k \in V_h \otimes \mathcal{P}_N(D)$. It is also possible to obtain a tensor product form discretisation matrix of $C(\underset{\approx}{\kappa}; \cdot, \cdot)$, *i.e.* consider $C(\underset{\approx}{\kappa}; \zeta_{jl}, \zeta_{ik})$ as follows:

$$\begin{aligned} C(\underset{\approx}{\kappa}; \zeta_{jl}, \zeta_{ik}) &= \int_{\Omega \times D} \left(\underset{\approx}{\kappa}^{n+1}(x) \underset{\sim}{q} X_j(\underset{\sim}{x}) Y_l(\underset{\sim}{q}) \right) \cdot \underset{\sim}{\nabla}_M (X_i(\underset{\sim}{x}) Y_k(\underset{\sim}{q})) \, \mathrm{d}\underset{\sim}{x} \, \mathrm{d}\underset{\sim}{q} \\ &= \sum_{s,t=1}^{d} \left(\int_\Omega \kappa_{st}^{n+1}(\underset{\sim}{x}) X_i(\underset{\sim}{x}) X_j(x) \, \mathrm{d}\underset{\sim}{x} \right) \left(\int_D q_t \, Y_l(\underset{\sim}{q}) \sqrt{M} \frac{\partial}{\partial q_s} \left(\frac{Y_k(\underset{\sim}{q})}{\sqrt{M}} \right) \mathrm{d}\underset{\sim}{q} \right). \end{aligned}$$

Therefore, we define the matrices $C_x^{st} \in \mathbb{R}^{N_\Omega \times N_\Omega}$ and $C_q^{st} \in \mathbb{R}^{N_D \times N_D}$ for $1 \leq s, t \leq d$ such that

$$\left(C_x^{st} \right)_{ij} := \int_\Omega \kappa_{st}^{n+1}(\underset{\sim}{x}) X_i(\underset{\sim}{x}) X_j(x) \, \mathrm{d}\underset{\sim}{x}, \tag{3.24}$$

$$\left(C_q^{st} \right)_{kl} := \int_D q_t \, Y_l(\underset{\sim}{q}) \sqrt{M} \frac{\partial}{\partial q_s} \left(\frac{Y_k(\underset{\sim}{q})}{\sqrt{M}} \right) \mathrm{d}\underset{\sim}{q}. \tag{3.25}$$

Hence, we can rewrite the term on the final line of (3.23) as $\sum_{s,t=1}^{d} C_x^{st} \otimes C_q^{st} \underset{\sim}{\gamma}^{n+1}$.

However, since this matrix expression for $C(\underset{\approx}{\kappa}; \cdot, \cdot)$ contains neither M_x nor M_q, we can no longer factorise the resulting equation in the same way as in (3.14), (3.15). That is, the term $C(\underset{\approx}{\kappa}; \cdot, \cdot)$ causes difficulties because its 'coefficient', $\underset{\approx}{\kappa}(\underset{\sim}{x})\underset{\sim}{q}$, depends on both the $\underset{\sim}{x}$- and $\underset{\sim}{q}$-directions.

This issue has been considered a number of times in the literature. For example, in the context of collocation-based alternating-direction schemes Celia & Pinder [21, 22] and Bialecki & Fernandes [16] developed methods that could handle equations with general variable coefficients. However, as indicated earlier, our focus is on developing a Galerkin-based framework, and therefore, again, the work of Douglas & Dupont is the most relevant here. In [28], Douglas & Dupont developed a "Laplace modification" scheme for the heat equation with general coefficients which involved discretising the equation

$$\frac{\partial u}{\partial t} = \nabla \cdot (a(x, y, t, u) \nabla u) + f(x, y, t, u),$$

as follows,

$$\left(\frac{u^{n+1}-u^n}{\Delta t}, v\right) + (a^n(u^n)\nabla u^n, \nabla v) + \lambda\left(\nabla(u^{n+1}-u^n), \nabla v\right) = (f^n(u^n), v)\,,$$

where λ is a constant scalar, which must satisfy a lower bound condition related to the supremum of $|a|$ in order to ensure the stability of the numerical method. This discretisation then allows the use of a standard Galerkin alternating-direction method, as in Example 3.2, because the term containing a can be moved to the right-hand side and treated as a source term.

However, it is not obvious how to apply this kind of approach to (3.23), because our problematic term is a convection term rather than a diffusion term. The most natural idea in the spirit of Douglas & Dupont would be to move the $C(\underset{\approx}{\kappa};\cdot,\cdot)$ term to the right-hand side of (3.23) and treat it explicitly in time. This idea is feasible, but for the purposes of practical computations, we would like to have the option of using a fully-implicit temporal discretisation. Indeed, the numerical results in Section 2.6.2 demonstrated that the semi-implicit temporal discretisation of the Fokker–Planck equation in which the term $C(\underset{\approx}{\kappa};\cdot,\cdot)$ was treated explicitly in time was less stable than the backward Euler discretisation, especially for problems in which the product $\mathrm{Wi}\,\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}$ is significantly larger than 1.

In order to circumvent this limitation, we develop a Galerkin alternating-direction approach that is an almagamation of the Douglas & Dupont framework and a new quadrature-based method. Using this approach, we can define either a fully-implicit in time or a semi-implicit in time alternating-direction method for the Fokker–Planck equation. We shall consider both options in detail in this chapter.

3.3.1 The hybrid alternating-direction scheme

The first ingredient of this scheme is a quadrature rule on Ω.

Let $\{(\underset{\sim}{x}_m, w_m), w_m > 0,\ \underset{\sim}{x}_m \in \overline{\Omega}, m = 1, \ldots, Q_\Omega\}$ define an element-based quadrature rule on the triangulation $\mathcal{T}_h$, where the $\underset{\sim}{x}_m$ are the quadrature points and the w_m are the corresponding weights. Therefore, for functions $f, g \in \mathrm{C}^0(\overline{\Omega})$, the quadrature sum is evaluated element-wise as follows,

$$\sum_{m=1}^{Q_\Omega} w_m f(\underset{\sim}{x}_m) g(\underset{\sim}{x}_m) = \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{Q_K} w_l^K f(\underset{\sim}{x}_l^K) g(\underset{\sim}{x}_l^K), \tag{3.26}$$

where Q_K is the number of quadrature points in element K. From now on, we will use the left-hand side of (3.26) as a shorthand for the right-hand side.

We now introduce two alternative hypotheses on the accuracy of the quadrature rule, Quadrature Hypothesis 1 (QH1) and Quadrature Hypothesis 2 (QH2).

Quadrature Hypothesis 1 (QH1). The quadrature rule satisfies

$$\sum_{m=1}^{Q_\Omega} w_m \kappa_{ij}(\underset{\sim}{x}_m) f(\underset{\sim}{x}_m) g(\underset{\sim}{x}_m) = \int_\Omega \kappa_{ij}(\underset{\sim}{x}) f(\underset{\sim}{x}) g(\underset{\sim}{x}) \, \mathrm{d}\underset{\sim}{x}, \tag{3.27}$$

for all $f, g \in V_h$ and for each component κ_{ij} of $\underset{\approx}{\kappa}$. $\diamond$

As discussed in Chapter 4, in the context of the Navier–Stokes–Fokker–Planck system, we compute the macroscopic velocity field, $\underset{\sim}{u}$ by solving the Navier–Stokes equations using a finite element method on the triangulation $\mathcal{T}_h$, *i.e.* the same triangulation that is used for the alternating-direction method for the Fokker–Planck equation. As a result, it is reasonable to assume that the components of $\underset{\approx}{\kappa} = \nabla_x \underset{\sim}{u}$ are represented by piecewise polynomials on $\mathcal{T}_h$ and in this case it is certainly possible to satisfy QH1 by choosing an appropriate element-based quadrature rule.

Quadrature Hypothesis 2 (QH2). The quadrature rule satisfies

$$\sum_{m=1}^{Q_\Omega} w_m f(\underset{\sim}{x}_m) g(\underset{\sim}{x}_m) = \int_\Omega f(\underset{\sim}{x}) g(\underset{\sim}{x}) \, \mathrm{d}\underset{\sim}{x}, \tag{3.28}$$

for all $f, g \in V_h$. $\diamond$

QH1 is a stronger hypothesis than QH2, and therefore in general we will require a larger value of Q_Ω in order to satisfy QH1. Some results in the following analysis will require QH1, whereas for others, QH2 will suffice. Refer to Section 3.8 for a discussion of specific quadrature rules that we use to satisfy QH1 and QH2 in practice.

Next, let $\hat{\psi}_{h,N} \in V_h \otimes \mathcal{P}_N(D)$ denote the numerical solution of the full Fokker–Planck equation. Recalling the bases from (3.18), $\hat{\psi}_{h,N}$ can be written in terms of coefficients $\{\hat{\psi}_{ik}\}$ as follows:

$$\hat{\psi}_{h,N} := \sum_{i=1}^{N_\Omega} \sum_{k=1}^{N_D} \hat{\psi}_{ik} X_i Y_k \in V_h \otimes \mathcal{P}_N(D). \tag{3.29}$$

Define the *line functions*, $\hat{\psi}_k$, for $k = 1, \ldots, N_D$ as follows:

$$\hat{\psi}_k := \sum_{i=1}^{N_\Omega} \hat{\psi}_{ik} X_i \in V_h, \tag{3.30}$$

and note that (3.29) can be rewritten using (3.30) as follows:

$$\hat{\psi}_{h,N}(\underset{\sim}{x}, \underset{\sim}{q}) = \sum_{k=1}^{N_D} \hat{\psi}_k(\underset{\sim}{x}) Y_k(\underset{\sim}{q}). \tag{3.31}$$

The formula (3.31) shall be useful in the discussion of the alternating-direction methods below.

As discussed above, the term $C(\underset{\approx}{\kappa};\cdot,\cdot)$ is the most problematic in terms of applying an alternating-direction method to the Fokker–Planck equation. Therefore we begin by considering how to use a quadrature-based scheme to derive an alternating-direction type of formulation of this term.

Suppose that QH1 is satisfied and that we have the line function decomposition (3.31) for $\hat{\psi}_{h,N}$, in which $\hat{\psi}_k \in V_h$ for $k = 1 \ldots, N_D$. Also, let $\zeta = X \times Y \in V_h \otimes \mathcal{P}_N(D)$. Then,

$$\begin{aligned}
C(\underset{\approx}{\kappa};\hat{\psi}_{h,N},\zeta) &= \int_{\Omega\times D} (\underset{\approx}{\kappa}\,\underset{\sim}{q}\,\hat{\psi}_{h,N}(\underset{\sim}{x},\underset{\sim}{q}))\cdot\underset{\sim}{\nabla}_M\zeta(\underset{\sim}{x},\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\,\mathrm{d}\underset{\sim}{x} \\
&= \int_D \sum_{k=1}^{N_D}\int_\Omega \left[\underset{\approx}{\kappa}\,\underset{\sim}{q}\,\hat{\psi}_k(\underset{\sim}{x})Y_k(\underset{\sim}{q})\right]\cdot\underset{\sim}{\nabla}_M\left(X(\underset{\sim}{x})Y(\underset{\sim}{q})\right)\;\mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q} \\
&= \int_D \sum_{k=1}^{N_D}\sum_{m=1}^{Q_\Omega} w_m\left[\underset{\approx}{\kappa}(\underset{\sim}{x}_m)\,\underset{\sim}{q}\,\hat{\psi}_k(\underset{\sim}{x}_m)Y_k(\underset{\sim}{q})\right]\cdot\underset{\sim}{\nabla}_M\left(X(\underset{\sim}{x}_m)Y(\underset{\sim}{q})\right)\;\mathrm{d}\underset{\sim}{q} \\
&= \sum_{m=1}^{Q_\Omega} w_m\,X(\underset{\sim}{x}_m)\left\{\sum_{k=1}^{N_D}\hat{\psi}_k(\underset{\sim}{x}_m)\left(\int_D(\underset{\approx}{\kappa}(\underset{\sim}{x}_m)\,\underset{\sim}{q}\,Y_k(\underset{\sim}{q}))\cdot\underset{\sim}{\nabla}_M Y(\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\right)\right\}. \qquad (3.32)
\end{aligned}$$

This shows the equivalence between the Galerkin formulation of $C(\underset{\approx}{\kappa};\cdot,\cdot)$ on $\Omega\times D$ and the quadrature sum over $m = 1,\ldots,Q_\Omega$ of the term

$$\sum_{k=1}^{N_D}\hat{\psi}_k(\underset{\sim}{x}_m)\left(\int_D(\underset{\approx}{\kappa}(\underset{\sim}{x}_m)\,\underset{\sim}{q}\,Y_k(\underset{\sim}{q}))\cdot\underset{\sim}{\nabla}_M Y(\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\right), \qquad (3.33)$$

which is the $\underset{\sim}{q}$-direction discretisation of $C(\underset{\approx}{\kappa};\cdot,\cdot)$.

Note that (3.33) is exactly the discretisation of the $\underset{\sim}{q}$-convection term that was used in the spectral method in Chapter 2, except that now $\underset{\approx}{\kappa}$ depends on $\underset{\sim}{x}\in\Omega$, and we sample $\underset{\approx}{\kappa}$ at the quadrature points $\underset{\sim}{x}_m$. Also, the coefficient vector in (3.33) corresponding to the quadrature point $\underset{\sim}{x}_m$ is the set of sampled line functions, $\hat{\psi}_k(\underset{\sim}{x}_m)$, $k = 1,\ldots,N_D$.

The preceding discussion relied on QH1, however we can use an analogous argument when only QH2 is assumed, in which case the quadrature rule is no longer exact for the $\underset{\approx}{\kappa}$-weighted integral in (3.27) and therefore we do not have equality between the second and third lines of (3.32). Instead, a quadrature error, E, is introduced as follows:

$$\sum_{m=1}^{Q_\Omega} w_m\kappa_{ij}(\underset{\sim}{x}_m)\,\hat{\psi}_k(\underset{\sim}{x}_m)X(\underset{\sim}{x}_m) = \int_\Omega \kappa_{ij}(\underset{\sim}{x})\,\hat{\psi}_k(\underset{\sim}{x})X(\underset{\sim}{x})\,\mathrm{d}\underset{\sim}{x} + E(\kappa_{ij},\hat{\psi}_k,X). \qquad (3.34)$$

Modifying (3.32) to include this error term, we obtain:

$$\sum_{m=1}^{Q_\Omega} w_m \, X(\underset{\sim}{x}_m) \left\{ \sum_{k=1}^{N_D} \hat{\psi}_k(\underset{\sim}{x}_m) \left(\int_D (\underset{\approx}{\kappa}(\underset{\sim}{x}_m) \, \underset{\sim}{q} \, Y_k(\underset{\sim}{q})) \cdot \underset{\sim}{\nabla}_M X(\underset{\sim}{q}) \, \mathrm{d}\underset{\sim}{q} \right) \right\}$$
$$= C(\underset{\approx}{\kappa}; \hat{\psi}_{h,N}, \zeta) + \sum_{k=1}^{N_D} \int_D \underset{\approx}{E}(\underset{\approx}{\kappa}, \hat{\psi}_k, X) \, \underset{\sim}{q} \, Y_k(\underset{\sim}{q}) \cdot \underset{\sim}{\nabla}_M Y(\underset{\sim}{q}) \, \mathrm{d}\underset{\sim}{q}, \quad (3.35)$$

where $\left(\underset{\approx}{E}(\underset{\approx}{\kappa}, \hat{\psi}_k, X)\right)_{ij} := E(\kappa_{ij}, \hat{\psi}_k, X)$. Of course, the precise nature of $\underset{\approx}{E}$ will depend on the choice of quadrature rule and the problem at hand. Nevertheless, if appropriate hypotheses on the rate of decay of $\underset{\approx}{E}$ are specified, it would be possible to consider the stability and convergence properties of an alternating-direction method that includes a quadrature error term of this form. However, for simplicity and brevity, we do not consider such quadrature error terms in the numerical analysis in this chapter. It is worth noting though that we develop a stability argument in Section 3.4 that only relies on QH2, and in which we do not need to consider quadrature error terms such as in (3.34).

It is clear from (3.32) that sampling functions at the quadrature points $\{\underset{\sim}{x}_m \in \overline{\Omega}, \; m = 1, \ldots, Q_\Omega\}$ will play an important role in the alternating-direction methods we define below. We will also require a reconstruction operator, which maps from a set of values at the quadrature points to a function in V_h. We now introduce this operator. To simplify notation, we first define the following discrete inner product and norm over Ω for $\{f_m\}, \{g_m\} \in \mathbb{R}^{Q_\Omega}$:

$$(\{f_m\}, \{g_m\})_{\ell^2(\Omega)} := \sum_{m=1}^{Q_\Omega} w_m f_m g_m, \quad \text{and} \quad \|\{f_m\}\|_{\ell^2(\Omega)} := (\{f_m\}, \{f_m\})_{\ell^2(\Omega)}^{\frac{1}{2}}. \quad (3.36)$$

Note that, by (3.27) or (3.28), for $f, g \in V_h$, $(\{f(\underset{\sim}{x}_m)\}, \{g(\underset{\sim}{x}_m)\})_{\ell^2(\Omega)} = (f, g)_{\mathrm{L}^2(\Omega)}$, where $(\cdot, \cdot)_{\mathrm{L}^2(\Omega)}$ is the standard L^2 inner product on Ω. Next we define the reconstruction operator $\mathcal{R} : \{f_m\} \in \mathbb{R}^{Q_\Omega} \mapsto \mathcal{R}\{f_m\} \in V_h$ such that

$$(\mathcal{R}\{f_m\}, X)_{\mathrm{L}^2(\Omega)} = (\{f_m\}, \{X(\underset{\sim}{x}_m)\})_{\ell^2(\Omega)} \qquad \forall \, X \in V_h. \quad (3.37)$$

Remark 3.3 *For any $\mathcal{R}\{f_m\} \in V_h$, there exist real numbers $\gamma_1, \ldots, \gamma_{N_\Omega}$ such that $\mathcal{R}\{f_m\} = \sum_{j=1}^{N_\Omega} \gamma_j X_j$. Letting $X = X_i$, $i = 1, \ldots, N_\Omega$ above it is clear that* (3.37) *is equivalent to the linear system $M_x \underset{\sim}{\gamma} = \underset{\sim}{F}$ where $M_x \in \mathbb{R}^{N_\Omega \times N_\Omega}$ is the V_h mass matrix, $\underset{\sim}{\gamma} = (\gamma_1, \ldots, \gamma_{N_\Omega})^T$, and $\underset{\sim}{F} \in \mathbb{R}^{N_\Omega}$ is such that $F_i = (\{f_m\}, \{X_i(\underset{\sim}{x}_m)\})_{\ell^2(\Omega)}$. The matrix M_x is non-singular, and therefore the reconstruction operator defined in* (3.37) *is well-defined.* $\diamond$

We are now in a position to discuss the alternating-direction Galerkin methods that are the focus of this chapter. We introduce two algorithms below, denoted method I and method II. Each method utilises a hybrid alternating-direction method, which combines the quadrature approach illustrated in (3.32) with a standard Douglas-Dupont type Galerkin alternating-direction method.

The distinction between method I and method II is that method I uses a semi-implicit spectral method in the $\underset{\sim}{q}$-direction (*i.e.* the term $C(\underset{\approx}{\kappa};\cdot,\cdot)$ is treated explicitly in time) whereas method II uses a fully-implicit temporal discretisation.

3.3.2 Method I: Semi-implicit scheme

Method I is initialised by computing the $\mathrm{L}^2(\Omega \times D)$ projection, $\hat{\psi}^0_{h,N}$, of the initial datum $\hat{\psi}_0 \in \mathrm{L}^2(\Omega \times D)$ onto $V_h \otimes \mathcal{P}_N(D)$, so that $\hat{\psi}^0_{h,N} \in V_h \otimes \mathcal{P}_N(D)$, satisfies

$$\left(\hat{\psi}_0,\, \zeta\right) = \left(\hat{\psi}^0_{h,N}, \zeta\right) \qquad \text{for all } \zeta \in V_h \otimes \mathcal{P}_N(D). \tag{3.38}$$

Then, as in (1.50), (1.51), this alternating-direction method consists of two stages at each time-step: the $\underset{\sim}{q}$-direction stage and the $\underset{\sim}{x}$-direction stage. We begin with the $\underset{\sim}{q}$-direction stage, in which we essentially use the Galerkin spectral method in D from Chapter 2.

Suppose $\hat{\psi}^n_{h,N} \in V_h \otimes \mathcal{P}_N(D)$. Then in the $\underset{\sim}{q}$-direction stage we compute $\hat{\psi}^{n*}_{h,N}(\underset{\sim}{x}_m, \cdot) \in \mathcal{P}_N(D)$ for each $m = 1, \ldots, Q_\Omega$ satisfying

$$\begin{aligned}\int_D \frac{\hat{\psi}^{n*}_{h,N}(\underset{\sim}{x}_m, \underset{\sim}{q}) - \hat{\psi}^n_{h,N}(\underset{\sim}{x}_m, \underset{\sim}{q})}{\Delta t}\, Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} + \frac{1}{2\mathrm{Wi}} \int_D \underset{\sim}{\nabla}_M \hat{\psi}^{n*}_{h,N}(\underset{\sim}{x}_m, \underset{\sim}{q}) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} \\ = \int_D (\underset{\approx}{\kappa}^n(\underset{\sim}{x}_m)\, \underset{\sim}{q}\, \hat{\psi}^n_{h,N}(\underset{\sim}{x}_m, \underset{\sim}{q})) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q},\end{aligned} \tag{3.39}$$

for $l = 1, \ldots, N_D$. (3.39) defines an $N_D \times N_D$ linear system at each quadrature point. In order to separate out the $\underset{\sim}{x}$- and $\underset{\sim}{q}$-direction dependencies more clearly, we rewrite this equation in terms of line functions using (3.30), *i.e.*:

$$\begin{aligned}\sum_{k=1}^{N_D} \hat{\psi}^{n*}_k(\underset{\sim}{x}_m) \left(\int_D Y_k(\underset{\sim}{q})\, Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} + \frac{\Delta t}{2\mathrm{Wi}} \int_D \underset{\sim}{\nabla}_M Y_k(\underset{\sim}{q}) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q}\right) \\ = \sum_{k=1}^{N_D} \hat{\psi}^n_k(\underset{\sim}{x}_m) \left(\int_D Y_k(\underset{\sim}{q})\, Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} + \Delta t \int_D (\underset{\approx}{\kappa}^n(\underset{\sim}{x}_m)\, \underset{\sim}{q}\, Y_k(\underset{\sim}{q})) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q}\right),\end{aligned} \tag{3.40}$$

for $l = 1, \ldots, N_D$. This system is solved at each quadrature point $\underset{\sim}{x}_m$, $m = 1, \ldots, Q_\Omega$.

Equation (3.40) shows that in the $\underset{\sim}{q}$-direction stage, the sampled values of the line functions, *i.e.* $\psi^{n*}_k(\underset{\sim}{x}_m)$, $k = 1, \ldots, N_D$, $m = 1, \ldots, Q_\Omega$, are the coefficients to be

computed. We determine these values by solving a different linear system at each quadrature point. Note that these linear systems are completely independent from one another. This independence enables parallel computation to be used very effectively in this context; this will be discussed in more detail later.

The $\underset{\sim}{q}$-direction stage is complete once the values $\psi_k^{n*}(\underset{\sim}{x}_m)$, $k = 1, \ldots, N_D$, $m = 1, \ldots, Q_\Omega$ have been computed, and then we can begin solving in the $\underset{\sim}{x}$-direction. In the $\underset{\sim}{x}$-direction stage, we use a finite element discretisation of the transport equation (1.51) to update the output data from the $\underset{\sim}{q}$-direction stage. That is, for a given k, we find $\hat{\psi}_k^{n+1} \in V_h$, satisfying:

$$\int_\Omega \hat{\psi}_k^{n+1} X_i \, \mathrm{d}\underset{\sim}{x} + \Delta t \int_\Omega \left(\underset{\sim}{u}^{n+1} \cdot \underset{\sim}{\nabla}_x \hat{\psi}_k^{n+1} \right) X_i \, \mathrm{d}\underset{\sim}{x} = \int_\Omega \mathcal{R}\{\hat{\psi}_k^{n*}(\underset{\sim}{x}_m)\} X_i \, \mathrm{d}\underset{\sim}{x}, \tag{3.41}$$

for $i = 1, \ldots, N_\Omega$.

Note, however, that based on (3.37), for the right-hand side in (3.41) we have:

$$\int_\Omega \mathcal{R}\{\hat{\psi}_k^{n*}(\underset{\sim}{x}_m)\} X_i \, \mathrm{d}\underset{\sim}{x} = \sum_{m=1}^{Q_\Omega} w_m \, \hat{\psi}_k^{n*}(\underset{\sim}{x}_m) \, X_i(\underset{\sim}{x}_m) =: F_i. \tag{3.42}$$

Hence we do not actually have to explicitly compute $\mathcal{R}\{\hat{\psi}_k^{n*}(\underset{\sim}{x}_m)\} \in V_h$ in order to solve (3.41), since it is equivalent to solve the following system:

$$\int_\Omega \hat{\psi}_k^{n+1} X_i \, \mathrm{d}\underset{\sim}{x} + \Delta t \int_\Omega \left(\underset{\sim}{u}^{n+1} \cdot \underset{\sim}{\nabla}_x \hat{\psi}_k^{n+1} \right) X_i \, \mathrm{d}\underset{\sim}{x} = F_i, \tag{3.43}$$

for $i = 1, \ldots, N_\Omega$. We solve (3.43) for each $k = 1, \ldots, N_D$, and, just as in the $\underset{\sim}{q}$-direction, these computations are decoupled from one another.

Once the $\underset{\sim}{x}$-direction computations are complete, we have the numerical solution at time level $n + 1$:

$$\hat{\psi}_{h,N}^{n+1} = \sum_{k=1}^{N_D} \hat{\psi}_k^{n+1} \, Y_k \in V_h \otimes \mathcal{P}_N(D).$$

Hence method I is defined by the initialisation (3.38), the $\underset{\sim}{q}$-direction spectral method (3.40) and the $\underset{\sim}{x}$-direction finite element method (3.43).

Before continuing further, we first verify that the $\underset{\sim}{q}$- and $\underset{\sim}{x}$-direction numerical methods are well-defined.

Lemma 3.4 *Let $A_q \in \mathbb{R}^{N_D \times N_D}$ denote the matrix appearing on the left-hand side of* (3.40)*,* i.e.

$$A_q := M_q + \frac{\Delta t}{2\mathrm{Wi}} S_q, \tag{3.44}$$

and let $A_x \in \mathbb{R}^{N_\Omega \times N_\Omega}$ *be the matrix from the left-hand side of* (3.41),

$$A_x := M_x + \Delta t T_x. \tag{3.45}$$

The matrices A_q *and* A_x *are non-singular.*

Proof. The result follows straightforwardly from the positive-definiteness of the bilinear forms, $\mathfrak{B}_q(\cdot,\cdot) : \mathcal{P}_N(D) \times \mathcal{P}_N(D) \mapsto \mathbb{R}$, and $\mathfrak{B}_x(\cdot,\cdot) : V_h \times V_h \mapsto \mathbb{R}$, defining A_q and A_x respectively.

Consider $\mathfrak{B}_q(X,X)$ for any $X \in \mathcal{P}_N(D) \setminus \{0\}$:

$$\mathfrak{B}_q(X,X) = \|X\|^2_{\mathrm{L}^2(D)} + \frac{\Delta t}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M X\|^2_{\mathrm{L}^2(D)} \geq \|X\|^2_{\mathrm{L}^2(D)} > 0. \tag{3.46}$$

Similarly, integrating by parts and utilising the enclosed flow and divergence free assumptions for $\mathfrak{B}_x(Y,Y)$ with $Y \in V_h \backslash \{0\}$, we have,

$$\mathfrak{B}_x(Y,Y) = \|Y\|^2_{\mathrm{L}^2(\Omega)} - \frac{\Delta t}{2} \int_\Omega (\underset{\sim}{\nabla}_x \cdot \underset{\sim}{u}^{n+1})\, Y^2 \, \mathrm{d}\underset{\sim}{x} = \|Y\|^2_{\mathrm{L}^2(\Omega)} > 0. \tag{3.47}$$

This completes the proof. □

In the next lemma we derive a Galerkin formulation posed on $\Omega \times D$ for method I. This will allow us to apply arguments analogous to those in Chapter 2 to the numerical analysis of method I.

Lemma 3.5 *Suppose the* $\underset{\sim}{x}$*-direction quadrature rule satisfies QH1. Method I is equivalent to the following fully-discrete formulation:*

Given $\hat{\psi}^0_{h,N} \in V_h \otimes \mathcal{P}_N(D)$ *defined as in* (3.38)*, for each* $n = 0, \ldots, N_T - 1$, $\hat{\psi}^{n+1}_{h,N} \in V_h \otimes \mathcal{P}_N(D)$ *satisfies*

$$\begin{aligned}\left(\frac{\hat{\psi}^{n+1}_{h,N} - \hat{\psi}^n_{h,N}}{\Delta t}, \zeta\right) + \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}^{n+1}_{h,N}\,,\, \zeta\right) + \frac{1}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M \hat{\psi}^{n+1}_{h,N}\,,\, \underset{\sim}{\nabla}_M \zeta\right) \\ + \frac{\Delta t}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}^{n+1}_{h,N}\right), \underset{\sim}{\nabla}_M \zeta\right) - \left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\, \hat{\psi}^n_{h,N}\,,\, \underset{\sim}{\nabla}_M \zeta\right) = 0,\end{aligned} \tag{3.48}$$

for all $\zeta \in V_h \otimes \mathcal{P}_N(D)$.

Proof. Multiplying (3.40) through by $X_i(\underset{\sim}{x}_m)$, where $X_i \in V_h$, and performing

the weighted sum according to (3.26) gives,

$$\sum_{k=1}^{N_D}(\{\hat{\psi}_k^{n*}(\underset{\sim}{x}_m)\},\{X_i(\underset{\sim}{x}_m)\})_{\ell^2(\Omega)}\left(\int_D Y_k(\underset{\sim}{q})\,Y_l(\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}+\frac{\Delta t}{2\mathrm{Wi}}\int_D \underset{\sim}{\nabla}_M Y_k(\underset{\sim}{q})\cdot\underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\right)$$
$$=\sum_{k=1}^{N_D}(\{\hat{\psi}_k^{n}(\underset{\sim}{x}_m)\},\{X_i(\underset{\sim}{x}_m)\})_{\ell^2(\Omega)}\left(\int_D Y_k(\underset{\sim}{q})\,Y_l(\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\right)$$
$$+\Delta t\sum_{m=1}^{Q_\Omega} w_m X_i(\underset{\sim}{x}_m)\left\{\sum_{k=1}^{N_D}\hat{\psi}_k^{n}(\underset{\sim}{x}_m)\left(\int_D(\underset{\approx}{\kappa}^n(\underset{\sim}{x}_m)\,\underset{\sim}{q}\,Y_k(\underset{\sim}{q}))\cdot\underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\right)\right\}. \quad (3.49)$$

Using the reconstruction operator, (3.37), with the ℓ^2 inner products and the argument of (3.32) on the term on the third line[2], we obtain the following formulation for $\mathcal{R}\hat{\psi}_{h,N}^{n*}\in V_h\otimes\mathcal{P}_N(D)$,

$$\int_{\Omega\times D}\frac{\mathcal{R}\hat{\psi}_{h,N}^{n*}(\underset{\sim}{x},\underset{\sim}{q})-\hat{\psi}_{h,N}^{n}(\underset{\sim}{x},\underset{\sim}{q})}{\Delta t}\zeta(\underset{\sim}{x},\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\,\mathrm{d}\underset{\sim}{x}+\frac{1}{2\mathrm{Wi}}\int_{\Omega\times D}\underset{\sim}{\nabla}_M\mathcal{R}\hat{\psi}_{h,N}^{n*}(\underset{\sim}{x},\underset{\sim}{q})\cdot\underset{\sim}{\nabla}_M\zeta(\underset{\sim}{x},\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\,\mathrm{d}\underset{\sim}{x}$$
$$=\int_{\Omega\times D}(\underset{\approx}{\kappa}^n(\underset{\sim}{x})\,\underset{\sim}{q}\,\hat{\psi}_{h,N}^{n}(\underset{\sim}{x},\underset{\sim}{q}))\cdot\underset{\sim}{\nabla}_M\zeta(\underset{\sim}{x},\underset{\sim}{q})\;\mathrm{d}\underset{\sim}{q}\,\mathrm{d}\underset{\sim}{x}, \quad (3.50)$$

where $\zeta=X_i\times Y_l$ is an element of $V_h\otimes\mathcal{P}_N(D)$ and the numerical solution at the intermediate "time level" $n*$ is defined as:

$$\mathcal{R}\hat{\psi}_{h,N}^{n*}:=\sum_{k=1}^{N_D}\mathcal{R}\{\hat{\psi}_k^{n*}(\underset{\sim}{x}_m)\}Y_k\in V_h\otimes\mathcal{P}_N(D). \quad (3.51)$$

Equation (3.50) is the Galerkin formulation of (3.39) on $\Omega\times D$ that is obtained by performing a quadrature sum over all Q_Ω quadrature points in Ω.

The $\underset{\sim}{x}$-direction stage is more straightforward to deal with; we use the classical Douglas-Dupont Galerkin alternating-direction approach for (3.41), since it does not contain any $\underset{\sim}{q}$-dependent coefficients.

Let $\mathcal{R}\{\hat{\psi}_k^{n*}(\underset{\sim}{x}_m)\}=\sum_{i=1}^{N_\Omega}\gamma_{ik}^{n*}X_i$ so that according to (3.51), the vector $\underset{\sim}{\gamma}^{n*}=(\gamma_{11}^{n*},\ldots,\gamma_{N_\Omega 1}^{n*},\gamma_{12}^{n*},\ldots,\gamma_{N_\Omega N_D}^{n*})\in\mathbb{R}^{N_D\,N_\Omega}$ defines $\mathcal{R}\hat{\psi}_{h,N}^{n*}$. Similarly, denote the coefficient vector for $\hat{\psi}_{h,N}^{n+1}$ as $\underset{\sim}{\gamma}^{n+1}\in\mathbb{R}^{N_D\,N_\Omega}$, and since the vector entries are ordered in blocks according to the $\underset{\sim}{q}$-direction degrees-of-freedom, it follows that (3.41) can be written as a linear system where the matrices are in tensor product form, *i.e.*:

$$(I_q\otimes M_x+\Delta t I_q\otimes T_x)\,\underset{\sim}{\gamma}^{n+1}=I_q\otimes M_x\underset{\sim}{\gamma}^{n*}, \quad (3.52)$$

[2]Note that $\hat{\psi}_k$ in the term on the last line of (3.49) must be at time level n for the argument of (3.32) to apply since it relies on the values $\{\hat{\psi}_k^n(\underset{\sim}{x}_m)\}$ interpolating a function in V_h.

where the discretisation matrices are as in (3.21) and (3.22), and I_q is the $N_D \times N_D$ identity matrix.

Equation (3.50) can be written in tensor product matrix form also:

$$\left(M_q \otimes M_x + \frac{\Delta t}{2\mathrm{Wi}} S_q \otimes M_x\right) \underset{\sim}{\gamma}^{n*} = M_q \otimes M_x \underset{\sim}{\gamma}^{n} + \Delta t C(\underset{\approx}{\kappa}^n; \hat{\psi}^n_{h,N}, \zeta_{il}), \tag{3.53}$$

where $\zeta_{il} = X_i \times Y_l \in V_h \otimes \mathcal{P}_N(D)$, for $1 \leq i \leq N_\Omega$ and $1 \leq l \leq N_D$. Also, M_q and S_q are defined in (3.19), (3.20), respectively.

Multiplying (3.52) by $(M_q \otimes I_x + \Delta t/(2\mathrm{Wi}) S_q \otimes I_x)$, where I_x is the $N_\Omega \times N_\Omega$ identity matrix, yields

$$\begin{aligned}\left(M_q \otimes M_x + \Delta t M_q \otimes T_x + \frac{\Delta t}{2\mathrm{Wi}} S_q \otimes M_x + \frac{(\Delta t)^2}{2\mathrm{Wi}} S_q \otimes T_x\right) \underset{\sim}{\gamma}^{n+1} \\ = \left(M_q \otimes M_x + \frac{\Delta t}{2\mathrm{Wi}} S_q \otimes M_x\right) \underset{\sim}{\gamma}^{n*}.\end{aligned} \tag{3.54}$$

Equating the left-hand side of (3.53) with the right-hand side of (3.54) gives:

$$\begin{aligned}\left(M_x \otimes M_q + \Delta t M_q \otimes T_x + \frac{\Delta t}{2\mathrm{Wi}} S_q \otimes M_x + \frac{(\Delta t)^2}{2\mathrm{Wi}} S_q \otimes T_x\right) \underset{\sim}{\gamma}^{n+1} \\ = M_q \otimes M_x \underset{\sim}{\gamma}^{n} + \Delta t C(\underset{\approx}{\kappa}^n; \hat{\psi}^n_{h,N}, \zeta_{il}).\end{aligned} \tag{3.55}$$

Equation (3.55) is equivalent to the inner product form in (3.48) and hence the proof is complete. □

Equation (3.48) will subsequently be referred to as the *equivalent one-step formulation* for method I. Note that (3.48) contains the cross-term,

$$\frac{\Delta t}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}^{n+1}_{h,N}\right), \underset{\sim}{\nabla}_M \zeta\right),$$

which is not present in the weak formulation (3.5). This is analogous to the alternating-direction formulation of the heat equation that was derived in Example 3.2, in which cross-terms of the form

$$\frac{\Delta t}{2}\left(\frac{\partial u_h^{n+1}}{\partial x}\frac{\partial v_h}{\partial y} + \frac{\partial u_h^{n+1}}{\partial y}\frac{\partial v_h}{\partial x}\right) \quad \text{and} \quad \frac{\Delta t}{2}\left(\frac{\partial u_h^{n}}{\partial x}\frac{\partial v_h}{\partial y} + \frac{\partial u_h^{n}}{\partial y}\frac{\partial v_h}{\partial x}\right),$$

were generated.

3.3.3 Method II: Fully-implicit scheme

Method II is very similar to method I, the sole difference being that the term $C(\underset{\approx}{\kappa};\cdot,\cdot)$ is now treated implicitly in time, and therefore we refer to method II as a fully-implicit scheme. We do not discuss the initialisation step or the $\underset{\sim}{x}$-direction scheme here because they are the same as in method I. Instead, we move immediately to discussing the $\underset{\sim}{q}$-direction stage of method II.

Using the line function notation of (3.40), the $\underset{\sim}{q}$-direction numerical method is defined as follows: Given the line functions $\hat{\psi}_k^n \in V_h$, $k = 1, \ldots, N_D$, determine the values $\hat{\psi}_k^{n*}(\underset{\sim}{x}_m)$ satisfying

$$\sum_{k=1}^{N_D} \hat{\psi}_k^{n*}(\underset{\sim}{x}_m) \left(\int_D Y_k(\underset{\sim}{q})\, Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} + \frac{\Delta t}{2\mathrm{Wi}} \int_D \underset{\sim}{\nabla}_M Y_k(\underset{\sim}{q}) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} \right.$$
$$\left. -\Delta t \int_D (\underset{\approx}{\kappa}^{n+1}(\underset{\sim}{x}_m)\, \underset{\sim}{q}\, Y_k(\underset{\sim}{q})) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} \right) = \sum_{k=1}^{N_D} \hat{\psi}_k^n(\underset{\sim}{x}_m) \int_D Y_k(\underset{\sim}{q})\, Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q}, \quad (3.56)$$

for all $l = 1, \ldots, N_D$, and for each quadrature point $\underset{\sim}{x}_m, m = 1, \ldots, Q_\Omega$.

Note that (3.56) is exactly the backward Euler Galerkin spectral method that was studied in Chapter 2. It follows as in Section 2.2 that for Δt sufficiently small the associated bilinear form is coercive, and therefore the linear system defined in (3.56) is non-singular.

Unfortunately we cannot derive an equivalent one-step Galerkin formulation for method II using the same reasoning as in Lemma 3.5 because the proof of that lemma relied on the term $C(\underset{\approx}{\kappa};\cdot,\cdot)$ being explicit-in-time (*cf.* footnote 2). In order to derive a one-step formulation for method II, we would need to recover an integral of $\mathcal{R}\{\psi_k^{n*}(\underset{\sim}{x}_m)\}$ over $\Omega \times D$ by performing the quadrature sum of the discretisation of $C(\underset{\approx}{\kappa};\cdot,\cdot)$ in (3.56). However, this is not possible because this would require a $\underset{\approx}{\kappa}$-weighted reconstruction operator, as distinct from the unweighted reconstruction operator defined in (3.37).

Nevertheless, even without an equivalent one-step formulation, we are still able to prove that method II is stable. This is shown in the next section.

Remark 3.6 *It is possible to modify method II to obtain a Crank-Nicolson scheme, for example, by adding the term*

$$-\frac{1}{2} \sum_{k=1}^{N_D} \hat{\psi}_k^n(\underset{\sim}{x}_m) \left(\frac{\Delta t}{2\mathrm{Wi}} \int_D \underset{\sim}{\nabla}_M Y_k(\underset{\sim}{q}) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} - \Delta t \int_D (\underset{\approx}{\kappa}^n(\underset{\sim}{x}_m)\, \underset{\sim}{q}\, Y_k(\underset{\sim}{q})) \cdot \underset{\sim}{\nabla}_M Y_l(\underset{\sim}{q})\, \mathrm{d}\underset{\sim}{q} \right)$$

to the right-hand side of (3.56)*, as well as adding the term*

$$-\frac{1}{2} \int_\Omega \left(\underset{\sim}{u}^n \cdot \underset{\sim}{\nabla}_x \hat{\psi}_k^n \right) X_i\, \mathrm{d}\underset{\sim}{x},$$

on the right-hand side of the $\underset{\sim}{x}$-direction equation.

However, we are ultimately interested in solving the coupled Navier–Stokes–Fokker–Planck system and, as discussed in Chapter 4, the scheme we use for solving this coupled system introduces an $\mathcal{O}(\Delta t)$ temporal discretisation error. Therefore, there will be no utility in using a Crank-Nicolson discretisation of the Fokker–Planck equation and hence we do not consider this idea any further. $\diamond$

3.4 Stability of methods I and II

First of all, we consider the stability of method I. In this case, the availability of an equivalent one-step method allows the use of standard energy analysis as in the proof of Lemma 3.7 below.

Following Chapter 2, we introduce the following right-hand side forcing terms,

$$\left(\mu^{n+1}, \zeta\right), \quad \left(\underset{\sim}{\nu}^{n+1}, \underset{\sim}{\nabla}_M \zeta\right), \tag{3.57}$$

where $\mu \in \mathrm{L}^2(\Omega \times D)$ and $\underset{\sim}{\nu} \in \mathrm{L}^2(\Omega \times D)^d$. Right-hand side terms of this form will be useful when we derive convergence estimates in Section 3.5.

Lemma 3.7 *If QH1 holds, so that we have the equivalent one-step formulation for method I given in Lemma 3.5, then letting $\Delta t = T/N_T$, $N_T \geq 1$, $\underset{\approx}{\kappa} \in (\mathrm{C}[0,T])^{d\times d}$, $\hat{\psi}^0_{h,N} \in \mathrm{L}^2(\Omega \times D)$, for $\hat{\psi}^s_{h,N} \in V_h \otimes \mathcal{P}_N(D)$ we have the following stability estimate:*

$$\begin{aligned}\|\hat{\psi}^s_{h,N}\|^2 + \sum_{n=0}^{s-1} \Delta t \left\| \frac{\hat{\psi}^{n+1}_{h,N} - \hat{\psi}^n_{h,N}}{\sqrt{\Delta t}} \right\|^2 + \sum_{n=0}^{s-1} \frac{\Delta t}{2\mathrm{Wi}} \|\nabla_M \hat{\psi}^{n+1}_{h,N}\|^2 \\ \leq \mathrm{e}^{Ks\Delta t} \left\{ \|\hat{\psi}^0_{h,N}\|^2 + \sum_{n=0}^{s-1} 2\Delta t \left(\|\mu^{n+1}\|^2 + 4\|\underset{\sim}{\nu}^{n+1}\|^2 \right) \right\},\end{aligned} \tag{3.58}$$

for all s such that $1 \leq s \leq N_T$, where $K = 2(1 + 4\mathrm{Wi}\, b\, |\underset{\approx}{\kappa}|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))})$.

Proof. Consider (3.48) with the right-hand side terms of (3.57):

$$\begin{aligned}\left(\frac{\hat{\psi}^{n+1}_{h,N} - \hat{\psi}^n_{h,N}}{\Delta t}, \zeta \right) + \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}^{n+1}_{h,N}\, , \, \zeta \right) + \frac{1}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M \hat{\psi}^{n+1}_{h,N}\, , \, \underset{\sim}{\nabla}_M \zeta \right) \\ + \frac{\Delta t}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}^{n+1}_{h,N} \right) , \underset{\sim}{\nabla}_M \zeta \right) - \left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\, \hat{\psi}^n_{h,N}\, , \, \underset{\sim}{\nabla}_M \zeta \right) \\ = \left(\mu^{n+1}, \zeta\right) + \left(\underset{\sim}{\nu}^{n+1}\, , \, \underset{\sim}{\nabla}_M \zeta\right),\end{aligned} \tag{3.59}$$

for all $\zeta \in V_h \otimes \mathcal{P}_N(D)$. Set $\zeta = \hat{\psi}_{h,N}^{n+1}$ in (3.59) to get

$$\left(\frac{\hat{\psi}_{h,N}^{n+1} - \hat{\psi}_{h,N}^{n}}{\Delta t}, \hat{\psi}_{h,N}^{n+1}\right) + \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}_{h,N}^{n+1}\,,\, \hat{\psi}_{h,N}^{n+1}\right) + \frac{1}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\|^2$$
$$+ \frac{\Delta t}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}_{h,N}^{n+1}\right)\,,\, \underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\right) - \left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\, \hat{\psi}_{h,N}^{n}\,,\, \underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\right)$$
$$= \left(\mu^{n+1}, \hat{\psi}_{h,N}^{n+1}\right) + \left(\underset{\sim}{\nu}^{n+1}\,,\, \underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\right). \tag{3.60}$$

The $\underset{\sim}{x}$-transport term vanishes because of (3.3) and (3.7). Similarly, the first term on the second line vanishes since

$$\left(\underset{\sim}{\nabla}_M \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}_{h,N}^{n+1}\right)\,,\, \underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\right)$$
$$= \int_{\Omega \times D} M \sum_{j=1}^{d} \left(\sum_{i=1}^{d} u_i \left(\frac{\partial}{\partial x_i} \frac{\partial}{\partial q_j} \frac{\hat{\psi}_{h,N}^{n+1}}{\sqrt{M}} \right) \left(\frac{\partial}{\partial q_j} \frac{\hat{\psi}_{h,N}^{n+1}}{\sqrt{M}} \right) \right) \mathrm{d}\underset{\sim}{x}\, \mathrm{d}\underset{\sim}{q}$$
$$= \frac{1}{2} \int_{\Omega \times D} M \sum_{j=1}^{d} \left(\sum_{i=1}^{d} u_i \frac{\partial}{\partial x_i} \left(\frac{\partial}{\partial q_j} \frac{\hat{\psi}_{h,N}^{n+1}}{\sqrt{M}} \right)^2 \right) \mathrm{d}\underset{\sim}{x}\, \mathrm{d}\underset{\sim}{q}$$
$$= -\frac{1}{2} \int_{\Omega \times D} M \sum_{j=1}^{d} \left((\underset{\sim}{\nabla}_x \cdot \underset{\sim}{u}) \left(\frac{\partial}{\partial q_j} \frac{\hat{\psi}_{h,N}^{n+1}}{\sqrt{M}} \right)^2 \right) \mathrm{d}\underset{\sim}{x}\, \mathrm{d}\underset{\sim}{q} = 0.$$

Applying the identity $2(a-b)a = a^2 - b^2 + (a-b)^2$ to the first term in (3.60), yields

$$\|\hat{\psi}_{h,N}^{n+1}\|^2 + \left\|\hat{\psi}_{h,N}^{n+1} - \hat{\psi}_{h,N}^{n}\right\|^2 + \frac{\Delta t}{\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\|^2 = \|\hat{\psi}_{h,N}^{n}\|^2$$
$$+ 2\Delta t \left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\, \hat{\psi}_{h,N}^{n}, \underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\right) + 2\Delta t \left(\mu^{n+1}, \hat{\psi}_{h,N}^{n+1}\right) + 2\Delta t \left(\underset{\sim}{\nu}^{n+1}\,,\, \underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\right)$$
$$=: \|\hat{\psi}_{h,N}^{n}\|^2 + T_1 + T_2 + T_3. \tag{3.61}$$

Treating T_1, T_2 and T_3 as in the proof of Lemma 2.4, we obtain:

$$(1 - \Delta t)\|\hat{\psi}_{h,N}^{n+1}\|^2 + \Delta t \left\| \frac{\hat{\psi}_{h,N}^{n+1} - \hat{\psi}_{h,N}^{n}}{\sqrt{\Delta t}} \right\|^2 + \frac{\Delta t}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\|^2 \tag{3.62}$$
$$\leq (1 + C_0 \Delta t)\|\hat{\psi}_{h,N}^{n}\|^2 + \Delta t \left(\|\mu^{n+1}\|^2 + 4\|\underset{\sim}{\nu}^{n+1}\|^2 \right),$$

where $C_0 := 4\mathrm{Wi}\, b\, |\underset{\approx}{\kappa}|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}$. Suppose that $\Delta t \leq 0.5$; then

$$\|\hat{\psi}_{h,N}^{n+1}\|^2 + \Delta t \left\| \frac{\hat{\psi}_{h,N}^{n+1} - \hat{\psi}_{h,N}^{n}}{\sqrt{\Delta t}} \right\|^2 + \frac{\Delta t}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \hat{\psi}_{h,N}^{n+1}\|^2$$
$$\leq \frac{1 + C_0 \Delta t}{1 - \Delta t} \|\hat{\psi}_{h,N}^{n}\|^2 + 2\Delta t \left(\|\mu^{n+1}\|^2 + 4\|\underset{\sim}{\nu}^{n+1}\|^2 \right)$$
$$\leq (1 + K\Delta t)\|\hat{\psi}_{h,N}^{n}\|^2 + 2\Delta t \left(\|\mu^{n+1}\|^2 + 4\|\underset{\sim}{\nu}^{n+1}\|^2 \right),$$

where $K = 2(1 + C_0) = 2(1 + 4\mathrm{Wi}\, b\, |\underset{\approx}{\kappa}|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))})$.

Summing over $n = 0, \ldots, s-1$ gives,

$$\|\hat{\psi}^s_{h,N}\|^2 + \sum_{n=0}^{s-1} \Delta t \left\| \frac{\hat{\psi}^{n+1}_{h,N} - \hat{\psi}^n_{h,N}}{\sqrt{\Delta t}} \right\|^2 + \sum_{n=0}^{s-1} \frac{\Delta t}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \hat{\psi}^{n+1}_{h,N}\|^2$$
$$\leq \left\{ \|\hat{\psi}^0_{h,N}\|^2 + \sum_{n=0}^{s-1} 2\Delta t \left(\|\mu^{n+1}\|^2 + 4\|\underset{\sim}{\nu}^{n+1}\|^2 \right) \right\} + K \sum_{n=0}^{s-1} \Delta t \|\hat{\psi}^n_{h,N}\|^2,$$

and applying a discrete Gronwall lemma yields (3.58). □

We cannot apply an analogous argument for method II due to the absence of an equivalent one-step method. However, by combining stability results for the $\underset{\sim}{q}$-direction and $\underset{\sim}{x}$-direction methods we can establish the stability of method II, as shown in the next lemma.

Lemma 3.8 *Suppose QH2 is satisfied and let $\Delta t = T/N_T$, $N_T \geq 1$. Then for $\hat{\psi}^n_{h,N} \in V_h \otimes \mathcal{P}_N(D)$ computed using alternating-direction method II we have*

$$\|\hat{\psi}^n_{h,N}\| \leq \mathrm{e}^{c_0 n \Delta t} \|\hat{\psi}^0_{h,N}\|. \tag{3.63}$$

for $1 \leq n \leq N_T$, where $c_0 := 1 + 4\mathrm{Wi}\, b\, |\underset{\approx}{\kappa}|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}$.

Proof. From the proof of Lemma 2.4, we have the following bound for (3.56) at a given quadrature point $\underset{\sim}{x}_m \in \overline{\Omega}$,

$$\|\hat{\psi}^{n*}(\underset{\sim}{x}_m, \cdot)\|^2_{\mathrm{L}^2(D)} \leq (1 + 2c_0\Delta t) \|\hat{\psi}^n(\underset{\sim}{x}_m, \cdot)\|^2_{\mathrm{L}^2(D)}. \tag{3.64}$$

Rewriting (3.64) in terms of a basis $\{Y_1, \ldots, Y_{N_D}\}$ of $\mathcal{P}_N(D)$, which, without loss of generality may be assumed to be orthogonal in the $\mathrm{L}^2(D)$ inner product, we obtain:

$$\sum_{k=1}^{N_D} \hat{\psi}^{n*}_k(\underset{\sim}{x}_m)^2 \|Y_k\|^2_{\mathrm{L}^2(D)} \leq (1 + 2c_0\Delta t) \sum_{k=1}^{N_D} \hat{\psi}^n_k(\underset{\sim}{x}_m)^2 \|Y_k\|^2_{\mathrm{L}^2(D)}. \tag{3.65}$$

Using (3.26) to sum (3.65) for $m = 1, \ldots, Q_\Omega$, and then employing (3.36), we have

$$\sum_{k=1}^{N_D} \|\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|^2_{\ell^2(\Omega)} \|Y_k\|^2_{\mathrm{L}^2(D)} \leq (1 + 2c_0\Delta t) \sum_{k=1}^{N_D} \|\{\hat{\psi}^n_k(\underset{\sim}{x}_m)\}\|^2_{\ell^2(\Omega)} \|Y_k\|^2_{\mathrm{L}^2(D)}. \tag{3.66}$$

Since $\hat{\psi}^n_{h,N} \in V_h \otimes \mathcal{P}_N(D)$, it follows that $\hat{\psi}^n_k \in V_h$, and therefore (as observed below (3.36)) the discrete $\ell^2(\Omega)$ norm on the right-hand side above is equal to the continuous $\mathrm{L}^2(\Omega)$ norm, so that

$$\begin{aligned}\sum_{k=1}^{N_D} \|\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|^2_{\ell^2(\Omega)} \, \|Y_k\|^2_{\mathrm{L}^2(D)} &\leq (1+2c_0\Delta t) \sum_{k=1}^{N_D} \|\hat{\psi}^n_k\|^2_{\mathrm{L}^2(\Omega)} \, \|Y_k\|^2_{\mathrm{L}^2(D)} \\ &= (1+2c_0\Delta t)\|\hat{\psi}^n_{h,N}\|^2. \end{aligned} \tag{3.67}$$

Also, by (3.3) and (3.7), it follows easily from (3.41) that:

$$\|\hat{\psi}^{n+1}_k\|^2_{\mathrm{L}^2(\Omega)} \leq \|\mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|^2_{\mathrm{L}^2(\Omega)}, \tag{3.68}$$

for each k. Multiplying through by $\|Y_k\|^2_{\mathrm{L}^2(D)}$ in (3.68) and summing over $k = 1, \ldots, N_D$ gives

$$\|\hat{\psi}^{n+1}_{h,N}\|^2 = \sum_{k=1}^{N_D} \|\hat{\psi}^{n+1}_k\|^2_{\mathrm{L}^2(\Omega)} \|Y_k\|^2_{\mathrm{L}^2(D)} \leq \sum_{k=1}^{N_D} \|\mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|^2_{\mathrm{L}^2(\Omega)} \|Y_k\|^2_{\mathrm{L}^2(D)}. \tag{3.69}$$

By taking $\{f_m\} = \{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}$ and $X = \mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\} \in V_h$ in (3.37) and applying the Cauchy-Schwarz inequality in the ℓ^2 inner product, we have

$$\begin{aligned}\|\mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|^2_{\mathrm{L}^2(\Omega)} &= \Big(\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}, \{\mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}(\underset{\sim}{x}_m)\}\Big)_{\ell^2(\Omega)} \\ &\leq \|\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|_{\ell^2(\Omega)} \, \|\mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|_{\ell^2(\Omega)} \\ &= \|\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|_{\ell^2(\Omega)} \, \|\mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|_{\mathrm{L}^2(\Omega)}, \end{aligned}$$

and therefore,

$$\|\mathcal{R}\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|_{\mathrm{L}^2(\Omega)} \leq \|\{\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)\}\|_{\ell^2(\Omega)}. \tag{3.70}$$

Combining (3.67), (3.69) and (3.70), gives,

$$\|\hat{\psi}^{n+1}_{h,N}\|^2 \leq (1+2c_0\Delta t)\|\hat{\psi}^n_{h,N}\|^2, \tag{3.71}$$

from which (3.63) follows easily on noting that $1 + 2c_0\Delta t \leq \mathrm{e}^{2c_0\Delta t}$. $\square$

Remark 3.9 *The argument in Lemma 3.8 can also be applied to method I and hence it follows that method I is stable when only QH2 is satisfied.*

3.5 Convergence analysis for method I, Part 1

In this section, the equivalent one-step scheme (3.48) and Lemma 3.7 are used to prove that the numerical solution obtained using method I converges to the weak solution of (3.5), (3.6). The convergence argument presented here is analogous to the approach in Section 2.3. Note that we need access to an equivalent one-step formulation to use this approach in the context of alternating-direction methods, and therefore we only consider the convergence analysis of method I.

Let $\hat{\psi}(\cdot,\cdot,t)$ be the weak solution of (3.5), (3.6) at time $t \in (0,T)$. To simplify the notation, we write $\hat{\psi}(t) := \hat{\psi}(\cdot,\cdot,t)$ throughout the rest of this section. As in Section 2.3, we define

$$e_{h,N}^n := \hat{\psi}(t^n) - \hat{\psi}_{h,N}^n = (\hat{\psi}(t^n) - \Pi\hat{\psi}(t^n)) + (\Pi\hat{\psi}(t^n) - \hat{\psi}_{h,N}^n) =: \eta^n + \xi^n,$$

where Π is a projection operator that projects onto $V_h \otimes \mathcal{P}_N(D)$. Π shall be defined later.

Noting that $\xi^n \in V_h \otimes \mathcal{P}_N(D)$, we apply the equivalent one-step formulation for method I, (3.48), to $\xi^n = \hat{\psi}(t^n) - \hat{\psi}_{h,N}^n - \eta^n$ and set $\zeta = \xi^{n+1}$, to obtain:

$$\begin{aligned}
&\left(\frac{\xi^{n+1}-\xi^n}{\Delta t}, \xi^{n+1}\right) + \left(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\xi^{n+1}, \xi^{n+1}\right) + \frac{1}{2\mathrm{Wi}}\|\underset{\sim}{\nabla}_M\xi^{n+1}\|^2 \\
&\qquad + \frac{\Delta t}{2\mathrm{Wi}}\left(\underset{\sim}{\nabla}_M(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\xi^{n+1}), \underset{\sim}{\nabla}_M\xi^{n+1}\right) - \left(\underset{\approx}{\kappa}^n\underset{\sim}{q}\xi^n, \underset{\sim}{\nabla}_M\xi^{n+1}\right) \\
&= \left(\frac{\hat{\psi}(t^{n+1})-\hat{\psi}(t^n)}{\Delta t}, \xi^{n+1}\right) + \left(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\hat{\psi}(t^{n+1}), \xi^{n+1}\right) + \frac{1}{2\mathrm{Wi}}\left(\underset{\sim}{\nabla}_M\hat{\psi}(t^{n+1}), \underset{\sim}{\nabla}_M\xi^{n+1}\right) \\
&\qquad + \frac{\Delta t}{2}\left(\underset{\sim}{\nabla}_M(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\hat{\psi}(t^{n+1})), \underset{\sim}{\nabla}_M\xi^{n+1}\right) - \left(\underset{\approx}{\kappa}^n\underset{\sim}{q}\hat{\psi}(t^n), \underset{\sim}{\nabla}_M\xi^{n+1}\right) \\
&- \left(\frac{\eta^{n+1}-\eta^n}{\Delta t}, \xi^{n+1}\right) - \left(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\eta^{n+1}, \xi^{n+1}\right) - \frac{1}{2\mathrm{Wi}}\left(\underset{\sim}{\nabla}_M\eta^{n+1}, \underset{\sim}{\nabla}_M\xi^{n+1}\right) \\
&\qquad - \frac{\Delta t}{2\mathrm{Wi}}\left(\underset{\sim}{\nabla}_M(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\eta^{n+1}), \underset{\sim}{\nabla}_M\xi^{n+1}\right) + \left(\underset{\approx}{\kappa}^n\underset{\sim}{q}\eta^n, \underset{\sim}{\nabla}_M\xi^{n+1}\right),
\end{aligned} \tag{3.72}$$

where the terms containing $\hat{\psi}_{h,N}^n$ and $\hat{\psi}_{h,N}^{n+1}$ vanish since $\hat{\psi}_{h,N}$ satisfies (3.48).

First of all we use the identities

$$\underset{\approx}{\kappa}^n = \underset{\approx}{\kappa}^{n+1} - \int_{t^n}^{t^{n+1}} \frac{\partial\underset{\approx}{\kappa}}{\partial t}\,\mathrm{d}t \qquad \text{and} \qquad \hat{\psi}^n = \hat{\psi}^{n+1} - \int_{t^n}^{t^{n+1}} \frac{\partial\hat{\psi}}{\partial t}\,\mathrm{d}t,$$

to obtain:

$$
\begin{aligned}
&\left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\hat{\psi}(t^n), \underset{\sim}{\nabla}_M \xi^{n+1}\right) \\
&= \left(\underset{\approx}{\kappa}^{n+1} \underset{\sim}{q}\hat{\psi}(t^{n+1}), \underset{\sim}{\nabla}_M \xi^{n+1}\right) - \left(\left(\int_{t^n}^{t^{n+1}} \frac{\partial \underset{\approx}{\kappa}}{\partial t}\,\mathrm{d}t\right) \underset{\sim}{q}\hat{\psi}(t^n), \underset{\sim}{\nabla}_M \xi^{n+1}\right) \\
&- \left(\underset{\approx}{\kappa}^{n+1} \underset{\sim}{q}\left(\int_{t^n}^{t^{n+1}} \frac{\partial \hat{\psi}}{\partial t}\,\mathrm{d}t\right), \underset{\sim}{\nabla}_M \xi^{n+1}\right) + \left(\left(\int_{t^n}^{t^{n+1}} \frac{\partial \underset{\approx}{\kappa}}{\partial t}\,\mathrm{d}t\right) \underset{\sim}{q}\left(\int_{t^n}^{t^{n+1}} \frac{\partial \hat{\psi}}{\partial t}\,\mathrm{d}t\right), \underset{\sim}{\nabla}_M \xi^{n+1}\right) \\
&=: \left(\underset{\approx}{\kappa}^{n+1} \underset{\sim}{q}\hat{\psi}(t^{n+1}), \underset{\sim}{\nabla}_M \xi^{n+1}\right) - \left(\underset{\sim}{K}_1, \underset{\sim}{\nabla}_M \xi^{n+1}\right) - \left(\underset{\sim}{K}_2, \underset{\sim}{\nabla}_M \xi^{n+1}\right) + \left(\underset{\sim}{K}_3, \underset{\sim}{\nabla}_M \xi^{n+1}\right).
\end{aligned}
$$

Now, considering only the terms containing $\hat{\psi}$ on the right-hand side of (3.72), we have:

$$
\begin{aligned}
&\left(\frac{\hat{\psi}(t^{n+1}) - \hat{\psi}(t^n)}{\Delta t}, \xi^{n+1}\right) + \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}(t^{n+1}), \xi^{n+1}\right) + \frac{1}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M \hat{\psi}(t^{n+1}), \underset{\sim}{\nabla}_M \xi^{n+1}\right) \\
&\qquad + \frac{\Delta t}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}(t^{n+1})), \underset{\sim}{\nabla}_M \xi^{n+1}\right) - \left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\hat{\psi}(t^n), \underset{\sim}{\nabla}_M \xi^{n+1}\right) \\
&= \left(\frac{\hat{\psi}(t^{n+1}) - \hat{\psi}(t^n)}{\Delta t} - \frac{\partial \hat{\psi}}{\partial t}(t^{n+1}), \xi^{n+1}\right) + \frac{\Delta t}{2\mathrm{Wi}} \left(\underset{\sim}{\nabla}_M (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}(t^{n+1})), \underset{\sim}{\nabla}_M \xi^{n+1}\right) \\
&\qquad + \left(\underset{\sim}{K}_1, \underset{\sim}{\nabla}_M \xi^{n+1}\right) + \left(\underset{\sim}{K}_2, \underset{\sim}{\nabla}_M \xi^{n+1}\right) - \left(\underset{\sim}{K}_3, \underset{\sim}{\nabla}_M \xi^{n+1}\right),
\end{aligned} \tag{3.73}
$$

where the fact that $\hat{\psi}$ satisfies (3.5), and the expansion of the term $\left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\hat{\psi}(t^n), \underset{\sim}{\nabla}_M \xi^{n+1}\right)$ from above, have been used on the right-hand side. Using (3.73) on the right-hand side of (3.72), we have:

$$
\begin{aligned}
&\left(\frac{\xi^{n+1} - \xi^n}{\Delta t}, \xi^{n+1}\right) + \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \xi^{n+1}, \xi^{n+1}\right) + \frac{1}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \xi^{n+1}\|^2 \\
&\qquad + \frac{\Delta t}{2} \left(\underset{\sim}{\nabla}_M (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \xi^{n+1}), \underset{\sim}{\nabla}_M \xi^{n+1}\right) - \left(\underset{\approx}{\kappa}^n \underset{\sim}{q}\xi^n, \underset{\sim}{\nabla}_M \xi^{n+1}\right) \\
&= \left(\mu^{n+1}, \xi^{n+1}\right) + \left(\underset{\sim}{\nu}^{n+1}, \underset{\sim}{\nabla}_M \xi^{n+1}\right),
\end{aligned} \tag{3.74}
$$

where

$$
\mu^{n+1} := \frac{\hat{\psi}(t^{n+1}) - \hat{\psi}(t^n)}{\Delta t} - \frac{\partial \hat{\psi}}{\partial t}(t^{n+1}) - \frac{\eta^{n+1} - \eta^n}{\Delta t} - \underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \eta^{n+1}, \tag{3.75}
$$

$$
\begin{aligned}
\underset{\sim}{\nu}^{n+1} := {} & \frac{\Delta t}{2\mathrm{Wi}} \underset{\sim}{\nabla}_M (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}(t^{n+1})) + \underset{\sim}{K}_1 + \underset{\sim}{K}_2 - \underset{\sim}{K}_3 - \frac{1}{2\mathrm{Wi}} \underset{\sim}{\nabla}_M \eta^{n+1} \\
& - \frac{\Delta t}{2\mathrm{Wi}} \underset{\sim}{\nabla}_M (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \eta^{n+1}) + \underset{\approx}{\kappa}^n \underset{\sim}{q}\eta^n.
\end{aligned} \tag{3.76}
$$

Therefore, applying the stability result (3.58) to (3.74) gives

$$
\|\xi^n\|^2 + \sum_{m=0}^{n-1} \frac{\Delta t}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \xi^{m+1}\|^2 \leq \mathrm{e}^{Kn\Delta t} \left\{ \|\xi^0\|^2 + \sum_{m=0}^{n-1} 2\Delta t \left(\|\mu^{m+1}\|^2 + 4\|\underset{\sim}{\nu}^{m+1}\|^2\right) \right\}. \tag{3.77}
$$

The next step is to bound the right-hand side of (3.77) in terms of norms of η and $\hat{\psi}$.

First of all, just as in Section 2.3, we have that $\|\xi^0\| \leq \|\eta^0\|$. Next we consider $\|\mu^{m+1}\|$:

$$\begin{aligned}\|\mu^{m+1}\|^2 &\leq 3\left\|\frac{\hat{\psi}(t^{m+1})-\hat{\psi}(t^m)}{\Delta t}-\frac{\partial\hat{\psi}}{\partial t}(t^{m+1})\right\|^2+3\left\|\frac{\eta^{m+1}-\eta^m}{\Delta t}\right\|^2+3\|\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\eta^{m+1}\|^2\\ &=: 3\,(I+II+III).\end{aligned} \tag{3.78}$$

For term I, applying Taylor's theorem with integral remainder yields

$$I \leq \Delta t\int_{t^m}^{t^{m+1}}\left\|\frac{\partial^2\hat{\psi}}{\partial t^2}(\cdot,\cdot,t)\right\|^2\,\mathrm{d}t,$$

and for term II we have the following bound:

$$II \leq \int_{\Omega\times D}\frac{1}{\Delta t}\int_{t^m}^{t^{m+1}}\left|\frac{\partial\eta}{\partial t}(\underset{\sim}{x},\underset{\sim}{q},t)\right|^2\,\mathrm{d}t\,\mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}=\frac{1}{\Delta t}\int_{t^m}^{t^{m+1}}\left\|\frac{\partial\eta}{\partial t}(\cdot,\cdot,t)\right\|^2\,\mathrm{d}t.$$

Term III is simple to bound by pulling out the supremum of $\underset{\sim}{u}$, as follows:

$$III=\int_{\Omega\times D}\left(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\eta^{m+1}\right)^2\,\mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\leq\|\underset{\sim}{u}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\|\underset{\sim}{\nabla}_x\eta^{m+1}\|^2. \tag{3.79}$$

Therefore,

$$\begin{aligned}\sum_{m=0}^{n-1}2\Delta t\|\mu^{m+1}\|^2 &\leq 6\sum_{m=0}^{n-1}\Delta t^2\int_{t^m}^{t^{m+1}}\left\|\frac{\partial^2\hat{\psi}}{\partial t^2}(\cdot,\cdot,t)\right\|^2\,\mathrm{d}t+6\sum_{m=0}^{n-1}\int_{t^m}^{t^{m+1}}\left\|\frac{\partial\eta}{\partial t}(\cdot,\cdot,t)\right\|^2\,\mathrm{d}t\\ &\quad+6\|\underset{\sim}{u}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\sum_{m=0}^{n-1}\Delta t\|\underset{\sim}{\nabla}_x\eta^{m+1}\|^2\\ &= 6\Delta t^2\left\|\frac{\partial^2\hat{\psi}}{\partial t^2}\right\|^2_{\mathrm{L}^2(0,t^m;\mathrm{L}^2(\Omega\times D))}+6\left\|\frac{\partial\eta}{\partial t}\right\|^2_{\mathrm{L}^2(0,t^m;\mathrm{L}^2(\Omega\times D))}\\ &\quad+6\|\underset{\sim}{u}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\|\underset{\sim}{\nabla}_x\eta\|^2_{\ell^2(0,t^m;\mathrm{L}^2(\Omega\times D))}.\end{aligned} \tag{3.80}$$

Next we derive upper bounds for the norms of the terms on the right-hand side of

(3.76). First of all, we consider the cross-term,

$$\begin{aligned}
\|\underset{\sim}{\nabla}_M(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\hat{\psi}(t^{n+1}))\|^2 &= \int_{\Omega\times D}\left|\underset{\sim}{\nabla}_M\left(\sum_{i=1}^d u_i\frac{\partial}{\partial x_i}\hat{\psi}(t^{n+1})\right)\right|^2 \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\\
&= \int_{\Omega\times D}\sum_{j=1}^d\left\{\sqrt{M}\frac{\partial}{\partial q_j}\left(\sum_{i=1}^d u_i\frac{\partial}{\partial x_i}\left(\frac{\hat{\psi}(t^{n+1})}{\sqrt{M}}\right)\right)\right\}^2 \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\\
&= \int_{\Omega\times D}\sum_{j=1}^d\left\{\sum_{i=1}^d u_i\frac{\partial}{\partial x_i}\left(\sqrt{M}\frac{\partial}{\partial q_j}\left(\frac{\hat{\psi}(t^{n+1})}{\sqrt{M}}\right)\right)\right\}^2 \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\\
&= \int_{\Omega\times D}\sum_{j=1}^d\left\{\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\left(\sqrt{M}\frac{\partial}{\partial q_j}\left(\frac{\hat{\psi}(t^{n+1})}{\sqrt{M}}\right)\right)\right\}^2 \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\\
&\le \int_{\Omega\times D}\sum_{j=1}^d\left(|\underset{\sim}{u}|^2\left|\underset{\sim}{\nabla}_x\left(\sqrt{M}\frac{\partial}{\partial q_j}\left(\frac{\hat{\psi}(t^{n+1})}{\sqrt{M}}\right)\right)\right|^2\right) \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\\
&\le \|u\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\psi(t^{n+1})\|^2. \qquad (3.81)
\end{aligned}$$

By the same reasoning as in (3.81), it follows that:

$$\|\underset{\sim}{\nabla}_M(\underset{\sim}{u}\cdot\underset{\sim}{\nabla}_x\eta^{n+1})\|^2 \le \|u\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\eta^{n+1}\|^2. \qquad (3.82)$$

Also, we have

$$\|\underset{\approx}{\kappa}^n\underset{\sim}{q}\eta^n\|^2 \le b\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\,\|\eta^n\|^2, \qquad (3.83)$$

and finally it remains to bound the norms of $\underset{\sim}{K}_1, \underset{\sim}{K}_2$ and $\underset{\sim}{K}_3$, for which we have,

$$\begin{aligned}
\|\underset{\sim}{K}_1\|^2 &= \int_{\Omega\times D}\left\{\left(\int_{t^n}^{t^{n+1}}\frac{\partial\underset{\approx}{\kappa}}{\partial t}\underset{\sim}{q}\,\mathrm{d}t\right)\hat{\psi}(t^n)\right\}^2 \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q} \qquad (3.84)\\
&\le \Delta t^2 b\left\|\frac{\partial\underset{\approx}{\kappa}}{\partial t}\right\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\left\|\hat{\psi}(t^n)\right\|^2,
\end{aligned}$$

$$\begin{aligned}
\|\underset{\sim}{K}_2\|^2 &= \int_{\Omega\times D}\left\{\underset{\approx}{\kappa}^{n+1}\underset{\sim}{q}\left(\int_{t^n}^{t^{n+1}}\frac{\partial\hat{\psi}}{\partial t}\,\mathrm{d}t\right)\right\}^2 \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\\
&\le \Delta t\, b\|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\int_{t^n}^{t^{n+1}}\left\|\frac{\partial\hat{\psi}}{\partial t}\right\|^2 \mathrm{d}t, \qquad (3.85)
\end{aligned}$$

and

$$\begin{aligned}
\|\underset{\sim}{K}_3\|^2 &= \int_{\Omega\times D}\left\{\left(\int_{t^n}^{t^{n+1}}\frac{\partial\underset{\approx}{\kappa}}{\partial t}\underset{\sim}{q}\,\mathrm{d}t\right)\left(\int_{t^n}^{t^{n+1}}\frac{\partial\hat{\psi}}{\partial t}\,\mathrm{d}t\right)\right\}^2 \mathrm{d}\underset{\sim}{x}\,\mathrm{d}\underset{\sim}{q}\\
&\le \Delta t^3 b\left\|\frac{\partial\underset{\approx}{\kappa}}{\partial t}\right\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))}\int_{t^n}^{t^{n+1}}\left\|\frac{\partial\hat{\psi}}{\partial t}\right\|^2 \mathrm{d}t, \qquad (3.86)
\end{aligned}$$

and it is convenient to bound $\underset{\sim}{K}_2$ and $\underset{\sim}{K}_3$ together as follows:

$$\|\underset{\sim}{K}_2\|^2 + \|\underset{\sim}{K}_3\|^2 \leq b\,\Delta t\,\|\underset{\approx}{\kappa}\|^2_{\mathrm{W}^{1,\infty}(0,T;\mathrm{L}^\infty(\Omega))} \int_{t^n}^{t^{n+1}} \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|^2 \mathrm{d}t.$$

Therefore,

$$\begin{aligned}
&\sum_{m=0}^{n-1} 8\Delta t \|\underset{\sim}{\nu}^{m+1}\|^2 \\
&\leq \sum_{m=0}^{n-1} 56\Delta t \left(\frac{\Delta t^2}{4\mathrm{Wi}^2} \left\| \underset{\sim}{\nabla}_M (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}(t^{m+1})) \right\|^2 + \frac{\Delta t^2}{4\mathrm{Wi}^2} \|\underset{\sim}{\nabla}_M (\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \eta^{m+1})\|^2 \right. \\
&\qquad \left. + \frac{1}{4\mathrm{Wi}^2} \|\underset{\sim}{\nabla}_M \eta^{m+1}\|^2 + \|\underset{\approx}{\kappa}^m \underset{\sim}{q} \eta^m\|^2 + \|\underset{\sim}{K}_1\|^2 + \|\underset{\sim}{K}_2\|^2 + \|\underset{\sim}{K}_3\|^2 \right) \\
&\leq \sum_{m=0}^{n-1} 56\Delta t \left(\frac{\Delta t^2}{4\mathrm{Wi}^2} \|u\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \left(\|\underset{\sim}{\nabla}_x \underset{\sim}{\nabla}_M \hat{\psi}(t^{m+1})\|^2 + \|\underset{\sim}{\nabla}_x \underset{\sim}{\nabla}_M \eta^{m+1}\|^2 \right) \right. \\
&\qquad + \frac{1}{4\mathrm{Wi}^2} \|\underset{\sim}{\nabla}_M \eta^{n+1}\|^2 + b \|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \|\eta^n\|^2 \\
&\qquad \left. + \Delta t^2 b \left\| \frac{\partial \underset{\approx}{\kappa}}{\partial t} \right\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \left\| \hat{\psi}(t^m) \right\|^2 + \Delta t\, b \|\underset{\approx}{\kappa}\|^2_{\mathrm{W}^{1,\infty}(0,T;\mathrm{L}^\infty(\Omega))} \int_{t^m}^{t^{m+1}} \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|^2 \mathrm{d}t \right) \\
&= \frac{14}{\mathrm{Wi}^2} \Delta t^2 \|u\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \left(\|\underset{\sim}{\nabla}_x \underset{\sim}{\nabla}_M \hat{\psi}\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} + \|\underset{\sim}{\nabla}_x \underset{\sim}{\nabla}_M \eta\|^2_{\ell^2(0,t^n;\ell^2(\Omega\times D))} \right) \\
&\qquad + \frac{14}{\mathrm{Wi}^2} \|\underset{\sim}{\nabla}_M \eta\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} + 56\, b \|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \|\eta\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} \\
&\qquad + 56\, b \Delta t^2 \left\| \frac{\partial \underset{\approx}{\kappa}}{\partial t} \right\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \left\| \hat{\psi} \right\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} \\
&\qquad + 56\, \Delta t^2\, b \|\underset{\approx}{\kappa}\|^2_{\mathrm{W}^{1,\infty}(0,T;\mathrm{L}^\infty(\Omega))} \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|^2_{\mathrm{L}^2(0,t^n;L^2(\Omega\times D))}.
\end{aligned} \tag{3.87}$$

We now combine the bounds in (3.77), (3.80) and (3.87) to get:

$$
\begin{aligned}
&\|\xi^n\|^2 + \sum_{m=0}^{n-1} \frac{\Delta t}{2\mathrm{Wi}} \|\underset{\sim}{\nabla}_M \xi^{m+1}\|^2 \\
&\qquad \leq \mathrm{e}^{Kn\Delta t} \Bigg\{ \|\eta^0\|^2 + 6\Delta t^2 \left\| \frac{\partial^2 \hat{\psi}}{\partial t^2} \right\|^2_{\mathrm{L}^2(0,t^n;\mathrm{L}^2(\Omega\times D))} + 6 \left\| \frac{\partial \eta}{\partial t} \right\|^2_{\mathrm{L}^2(0,t^n;\mathrm{L}^2(\Omega\times D))} \\
&\quad + 6\|\underset{\sim}{u}\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \|\underset{\sim}{\nabla}_x \eta\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} \\
&\quad + \frac{14}{\mathrm{Wi}^2} \Delta t^2 \|u\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \left(\|\underset{\sim}{\nabla}_x \underset{\sim}{\nabla}_M \hat{\psi}\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} + \|\underset{\sim}{\nabla}_x \underset{\sim}{\nabla}_M \eta\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} \right) \\
&\quad + \frac{14}{\mathrm{Wi}^2} \|\underset{\sim}{\nabla}_M \eta\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} + 56\, b \|\underset{\approx}{\kappa}\|^2_{\mathrm{L}^\infty(0,t^n;\mathrm{L}^\infty(\Omega))} \|\eta\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} \\
&\quad + 56\, b \Delta t^2 \left\| \frac{\partial \underset{\approx}{\kappa}}{\partial t} \right\|^2_{\mathrm{L}^\infty(0,T;\mathrm{L}^\infty(\Omega))} \left\| \hat{\psi} \right\|^2_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} \\
&\quad + 56\, \Delta t^2\, b \|\underset{\approx}{\kappa}\|^2_{\mathrm{W}^{1,\infty}(0,T;\mathrm{L}^\infty(\Omega))} \left\| \frac{\partial \hat{\psi}}{\partial t} \right\|^2_{\mathrm{L}^2(0,t^n;L^2(\Omega\times D))} \Bigg\}.
\end{aligned} \tag{3.88}
$$

Now, just as in Chapter 2, we need to bound the terms containing η in (3.88). This is considered in the next section.

3.6 Approximation results on $\Omega \times D$

In order to use the approximation results from Section 2.4, we restrict our attention to the $d = 2$ case here although, of course, analogus results could be obtained for the $d = 3$ case. We denote the projection operator considered in Section 2.4 (referred to there as $\hat{\Pi}_N$) by $\Pi_q : \mathcal{H}^{1,1}(D) \to \mathcal{P}_N(D)$. Also, we consider a quasi-interpolation operator, $\mathcal{I}_x : \mathrm{L}^1(\Omega) \to V_h$, which is a generalisation of the standard finite element interpolant such that the quasi-interpolant is well-defined for non-smooth functions; we refer to Section 4.8 of [19] for the details of the definition of this operator (alternatively, see [25] or [74]).

We have the following result for $\mathcal{I}_x$ (*cf.* Theorem (4.8.12) in [19]):

Theorem 3.10 *Suppose that $\mathcal{T}_h$ is non-degenerate in the sense that there exists $\rho > 0$ such that for all $K \in \mathcal{T}_h$, $\mathrm{diam}(B_K) \geq \rho\, \mathrm{diam}(K)$, where B_K is the largest ball contained in K. Suppose also that the set of shape functions for each element $K \in \mathcal{T}_h$ contains all polynomials of degree less than m. Then, there exists a positive constant C such that*

$$
\left(\sum_{K\in\mathcal{T}_h} h_K^{p(s-k)} \|v - \mathcal{I}_x\, v\|^p_{\mathrm{W}^{s,p}(K)} \right)^{1/p} \leq C |v|_{\mathrm{W}^{k,p}(\Omega)},
$$

for all $v \in \mathrm{W}^{k,p}(\Omega)$, $0 \le k \le m$, $1 \le p \le \infty$, $0 \le s \le k$, *where* $h_K := \mathrm{diam}(K)$.

Corollary 3.11 *(*cf. *Corollary 4.8.15 in [19]) Setting* $s = k$ *in Theorem 3.10, it follows that*

$$\|\mathcal{I}_x\, v\|_{\mathrm{W}^{k,p}(\Omega)} \le C|v|_{\mathrm{W}^{k,p}(\Omega)} \qquad \forall v \in \mathrm{W}^{k,p}(\Omega), \tag{3.89}$$

for $0 \le s, k \le m$, *where* m *is as in Theorem 3.10, and* $1 \le p \le \infty$. *Also, letting* $h = \max_{K\in\mathcal{T}_h}\mathrm{diam}(K)$ *in Theorem 3.10, we obtain*

$$\|v - \mathcal{I}_x v\|_{\mathrm{W}^{s,p}(\Omega)} \le Ch^{k-s}|v|_{\mathrm{W}^{k,p}(\Omega)}, \tag{3.90}$$

for $0 \le s \le k$, $0 \le k \le m$, *and* m, p *as in* (3.89).

For the projection operator Π_q, recall from Section 2.4 that:

$$\|\hat{\psi} - \Pi_q\hat{\psi}\|_{\mathrm{H}^1_0(D;M)} \le C_1 N_r^{-k}\|\hat{\psi}\|_{\mathcal{H}^{k+1}_r(D)} \;+\; C_2 N_\theta^{-l}\|\hat{\psi}\|_{\mathcal{H}^{l+1}_\theta(D)}, \tag{3.91}$$

and

$$\|\hat{\psi} - \Pi_q\hat{\psi}\|_{\mathrm{L}^2(D)} \le C_1 N_r^{-k}\|\hat{\psi}\|_{\mathcal{H}^{k}_r(D)} \;+\; C_2 N_\theta^{-l}\|\hat{\psi}\|_{\mathcal{H}^{l}_\theta(D)}. \tag{3.92}$$

Now, let the projection operator $\Pi : \mathrm{L}^1(\Omega; \mathcal{H}^{1,1}(D)) \to V_h \otimes \mathcal{P}_N(D)$ be defined as

$$\Pi := \mathcal{I}_x\, \Pi_q = \Pi_q\, \mathcal{I}_x,$$

so that $\eta := \hat{\psi} - \Pi\hat{\psi}$. We will use the approximation properties listed above for Π_q and $\mathcal{I}_x$ to derive bounds for the terms $\|\eta\|$, $\|\underset{\sim}{\nabla}_x\eta\|$, $\|\underset{\sim}{\nabla}_M\eta\|$ and $\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\eta\|$ that appear on the right-hand side of (3.88).

First of all, consider $\|\eta\|$:

$$\|\eta\| = \|\hat{\psi} - \mathcal{I}_x\Pi_q\hat{\psi}\| \le \|\hat{\psi} - \mathcal{I}_x\hat{\psi}\| + \|\mathcal{I}_x\hat{\psi} - \Pi_q\mathcal{I}_x\hat{\psi}\| =: I + II.$$

From (3.90), we have that

$$I = \left(\int_D \|\hat{\psi} - \mathcal{I}_x\hat{\psi}\|^2_{\mathrm{L}^2(\Omega)}\, \mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}} \le Ch^s \left(\int_D |\hat{\psi}|^2_{\mathrm{H}^s(\Omega)}\, \mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}}.$$

Also,

$$\begin{aligned} II &= \left(\int_\Omega \|\mathcal{I}_x\hat{\psi} - \Pi_q\mathcal{I}_x\hat{\psi}\|^2_{\mathrm{L}^2(D)}\, \mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \\ &\le C_1 N_r^{-k}\left(\int_\Omega \|\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^k_r(D)}\, \mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \;+\; C_2 N_\theta^{-l}\left(\int_\Omega \|\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^l_\theta(D)}\, \mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \\ &\le C_1 N_r^{-k}\left(\int_\Omega \|\hat{\psi}\|^2_{\mathcal{H}^k_r(D)}\, \mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \;+\; C_2 N_\theta^{-l}\left(\int_\Omega \|\hat{\psi}\|^2_{\mathcal{H}^l_\theta(D)}\, \mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}}, \end{aligned}$$

where we used (3.89) with $k = 0,\ p = 2$ to obtain the last line.

We treat $\|\underset{\sim}{\nabla}_x\eta\|$ similarly:

$$\begin{aligned}
\|\underset{\sim}{\nabla}_x\eta\| &\leq \|\underset{\sim}{\nabla}_x\hat{\psi} - \underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi}\| + \|\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi} - \Pi_q\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi}\| \\
&\leq Ch^s\left(\int_D |\hat{\psi}|^2_{\mathrm{H}^{s+1}(\Omega)}\,\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}} \\
&\quad + C_1N_r^{-k}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^k_r(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \ + \ C_2N_\theta^{-l}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^l_\theta(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \\
&\leq Ch^s\left(\int_D |\hat{\psi}|^2_{\mathrm{H}^{s+1}(\Omega)}\,\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}} \\
&\quad + C_1N_r^{-k}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\hat{\psi}\|^2_{\mathcal{H}^k_r(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \ + \ C_2N_\theta^{-l}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\hat{\psi}\|^2_{\mathcal{H}^l_\theta(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}}.
\end{aligned}$$

Next, we have

$$\begin{aligned}
\|\underset{\sim}{\nabla}_M\eta\| &\leq \|\underset{\sim}{\nabla}_M\hat{\psi} - \mathcal{I}_x\underset{\sim}{\nabla}_M\hat{\psi}\| + \|\underset{\sim}{\nabla}_M\mathcal{I}_x\hat{\psi} - \underset{\sim}{\nabla}_M\Pi_q\mathcal{I}_x\hat{\psi}\| \\
&\leq Ch^s\left(\int_D |\underset{\sim}{\nabla}_M\hat{\psi}|^2_{\mathrm{H}^s(\Omega)}\,\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}} \\
&\quad + C_1N_r^{-k}\left(\int_\Omega \|\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^{k+1}_r(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \ + \ C_2N_\theta^{-l}\left(\int_\Omega \|\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^{l+1}_\theta(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \\
&\leq Ch^s\left(\int_D |\underset{\sim}{\nabla}_M\hat{\psi}|^2_{\mathrm{H}^s(\Omega)}\,\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}} \\
&\quad + C_1N_r^{-k}\left(\int_\Omega \|\hat{\psi}\|^2_{\mathcal{H}^{k+1}_r(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \ + \ C_2N_\theta^{-l}\left(\int_\Omega \|\hat{\psi}\|^2_{\mathcal{H}^{l+1}_\theta(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}}.
\end{aligned}$$

Finally, we derive a bound for the cross-term $\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\eta\|$ as follows:

$$\begin{aligned}
\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\eta\| &\leq \|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\hat{\psi} - \underset{\sim}{\nabla}_x\mathcal{I}_x\underset{\sim}{\nabla}_M\hat{\psi}\| + \|\underset{\sim}{\nabla}_M\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi} - \underset{\sim}{\nabla}_M\Pi_q\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi}\| \\
&\leq Ch^s\left(\int_D |\underset{\sim}{\nabla}_M\hat{\psi}|^2_{\mathrm{H}^{s+1}(\Omega)}\,\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}} \\
&\quad + C_1N_r^{-k}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^{k+1}_r(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \ + \ C_2N_\theta^{-l}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\mathcal{I}_x\hat{\psi}\|^2_{\mathcal{H}^{l+1}_\theta(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \\
&\leq Ch^s\left(\int_D |\underset{\sim}{\nabla}_M\hat{\psi}|^2_{\mathrm{H}^{s+1}(\Omega)}\,\mathrm{d}\underset{\sim}{q}\right)^{\frac{1}{2}} \\
&\quad + C_1N_r^{-k}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\hat{\psi}\|^2_{\mathcal{H}^{k+1}_r(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}} \ + \ C_2N_\theta^{-l}\left(\int_\Omega \|\underset{\sim}{\nabla}_x\hat{\psi}\|^2_{\mathcal{H}^{l+1}_\theta(D)}\,\mathrm{d}\underset{\sim}{x}\right)^{\frac{1}{2}}.
\end{aligned}$$

Therefore, we have the following optimal order bounds for the terms on the right-hand side of (3.88):

$$\|\eta^0\| \leq Ch^s\|\hat{\psi}_0\|_{\mathrm{H}^s(\Omega;\mathrm{L}^2(D))} + C_1N_r^{-k}\|\hat{\psi}_0\|_{\mathrm{L}^2(\Omega;\mathcal{H}^k_r(D))} \ + \ C_2N_\theta^{-l}\|\hat{\psi}_0\|_{\mathrm{L}^2(\Omega;\mathcal{H}^l_\theta(D))},$$

$$\begin{aligned}\|\eta\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} &\leq Ch^s \left\|\hat{\psi}\right\|_{\ell^2(0,t^n;\mathrm{H}^s(\Omega;\mathrm{L}^2(D)))} + C_1 N_r^{-k} \left\|\hat{\psi}\right\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D)))} \\ &\quad + C_2 N_\theta^{-l} \left\|\hat{\psi}\right\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D)))},\end{aligned}$$

$$\begin{aligned}\left\|\frac{\partial\eta}{\partial t}\right\|_{\mathrm{L}^2(0,t^n;\mathrm{L}^2(\Omega\times D))} &\leq Ch^s \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,t^n;\mathrm{H}^s(\Omega;\mathrm{L}^2(D)))} + C_1 N_r^{-k} \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,t^n;\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D)))} \\ &\quad + C_2 N_\theta^{-l} \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,t^n;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D)))},\end{aligned}$$

$$\begin{aligned}\|\underset{\sim}{\nabla}_x\eta\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} &\leq Ch^s\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{H}^{s+1}(\Omega;\mathrm{L}^2(D)))} \\ &\quad + C_1 N_r^{-k}\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{H}^1(\Omega;\mathcal{H}_r^k(D)))} + C_2 N_\theta^{-l}\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^l(D)))},\end{aligned}$$

$$\begin{aligned}\|\underset{\sim}{\nabla}_M\eta\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} &\leq Ch^s\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{H}^s(\Omega;\mathrm{H}_0^1(D;M)))} \\ &\quad + C_1 N_r^{-k}\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega;\mathcal{H}_r^{k+1}(D)))} + C_2 N_\theta^{-l}\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^{l+1}(D)))},\end{aligned}$$

$$\begin{aligned}\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\eta\|_{\ell^2(0,t^n;\mathrm{L}^2(\Omega\times D))} &\leq Ch^s\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{H}^{s+1}(\Omega;\mathrm{H}_0^1(D;M)))} \\ &\quad + C_1 N_r^{-k}\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{H}^1(\Omega;\mathcal{H}_r^{k+1}(D)))} + C_2 N_\theta^{-l}\|\hat{\psi}\|_{\ell^2(0,t^n;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^{l+1}(D)))}.\end{aligned}$$

3.7 Convergence analysis for method I, Part 2

Putting the estimates derived above into (3.88), with appropriate constants C_1, C_2 C_3 and C_4, we obtain:

$$\begin{aligned}&\|\xi\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega\times D))} + \|\underset{\sim}{\nabla}_M\xi\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))} \\ &\quad\leq C_1 h^s\Big(\|\hat{\psi}_0\|_{\mathrm{H}^s(\Omega;\mathrm{L}^2(D))} + \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{H}^s(\Omega;\mathrm{L}^2(D)))} + \left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{H}^s(\Omega;\mathrm{H}_0^1(D;M)))} \\ &\qquad\qquad + \left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{H}^{s+1}(\Omega;\mathrm{L}^2(D)))}\Big) \\ &\qquad + C_2 N_r^{-k}\Big(\|\hat{\psi}_0\|_{\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D))} + \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D)))} + \|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_r^k(D)))} \\ &\qquad\qquad + \|\hat{\psi}\|_{\ell^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^{k+1}(D)))}\Big) \\ &\qquad + C_3 N_\theta^{-l}\Big(\|\hat{\psi}_0\|_{\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D))} + \left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D)))} + \|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^l(D)))} \\ &\qquad\qquad + \left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^{l+1}(D)))}\Big) \\ &\qquad + C_4\Delta t\Big(\left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))} + \left\|\hat{\psi}\right\|_{\mathrm{H}^2(0,T;\mathrm{L}^2(\Omega\times D))} + \|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\hat{\psi}\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))} \\ &\qquad\qquad + N_r^{-k}\|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_r^{k+1}(D)))} + N_\theta^{-l}\|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^{l+1}(D)))}\Big).\end{aligned}$$

Hence, by the triangle inequality:

$$
\begin{aligned}
&\|\hat{\psi}-\hat{\psi}_{h,N}\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega\times D))}+\|\underset{\sim}{\nabla}_M(\hat{\psi}-\hat{\psi}_{h,N})\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))}\\
&\quad\le \|\xi\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega\times D))}+\|\underset{\sim}{\nabla}_M\xi\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))}+\|\eta\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega\times D))}+\|\underset{\sim}{\nabla}_M\eta\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))}\\
&\qquad\le C_1h^s\Big(\|\hat{\psi}\|_{\ell^\infty(0,T;\mathrm{H}^s(\Omega;\mathrm{L}^2(D)))}+\left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{H}^s(\Omega;\mathrm{L}^2(D)))}+\left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{H}^s(\Omega;\mathrm{H}_0^1(D;M)))}\\
&\qquad\qquad+\left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{H}^{s+1}(\Omega;\mathrm{L}^2(D)))}\Big)\\
&\qquad+C_2N_r^{-k}\Big(\|\hat{\psi}\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D)))}+\left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D)))}+\|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_r^k(D)))}\\
&\qquad\qquad+\|\hat{\psi}\|_{\ell^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^{k+1}(D)))}\Big)\\
&\qquad+C_3N_\theta^{-l}\Big(\|\hat{\psi}\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D)))}+\left\|\frac{\partial\hat{\psi}}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D)))}+\|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^l(D)))}\\
&\qquad\qquad+\left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^{l+1}(D)))}\Big)\\
&\qquad+C_4\Delta t\Big(\left\|\hat{\psi}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))}+\left\|\hat{\psi}\right\|_{\mathrm{H}^2(0,T;\mathrm{L}^2(\Omega\times D))}+\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\hat{\psi}\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))}\\
&\qquad\qquad+N_r^{-k}\|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_r^{k+1}(D)))}+N_\theta^{-l}\|\hat{\psi}\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^{l+1}(D)))}\Big).
\end{aligned}
\tag{3.93}
$$

Therefore, with $\psi_{h,N}=\sqrt{M}\hat{\psi}_{h,N}$, the estimate analogous to (2.50) for alternating-direction method I is the following:

$$
\begin{aligned}
&\|\psi-\psi_{h,N}\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega;\mathfrak{H}))}+\|\psi-\psi_{h,N}\|_{\ell^2(0,T;\mathrm{L}^2(\Omega;\mathfrak{K}))}\\
&\qquad\le C_1h^s\Big(\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^\infty(0,T;\mathrm{H}^s(\Omega;\mathrm{L}^2(D)))}+\left\|\frac{1}{\sqrt{M}}\frac{\partial\psi}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{H}^s(\Omega;\mathrm{L}^2(D)))}+\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{H}^s(\Omega;\mathrm{H}_0^1(D;M)))}\\
&\qquad\qquad+\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{H}^{s+1}(\Omega;\mathrm{L}^2(D)))}\Big)\\
&\qquad+C_2N_r^{-k}\Big(\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D)))}+\left\|\frac{1}{\sqrt{M}}\frac{\partial\psi}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^k(D)))}+\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_r^k(D)))}\\
&\qquad\qquad+\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_r^{k+1}(D)))}\Big)\\
&\qquad+C_3N_\theta^{-l}\Big(\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D)))}+\left\|\frac{1}{\sqrt{M}}\frac{\partial\psi}{\partial t}\right\|_{\mathrm{L}^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^l(D)))}+\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^l(D)))}\\
&\qquad\qquad+\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega;\mathcal{H}_\theta^{l+1}(D)))}\Big)\\
&\qquad+C_4\Delta t\Big(\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))}+\left\|\frac{\psi}{\sqrt{M}}\right\|_{\mathrm{H}^2(0,T;\mathrm{L}^2(\Omega\times D))}+\left\|\underset{\sim}{\nabla}_x\underset{\sim}{\nabla}_M\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))}\\
&\qquad\qquad+N_r^{-k}\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_r^{k+1}(D)))}+N_\theta^{-l}\left\|\frac{\psi}{\sqrt{M}}\right\|_{\ell^2(0,T;\mathrm{H}^1(\Omega;\mathcal{H}_\theta^{l+1}(D)))}\Big),
\end{aligned}
\tag{3.94}
$$

for $s, k, l \geq 1$, provided that ψ is such that the right-hand side is finite. Note than an obvious difference between (3.94) and (2.50) is that in (3.94) we require

$$\left\| \underset{\sim}{\nabla}_x \underset{\sim}{\nabla}_M \frac{\psi}{\sqrt{M}} \right\|_{\ell^2(0,T;\mathrm{L}^2(\Omega\times D))} < \infty.$$

This regularity condition is necessitated by the presence of the cross term,

$$\left(\underset{\sim}{\nabla}_M \left(\underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \hat{\psi}_{h,N}^{n+1} \right), \underset{\sim}{\nabla}_M \zeta \right),$$

in (3.48).

Remark 3.12 *Looking at* (3.94)*, it could be argued that there is a mismatch between the convergence rates of the finite element method in* Ω *and the spectral method in* D*, in the sense that the spectral method will generally be far more accurate. This is a reasonable point, but we believe that in practice the numerical method analysed here is appropriate. First of all, while in general a finite element scheme will have a low-order convergence rate, its flexibility is invaluable when it comes to meshing physical space domains that may be complicated. Moreoever, we do not have a diffusion operator in the* $\underset{\sim}{x}$*-direction, so it is not obvious that* ψ *will be highly smooth in* Ω*.*

Nevertheless, it is certainly also reasonable to use a higher-order method for solving the transport equation in physical space, for example, Chauvière & Lozinski used a spectral element method for this purpose in [23, 24]. Note that the analysis in this chapter would carry over essentially unchanged if we replaced the finite element discretisation of (3.41) *by a higher-order method.*

On the other hand, the $\underset{\sim}{q}$*-direction is much better suited to the use of a high-order method since* D *is always a ball in* $\mathbb{R}^d$*, and, as seen in Section 2.6, at least for the FENE potential, the solution profiles in* D *are generally very smooth. Note that in practice the spectral convergence of the* $\underset{\sim}{q}$*-direction numerical method means that the discrete space* $\mathcal{P}_N(D)$ *need only have a rather low dimensionality. This is highly advantageous because (a) each* $\underset{\sim}{q}$*-direction solve requires relatively modest computational resources and (b) a reduction in the dimensionality of* $\mathcal{P}_N(D)$ *reduces the number of* $\underset{\sim}{x}$*-direction solves that need to be performed each time-step (*cf. (3.41)*).*

Remark 3.13 *In the preceding argument, we made use of the (pointwise) divergence-free assumption,* (3.3)*. This assumption was made to simplify the argument, but it is not essential, i.e. it follows from* (3.4) *that* $\underset{\sim}{\nabla}_x \cdot \underset{\sim}{u} \in \mathrm{L}^\infty(\Omega)$*, hence if we allowed* $\underset{\sim}{\nabla}_x \cdot \underset{\sim}{u}$ *to be non-zero the preceding convergence argument could be modified to use the norm* $\|\underset{\sim}{\nabla}_x \cdot \underset{\sim}{u}\|_{\mathrm{L}^\infty(\Omega)}$ *instead of* (3.3)*.*

Now, following the discussion in Section 2.6.1, we consider the convergence of $\underset{\approx}{\tau}$. In order to coincide with Section 2.6.1, here we consider only the FENE spring force and the case in which $d = 2$.

Using Parseval's identity from Chapter 2, we write the weak solution $\hat{\psi}(\underset{\sim}{x}, \underset{\sim}{q}, t) = \tilde{\psi}(\underset{\sim}{x}, r, \theta, t)$ as follows:

$$\tilde{\psi}(\underset{\sim}{x}, r, \theta, t) = \tilde{\psi}_1(\underset{\sim}{x}, r, t) + r \sum_{l=1}^{\infty} \left(\tilde{A}_l(\underset{\sim}{x}, r, t) \cos(2l\theta) + \tilde{B}_l(\underset{\sim}{x}, r, t) \sin(2l\theta) \right), \tag{3.95}$$

and supposing we use basis $\mathcal{A}$ in the $\underset{\sim}{q}$-direction, we define the numerical solution as:

$$\tilde{\psi}_{h,N}(\underset{\sim}{x}, r, \theta) = (1-r) \sum_{k=0}^{N_r-1} \tilde{\Psi}_{0,k}(\underset{\sim}{x}) P_k(r) + r(1-r) \sum_{i=0}^{1} \sum_{l=1}^{N_\theta} \sum_{k=0}^{N_r-1} \tilde{\Psi}^i_{l,k}(\underset{\sim}{x}) P_k(r) \Phi_{il}(\theta),$$

where $\tilde{\Psi}_{0,k}, \tilde{\Psi}^i_{l,k} \in V_h$ are line functions as in (3.30).

Therefore, proceeding as in Section 2.6, we obtain

$$\begin{aligned}
&\|\tau_{11}(\hat{\psi}(t^n)) - \tau_{11}(\hat{\psi}^n_{h,N})\|^2_{\mathrm{L}^2(\Omega)} \\
&\leq C_* \int_\Omega \left\| \tilde{\psi}_1(\underset{\sim}{x}, r, t^n) - (1-r) \sum_{k=0}^{N_r-1} \tilde{\Psi}^n_{0,k}(\underset{\sim}{x}) P_k(r) \right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} \mathrm{d}\underset{\sim}{x} \\
&\quad + \frac{C_*}{4} \int_\Omega \left\| r\tilde{A}_1(\underset{\sim}{x}, r, t^n) - r(1-r) \sum_{k=0}^{N_r-1} \tilde{\Psi}^{0,n}_{1,k}(\underset{\sim}{x}) P_k(r) \right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} \mathrm{d}\underset{\sim}{x},
\end{aligned} \tag{3.96}$$

where C_* is defined in (2.68).

Also, the analogue of (2.69) here is:

$$\begin{aligned}
&\|\hat{\psi}(\cdot, \cdot, t^n) - \hat{\psi}^n_N(\cdot, \cdot)\|^2_{\mathrm{L}^2(\Omega \times D)} \\
&= 2\pi b \int_\Omega \left\| \tilde{\psi}_1(\underset{\sim}{x}, r, t^n) - (1-r) \sum_{k=0}^{N_r-1} \tilde{\Psi}^n_{0,k}(\underset{\sim}{x}) P_k(r) \right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} \mathrm{d}\underset{\sim}{x} \\
&\quad + \pi b \sum_{l=1}^{N_\theta} \int_\Omega \left\| r\tilde{A}_l(\underset{\sim}{x}, r, t^n) - r(1-r) \sum_{k=0}^{N_r-1} \tilde{\Psi}^{0,n}_{l,k}(\underset{\sim}{x}) P_k(r) \right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} \mathrm{d}\underset{\sim}{x} \\
&\quad + \pi b \sum_{l=1}^{N_\theta} \int_\Omega \left\| r\tilde{B}_l(\underset{\sim}{x}, r, t^n) - r(1-r) \sum_{k=0}^{N_r-1} \tilde{\Psi}^{1,n}_{l,k}(\underset{\sim}{x}) P_k(r) \right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} \mathrm{d}\underset{\sim}{x} \\
&\quad + \pi b \sum_{l=N_\theta+1}^{\infty} \int_\Omega \left(\left\| r\tilde{A}_l(\underset{\sim}{x}, r, t^n) \right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} + \left\| r\tilde{B}_l(\underset{\sim}{x}, r, t^n) \right\|^2_{\mathrm{L}^2_{\tilde{w}}(0,1)} \right) \mathrm{d}\underset{\sim}{x},
\end{aligned} \tag{3.97}$$

and hence, once again, the τ_{11} error only contains two terms from the infinite series in (3.97), and as in (2.70), we have

$$\|\tau_{11}(\hat{\psi}) - \tau_{11}(\hat{\psi}_{h,N})\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega))} \leq \sqrt{\frac{C_*}{2\pi b}}\, \|\hat{\psi} - \hat{\psi}_{h,N}\|_{\ell^\infty(0,T;\mathrm{L}^2(\Omega\times D))}. \tag{3.98}$$

Note that since the line functions $\tilde{\Psi}^n_{0,k}$ and $\tilde{\Psi}^{0,n}_{1,k}$ in (3.96) are computed by solving (3.41) using the $\underset{\sim}{x}$-direction finite element method, we expect an $\mathcal{O}(h^s)$ error to dominate the spatial convergence rate of $\underset{\approx}{\tau}$, just as in (3.94). However, by comparing (3.96) and (3.97), we can see that only relatively few terms in the $\underset{\sim}{q}$-direction spectral expansion of $\hat{\psi}_{h,N}$ contribute to the τ_{11} error. Hence, this suggests that the accuracy of $\underset{\approx}{\tau}$ will be less sensitive to the resolution of the $\underset{\sim}{q}$-direction spectral method than the accuracy of $\hat{\psi}_{h,N}$. In Section 3.9 we show that this is indeed the case in practice.

3.8 Implementation of methods I and II

In this section we consider the implementation of the $\underset{\sim}{q}$-direction spectral method and the $\underset{\sim}{x}$-direction finite element method in Sections 3.8.1 and 3.8.2, respectively, and then in Section 3.8.3 we discuss the $\underset{\sim}{x}$-direction quadrature rule used to integrate these two methods into a single alternating-direction algorithm. Finally, we consider the parallel implementation of the alternating-direction methods in Section 3.8.4.

3.8.1 The $\underset{\sim}{q}$-direction stage

We note first of all that from an implementational point of view method I and method II are almost identical; the only difference between the two methods is that method I uses a semi-implicit temporal discretisation whereas method II uses the backward Euler scheme.

Therefore, letting $\hat{\underset{\sim}{\psi}}^{n*}(\underset{\sim}{x}_m) \in \mathbb{R}^{N_D}$ be the vector with k^{th} entry equal to $\hat{\psi}^{n*}_k(\underset{\sim}{x}_m)$ and defining $\hat{\underset{\sim}{\psi}}^{n}(\underset{\sim}{x}_m)$ analogously, the set of $\underset{\sim}{q}$-direction linear systems to be solved at time-level n for method I is:

$$\left(M_q + \frac{\Delta t}{2\mathrm{Wi}} S_q\right) \hat{\underset{\sim}{\psi}}^{n*}(\underset{\sim}{x}_m) = \left(M_q + \Delta t\, C^m_q\right) \hat{\underset{\sim}{\psi}}^{n}(\underset{\sim}{x}_m), \tag{3.99}$$

for $m = 1, \ldots, Q_\Omega$, whereas for method II we solve:

$$\left(M_q + \frac{\Delta t}{2\mathrm{Wi}} S_q - \Delta t\, C^m_q\right) \hat{\underset{\sim}{\psi}}^{n*}(\underset{\sim}{x}_m) = M_q \hat{\underset{\sim}{\psi}}^{n}(\underset{\sim}{x}_m), \tag{3.100}$$

for $m = 1, \ldots, Q_\Omega$. The matrices M_q, S_q and C^m_q in (3.99) and (3.100) are as defined in (2.54), where $\underset{\approx}{\kappa}$ in C^m_q is sampled at $\underset{\sim}{x}_m$. These matrices depend on the choice of

basis of $\mathcal{P}_N(D)$; refer to Section 2.6 for a discussion of the construction of bases $\mathcal{A}$ and $\mathcal{B}$ for the $d = 2$ case, and basis $\mathcal{C}$ in the case of $d = 3$.

It is clear that for both method I and method II, we must solve an $N_D \times N_D$ linear system Q_Ω times per time-step in the $\underset{\sim}{q}$-direction. Q_Ω can be very large in practice. For example, in Section 3.9 we consider some computations for which Q_Ω is on the order of 10^4. The use of parallel computation can be very helpful in this situation because the $\underset{\sim}{q}$-direction linear solves are independent and therefore it is straightforward to perform them in parallel (we discuss this in detail in Section 3.8.4).

It is also interesting to note that method I requires significantly less computational effort in each time-step than method II because the matrix on the left-hand side in (3.99) is constant for all m and therefore we need only perform one LU-factorisation per time-step with method I, whereas the linear system in (3.100) must be reassembled and solved afresh at each quadrature point $\underset{\sim}{x}_m$ since in general $\underset{\approx}{\kappa}(\underset{\sim}{x}_m)$ varies from one quadrature point to the next. On the other hand, the numerical experiments in Section 2.6.2 indicate that the backward Euler temporal discretisation of the $\underset{\sim}{q}$-direction equation is more stable, and it allows one to take larger time-steps, especially for larger values of Wi or $\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(\Omega)}$. Hence, there is a familiar trade-off in efficiency: each time-step is faster with method I, but we can take larger time-steps with method II. Therefore the optimal choice of numerical method depends on the problem at hand.

Remark 3.14 *The alternating direction method used by Chauvière & Lozinski in [24] is similar to method II in that it treats the $\underset{\approx}{\kappa}$ convection term implicitly in time. In the follow-up papers [23, 60] the same authors developed a fast solver approach in which the computational work required for each $\underset{\sim}{q}$-direction solve was significantly reduced. However, their fast solver was based on an assumption that $\underset{\approx}{\kappa}$ arises from a two-dimensional velocity field (*i.e. *that $\Omega \subset \mathbb{R}^2$) whereas in this thesis we are interested in developing numerical methods that are suitable for $\Omega \subset \mathbb{R}^3$.*

The $\underset{\sim}{q}$-direction solvers for methods I and II were implemented in the `C++` programming language and `PETSc` [5] was used to perform the linear algebra operations. `PETSc` was a natural choice in this context because it is designed for use on parallel architectures, which is a feature we made extensive use of.

3.8.2 The $\underset{\sim}{x}$-direction stage

In the $\underset{\sim}{x}$-direction, methods I and II are identical: For each line function, $\hat{\psi}_k^{n*}$, $k = 1, \ldots, N_D$, we solve the transport equation (3.43). This involves solving an $N_\Omega \times N_\Omega$

linear system N_D times, although the system matrix $M_x + \Delta t\, T_x$ only needs to be assembled once per time-step.

In our implementation, we used an $\mathrm{H}^1(\Omega)$-conforming finite element method with quadratic shape functions to perform the $\underset{\sim}{x}$-direction computations, and we used GMRES to solve the resulting linear systems. Hence, assuming sufficient regularity for $\psi/\sqrt{M}$, we can set $s = 2$ in (3.94), which yields $\mathcal{O}(h^2)$ terms in the error estimate. Note that in order to strengthen the norm in which the $\underset{\sim}{x}$-direction solver is stable, Chauviére & Lozinski used an SUPG scheme to discretise the transport equation in [24]. It would be straightforward to integrate such a scheme into our alternating-direction framework, but since the analysis in the preceding sections was performed for a standard Galerkin formulation in the $\underset{\sim}{x}$-direction, for consistency, we prefer to use the Galerkin method in practice also. Moreover, our numerical results in Section 3.9 and in Chapter 4 demonstrate that the standard Galerkin formulation performs well in practice.

This method was implemented using the free, open source `C++` finite element library `libMesh` [47]. Note also that the $\underset{\sim}{x}$-direction computations are independent from one another, and hence parallel computation can again be used effectively.

3.8.3 The $\underset{\sim}{x}$-direction quadrature rule

We have a great deal of freedom in the choice of the $\underset{\sim}{x}$-direction quadrature rule. From the analytical point of view, it is preferable to choose a quadrature rule that satisfies QH1, since then, at least with method I, we have access to the equivalent one-step formulation (3.48), which was the foundation of the convergence analysis of Section 3.7. However, Lemma (3.67) also shows that only QH2 is required for the stability of method I and method II. In practice, the overall computation time depends very strongly on Q_Ω and hence it is often desirable to only satisfy QH2 in order to keep Q_Ω as low as possible.

We now discuss some quadrature rules with which we can satisfy either QH2 or both QH1 and QH2 (recall that QH1 is a stronger hypothesis than QH2). Of course, the quadrature rules depend on the element type and the dimension; we will consider triangles and quadrilaterals in two dimensions and tetrahedra and hexahedra in three dimensions. We discuss element-based quadrature rules only. By combining the quadrature rule on each element of $\mathcal{T}_h$ we obtain a global formula as in (3.26).

We assume that each element $K \in \mathcal{T}_h$ is an affine mapping of some canonical element $\hat{K}$. Hence we only need to consider quadrature rules on $\hat{K}$.

Tensor product elements: In this case, we consider $\overline{\hat{K}}$ to be either the square $[-1,1]^2$ or the cube $[-1,1]^3$. Let $\{\hat{x}_1,\ldots,\hat{x}_n\}$ and $\{\hat{w}_1,\ldots,\hat{w}_n\}$ define the points and weights of a Gaussian quadrature rule, such that $\hat{x}_i \in (-1,1)$ and $\hat{w}_i > 0$ for each i (*e.g.* see Chapter 10 of [77]). It is well known that a Gaussian quadrature rule with n points in one dimension is optimal in the sense that it integrates polynomials of degree $2n-1$ on $\hat{x} \in [-1,1]$ exactly.

For tensor product finite elements defined on the reference square $[-1,1]^2$, the natural choice of quadrature rule is a tensor product Gaussian rule. For example, following [85], we use the quadrature points:

$$\{(\hat{x}_1,\hat{x}_1),(\hat{x}_1,\hat{x}_2),\ldots,(\hat{x}_1,\hat{x}_n),(\hat{x}_2,\hat{x}_1),\ldots,(\hat{x}_n,\hat{x}_n)\},$$

and corresponding weights:

$$\{\hat{w}_1\,\hat{w}_1\,,\ \hat{w}_1\,\hat{w}_2,\ldots,\hat{w}_1\,\hat{w}_n\,,\ \hat{w}_2\,\hat{w}_1,\ldots,\hat{w}_n\,\hat{w}_n\}.$$

This quadrature rule involves $Q_{\hat{K}} = n^2$ points and weights and exactly integrates polynomials on $[-1,1]^2$ of degree $2n-1$ in each direction. A three dimensional tensor product Gauss quadrature rule on $[-1,1]^3$ can be defined analogously.

It is clear from the discussion above that we can construct tensor product Gauss quadrature rules to exactly integrate polynomials of arbitrarily high degree on $[-1,1]^2$ or $[-1,1]^3$. We now consider how many quadrature points we require to satisfy QH1 or QH2 on tensor product elements in two and three dimensions.

In the computations considered in Section 3.9 and in Chapter 4, we use tensor product quadratic shape functions on each element $K \in \mathcal{T}_h$ for $\hat{\psi}_{h,N}$ and for $\underset{\sim}{u}$. Hence the components of $\underset{\approx}{\kappa} = \underset{\sim}{\nabla}_x \underset{\sim}{u}$ can also be quadratic in each direction. Therefore, in order to satisfy QH1, we need to be able to exactly integrate polynomials of degree six, and for QH2 we need to integrate polynomials of degree four exactly. Let p denote the highest degree polynomial that can be exactly integrated by a quadrature rule. We use the following tensor product quadrature rules on the reference square and cube:

- QH1, $p = 7$: $Q_{\hat{K}} = 16$ on $\overline{\hat{K}} = [-1,1]^2$, and $Q_{\hat{K}} = 64$ on $\overline{\hat{K}} = [-1,1]^3$.
- QH2, $p = 5$: $Q_{\hat{K}} = 9$ on $\overline{\hat{K}} = [-1,1]^2$, and $Q_{\hat{K}} = 27$ on $\overline{\hat{K}} = [-1,1]^3$.

These quadrature rules are implemented in the `libMesh` software package.

Simplices: In this case we assume that $\hat{K}$ is either a triangle in two dimensions or a tetrahedron in three dimensions. We again consider quadratic shape functions

for $\underset{\sim}{y}$ and $\hat{\psi}_{h,N}$, but since we are no longer using tensor product finite elements, the components of $\underset{\approx}{\kappa} = \underset{\sim}{\nabla}_x \underset{\sim}{y}$ are only linear functions in this case, so that in order to satisfy QH1 we need to exactly integrate fifth degree polynomials. To satisfy QH2, we need to exactly integrate degree four polynomials, as in the tensor product case.

In our computations, we used the following quadrature rules, which are implemented in the `libMesh` software package on triangles and tetrahedra:

- QH1 on triangles, $p = 5$: $Q_{\hat{K}} = 7$ [81].
- QH2 on triangles, $p = 4$: $Q_{\hat{K}} = 6$ [62].
- QH1 & QH2 on tetrahedra, $p = 5$: $Q_{\hat{K}} = 14$ [81].

Note that there is a fourth order 11 point quadrature rule on tetrahedra from [44] that is implemented in `libMesh` also, but it contains a negative weight and therefore we cannot use it for our alternating-direction method since we need the quadrature rule to define an inner product, *cf.* (3.36). Therefore we use the same $p = 5$ rule on tetrahedra for both QH1 and QH2.

3.8.4 Parallel implementation of the alternating-direction method

It is clear that the computational effort required to solve the high-dimensional Fokker–Planck equation can be very large, particularly in the case $d = 3$. Parallel computation is a key ingredient in the alternating-direction framework developed in this thesis, since it makes many problems tractable that would otherwise be well beyond our reach. As indicated above, methods I and II are very well suited to implentation on a parallel architecture; indeed these algorithms are "embarassingly parallel" in the sense that they involve performing a large number of independent solves in each time-step.

More specifically, suppose we use N_{proc} processors ($N_{\text{proc}} \geq 1$) to solve a problem (using either method I or II) with parameters N_D, N_Ω denoting the number of basis functions in the $\underset{\sim}{q}$-direction and $\underset{\sim}{x}$-direction, respectively, and Q_Ω defining the number of quadrature points in Ω, as in (3.26). At time-level n, we store a dense matrix $D^n \in \mathbb{R}^{Q_\Omega \times N_D}$, where $(D^n)_{ij} = \hat{\psi}_j^n(\underset{\sim}{x}_i)$, and $\hat{\psi}_j^n \in V_h$ is a line function as in (3.30). The entries of D^n uniquely determine $\hat{\psi}_{h,N}^n \in V_h \otimes \mathcal{P}_N(D)$. In practice D^n can be a very large matrix, so we partition it among the processors so that each processor stores a subset of the rows (for $\underset{\sim}{q}$-direction solves) or columns (for $\underset{\sim}{x}$-direction solves) of D^n. We would like these submatrices to be equally sized to obtain ideal load balancing between processors, but depending on Q_Ω, N_D and N_{proc}, this is often not possible. However, to simplify the discussion here, we will assume for the remainder of this section that

N_{proc} is a common divisor of Q_Ω and N_D and hence that the submatrices are equally sized.

Now, let us consider the $\underset{\sim}{q}$-direction computations at time-level n (we do not distinguish between methods I and II here because, from the point of view of the current discussion, they are identical). We distribute D^n so that each processor stores Q_Ω/N_{proc} rows of the matrix. Then, simultaneously, each processor solves the Q_Ω/N_{proc} $\underset{\sim}{q}$-direction problems corresponding to its rows in D^n and updates the data in the matrix. In this manner, D^n is updated to D^{n*} where $(D^{n*})_{ij} = \hat{\psi}_j^{n*}(\underset{\sim}{x}_i)$.

Next, we perform the $\underset{\sim}{x}$-direction computations. First of all, however, we need to redistribute D^{n*} so that each processor stores N_D/N_{proc} columns of the matrix.[3] This involves a global communication operation between all of the processors, which can be time consuming. The time required to perform this parallel communication step depends on the problem size and the number of processors being used. We discuss this issue with regard to some practical computations in Section 3.9, where we show that by selecting N_{proc} appropriately it is generally possible to ensure that the matrix redistribution steps take only a small proportion of the overall computation time.

So, once this matrix redistribution is complete, the $\underset{\sim}{x}$-direction computations on each processor proceed in the same way as in the $\underset{\sim}{q}$-direction. That is, each processor works sequentially through its N_D/N_{proc} columns, first solving (3.43), and then sampling the resulting line function $\hat{\psi}_k^{n+1}$ at $\underset{\sim}{x}_m$ for $m = 1, \ldots, Q_\Omega$ and writing these values back into the matrix. This yields the updated matrix D^{n+1} on completion of all of the $\underset{\sim}{x}$-direction solves.

This process is performed for each time-step, $n = 1, \ldots, N_T$. Note that for computations with the Navier–Stokes–Fokker–Planck system we will need to compute the extra-stress tensor $\underset{\approx}{\tau}$ also. This can be easily included into the framework described above. Suppose we have just finished the $\underset{\sim}{x}$-direction solves so that D^{n+1} has been computed and is stored column-wise so that each processor holds N_D/N_{proc} columns of the matrix. Then to begin the next time-step, we redistribute D^{n+1} again so that each processor holds Q_Ω/N_{proc} rows. Once the redistribution is complete and before we begin the $\underset{\sim}{q}$-direction solves, for each $m = 1, \ldots, Q_\Omega$ we compute and store the values $\underset{\approx}{\tau}^{n+1}(\underset{\sim}{x}_m) \in \mathbb{R}^{d\times d}$ using (1.45) on the $\underset{\sim}{q}$-direction cross-section $\hat{\psi}_{h,N}^{n+1}(\underset{\sim}{x}_m, \cdot) \in \mathcal{P}_N(D)$; this is again done row by row, and hence each processor only performs Q_Ω/N_{proc} computations with Kramers expression. Using (3.37), we can reconstruct $\mathcal{R}\{\underset{\approx}{\tau}^{n+1}(\underset{\sim}{x}_m)\} \in (V_h)^{d\times d}$, which can be used in the right-hand side of (1.42).

[3] In our implementation, we performed this redistribution using `PETSc`'s transpose operation for parallel dense matrices.

3.9 Numerical results

In this section, we present some numerical results for the alternating-direction approach considered in this chapter applied to a model problem for the FENE Fokker–Planck equation in the $d = 2$ case. We take $\underset{\sim}{u}$ to be the solution of the steady incompressible Navier–Stokes equations with $\mathrm{Re} = 1$, and with forcing term $f(x, y) = (5\sin(2\pi y), -5\sin(2\pi x))$, in the domain $\Omega = (0, 1)^2$. In this case, $\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(\Omega)} \approx 2$. We imposed the Dirichlet boundary condition $\underset{\sim}{u} = 0$ on $\partial\Omega$, which ensures that (3.7) is satisfied. The streamlines of $\underset{\sim}{u}$ are shown in Figure 3.1, and we take $\underset{\sim}{u}$ to be constant in time throughout $t \in (0, T]$. This velocity field was obtained by solving the Navier–Stokes equations using the Taylor–Hood finite element scheme with quadratic shape functions for $\underset{\sim}{u}$ and linear shape functions for the pressure (this numerical method is discussed in more detail in Section 4.2), and we use the same finite element mesh, $\mathcal{T}_h$, for the Navier–Stokes equations as for the alternating-direction method, and hence $\underset{\sim}{u} \in V_h$. Note that in general the Taylor-Hood scheme for the Navier–Stokes equations does not yield a (pointwise) divergence-free velocity field, and hence the assumption (3.3) is not satisfied for the computational results in this section. However, as noted in Remark 3.13, the analysis developed in this chapter can be extended essentially unchanged to the case in which $\underset{\sim}{u}$ is not divergence-free.

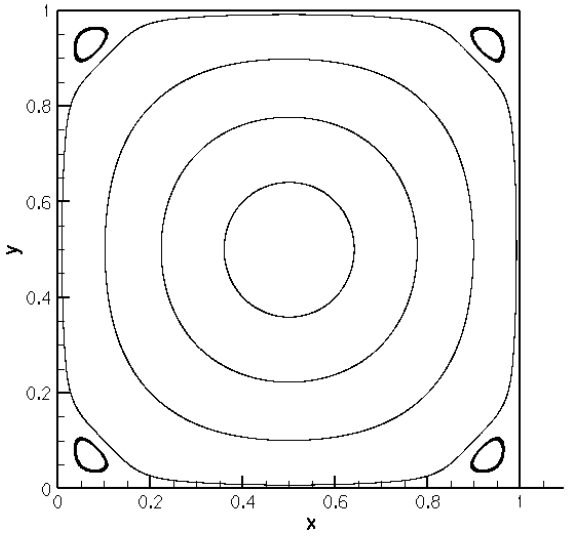
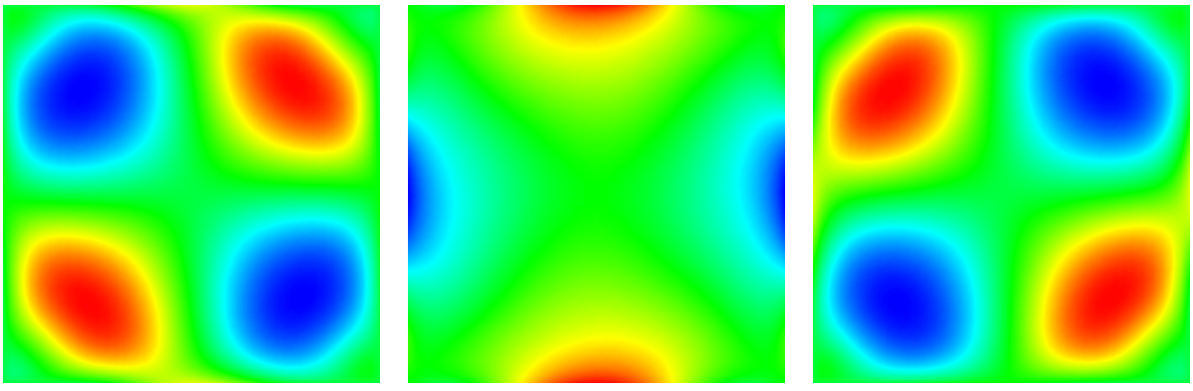


Figure 3.1: Streamlines of the macroscopic velocity field $\underset{\sim}{u}$ driving the enclosed flow model problem. The velocity field is the solution of the steady Navier–Stokes equation with $\mathrm{Re} = 1$ on $\Omega = (0, 1)^2$ with forcing $f(x, y) = (5\sin(2\pi y), -5\sin(2\pi x))$.

We now consider computations using methods I and II for the model problem

described above, with the parameters Wi $= 1$ and $b = 12$. Also, in each of the computations discussed below, we used the initial condition $\hat{\psi}^0_{h,N}(\underset{\sim}{x}, \underset{\sim}{q}) = \sqrt{M(\underset{\sim}{q})}$, where M is the normalised Maxwellian and we ensured that $N_r \geq 6$, since according to Remark 2.17, that guarantees that $\sqrt{M} \in \mathcal{P}_N(D)$ in this case. Our goal is to compare the performance of methods I and II, and to study the convergence of these methods under mesh refinement. All of the computations in this section were performed on the Lonestar parallel computer at the Texas Advanced Computing Center (TACC), `http://www.tacc.utexas.edu`, and we used the parallel implementation of the alternating direction method described in Section 3.8.

We do not know the exact solution of the Fokker–Planck equation with the velocity field in Figure 3.1 and therefore in order to obtain quantitative convergence results we first computed a "reference solution", $\hat{\psi}_{\text{ref}}$, and corresponding polymeric extra-stress tensor, $\underset{\approx}{\tau}_{\text{ref}}$, using method I with basis $\mathcal{A}$ in the $\underset{\sim}{q}$-direction and with a quadrature rule on Ω that satisfied QH1. We obtained this reference solution using a highly refined discrete space, $(V_h \otimes \mathcal{P}_N(D))_{\text{ref}}$, for which $\mathcal{T}_h$ was a 40×40 uniform mesh of square finite elements and $(N_r, N_\theta) = (14, 14)$. In order to satisfy QH1 in this case we required $Q_{\hat{K}} = 16$, and hence $Q_\Omega = 25600$ (*cf.* Section 3.8.3). We took 200 time-steps with $\Delta t = 10^{-3}$ so that $T = 0.2$; this value of Δt is sufficiently small so that temporal discretisation error does not contaminate the spatial convergence results presented below. The components of $\underset{\approx}{\tau}_{\text{ref}}$ at $T = 0.2$ are shown in Figure 3.2.



$\tau_{\text{ref},11}$ $\tau_{\text{ref},12}$ $\tau_{\text{ref},22}$

Figure 3.2: The components of $\underset{\approx}{\tau}_{\text{ref}}$ at $T = 0.2$. Note that we do not show $\tau_{\text{ref},21}$ since it is identical to $\tau_{\text{ref},12}$. In the $\tau_{\text{ref},11}$ and $\tau_{\text{ref},22}$ plots, the values range from 0.882 (blue) to 1.15 (red), and in the $\tau_{\text{ref},12}$ plot we have -0.229 (blue) to 0.229 (red).

In order to obtain convergence data, we then computed $\hat{\psi}_{h,N}$ and the corresponding stress tensor $\underset{\approx}{\tau}$ for several coarser discrete spaces than $(V_h \otimes \mathcal{P}_N(D))_{\text{ref}}$. First of all

we carried out this process using the same numerical method with which we obtained the reference solution, *i.e.* method I with basis $\mathcal{A}$ and a quadrature rule that satisfied QH1. The solution data obtained from these computations are denoted $\hat{\psi}_{\mathrm{I}}$ and $\underset{\approx}{\tau}_{\mathrm{I}}$ below. Then, we also computed a corresponding set of numerical solutions on the same discrete spaces, but using method II with basis $\mathcal{A}$ and a quadrature rule that only satisfied QH2.[4] We denote the solution data in this second case by $\hat{\psi}_{\mathrm{II}}$ and $\underset{\approx}{\tau}_{\mathrm{II}}$.

The numerial results for $\hat{\psi}_{\mathrm{I}}$ and $\underset{\approx}{\tau}_{\mathrm{I}}$ were obtained using a numerical method that satisfies all of the hypotheses required by the convergence estimates in Section 3.7 (except the divergence-free assumption on $\underset{\sim}{u}$, but, as mentioned above, this assumption is not essential; we only used it in order to simplify the analysis in this chapter). Hence, the $\hat{\psi}_{\mathrm{I}}$ and $\underset{\approx}{\tau}_{\mathrm{I}}$ convergence data in the table allow us to compare the theoretical estimates with practical convergence results. Also, the numerical results enable us to compare the convergence behaviour of method I with QH1 to method II with QH2. These two methods are very similar to one another hence we expect to observe the same convergence behaviour in the two cases, but it is important to provide experimental evidence that these two methods converge to the same solution, and at the same rate, in practice because strictly speaking the convergence analysis in this chapter is only valid for method I with QH1.

The convergence estimates (3.94) and (3.98) indicate that if the error due to the $\underset{\sim}{q}$-direction spectral method is negligible compared to the error from the $\underset{\sim}{x}$-direction finite element method, we should obtain $\mathcal{O}(h^2)$ convergence rates for both $\hat{\psi}$ and $\underset{\approx}{\tau}$ as $\mathcal{T}_h$ is refined. Table 3.1 gives the relative errors $\|\hat{\psi}_{\mathrm{I}} - \hat{\psi}_{\mathrm{ref}}\|_{\mathrm{L}^2(\Omega\times D)}/\|\hat{\psi}_{\mathrm{ref}}\|_{\mathrm{L}^2(\Omega\times D)}$ and $\|\hat{\psi}_{\mathrm{II}} - \hat{\psi}_{\mathrm{ref}}\|_{\mathrm{L}^2(\Omega\times D)}/\|\hat{\psi}_{\mathrm{ref}}\|_{\mathrm{L}^2(\Omega\times D)}$ as well as $\|\tau_{\mathrm{I},11} - \tau_{\mathrm{ref},11}\|_{\mathrm{L}^2(\Omega)}/\|\tau_{\mathrm{ref},11}\|_{\mathrm{L}^2(\Omega)}$ and $\|\tau_{\mathrm{II},11} - \tau_{\mathrm{ref},11}\|_{\mathrm{L}^2(\Omega)}/\|\tau_{\mathrm{ref},11}\|_{\mathrm{L}^2(\Omega)}$, at $T = 0.2$, for the discrete spaces that we considered.

In order to gain further insight into the convergence behaviour of the numerical methods, we plotted the data in Table 3.1 in Figures 3.3 and 3.4.

In Figure 3.3, the convergence results for $\hat{\psi}_{\mathrm{I}}$ and $\hat{\psi}_{\mathrm{II}}$ with $(N_r, N_\theta) = (6, 6)$ and $(N_r, N_\theta) = (10, 10)$ are plotted on a log-log scale. We have also included a plot of h^2 to show how the decay of the computed errors compare to the expected asymptotic rate. First of all, it is clear from the figure that the two numerical methods behave very similarly; the lines from $\hat{\psi}_{\mathrm{I}}$ and $\hat{\psi}_{\mathrm{II}}$ are almost indistinguishable. Also, Figure 3.3 shows that we obtain $\mathcal{O}(h^2)$ convergence when $(N_r, N_\theta) = (10, 10)$. However, when $(N_r, N_\theta) = (6, 6)$, the plots plateau, which indicates that the error due to the spectral method dominates the $\mathcal{O}(h^2)$ finite element error when $\mathcal{T}_h$ is a 20×20 mesh.

[4]Recall that we only require $Q_{\hat{K}} = 9$ to satisfy QH2 on square finite elements.

$\mathcal{T}_h$	$(N_r,\ N_\theta)$	$\hat{\psi}_{\mathrm{I}}$ error	$\tau_{\mathrm{I},11}$ error	$\hat{\psi}_{\mathrm{II}}$ error	$\tau_{\mathrm{II},11}$ error
5×5	$(6, 6)$	2.07×10^{-2}	1.63×10^{-2}	2.08×10^{-2}	1.63×10^{-2}
5×5	$(8, 8)$	2.05×10^{-2}	1.63×10^{-2}	2.06×10^{-2}	1.63×10^{-2}
5×5	$(10, 10)$	2.05×10^{-2}	1.63×10^{-2}	2.06×10^{-2}	1.63×10^{-2}
10×10	$(6, 6)$	6.25×10^{-3}	4.22×10^{-3}	6.30×10^{-3}	4.24×10^{-3}
10×10	$(8, 8)$	5.62×10^{-3}	4.22×10^{-3}	5.65×10^{-3}	4.23×10^{-3}
10×10	$(10, 10)$	5.54×10^{-3}	4.22×10^{-3}	5.58×10^{-3}	4.23×10^{-3}
20×20	$(6, 6)$	3.29×10^{-3}	9.95×10^{-4}	3.40×10^{-3}	1.07×10^{-3}
20×20	$(8, 8)$	1.80×10^{-3}	9.90×10^{-4}	1.89×10^{-3}	1.04×10^{-3}
20×20	$(10, 10)$	1.52×10^{-3}	9.90×10^{-4}	1.67×10^{-3}	1.04×10^{-3}

Table 3.1: Convergence of $\hat{\psi}$ and τ_{11} with respect to the reference solution $\hat{\psi}_{\mathrm{ref}}$ and reference polymeric stress tensor $\tau_{\mathrm{ref},11}$ for a series of increasingly refined discrete spaces. The errors are calculated in the L^2 norm at $T = 0.2$, and are normalised by dividing by $\|\hat{\psi}_{\mathrm{ref}}(\cdot, \cdot, T)\|_{\mathrm{L}^2(\Omega \times D)} = 0.31$ and $\|\tau_{\mathrm{ref},11}(\cdot, T)\|_{\mathrm{L}^2(\Omega)} = 1.04$.

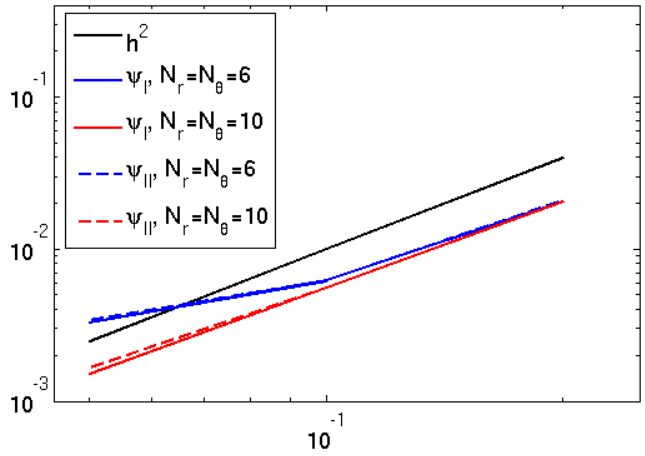


Figure 3.3: Plots of the $\hat{\psi}_{\mathrm{I}}$ and $\hat{\psi}_{\mathrm{II}}$ convergence data in Table 3.1. The black line shows the expected asymptotic decay rate, h^2, and the blue and red lines show the convergence of the two numerical methods when (N_r, N_θ) is fixed at $(6, 6)$ and $(10, 10)$, respectively.

The $\tau_{\mathrm{I},11}$ and $\tau_{\mathrm{II},11}$ convergence data is plotted in Figure 3.4. The data in Table 3.1 is almost identical for $(N_r, N_\theta) = (6, 6), (8, 8)$ and $(10, 10)$, and therefore we only show the $(N_r, N_\theta) = (6, 6)$ data in the figure. The plot shows that we obtained $\mathcal{O}(h^2)$ convergence for both $\tau_{\mathrm{I},11}$ and $\tau_{\mathrm{II},11}$ as $\mathcal{T}_h$ is refined from a 5×5 mesh to 20×20 mesh, when $(N_r, N_\theta) = (6, 6)$. This is markedly different from the convergence behaviour of $\hat{\psi}_{h,N}$, in which the $\underset{\sim}{q}$-direction spectral error for $(N_r, N_\theta) = (6, 6)$ dominated the finite element error on the 20×20 $\underset{\sim}{x}$-direction mesh. Therefore, this indicates that, just as in Section 2.6, the D domain spectral method exhibits superconvergence for $\underset{\approx}{\tau}$ compared

to $\hat{\psi}$. This behaviour is dictated by (3.96), which indicates that only a small fraction of the terms in the expansion of $\hat{\psi}_{h,N}$ in terms of spectral basis functions contribute to the error in $\underset{\approx}{\tau}$. As has been noted earlier, the superconvergence of $\underset{\approx}{\tau}$ is extremely beneficial in the context of micro-macro computations for simulating dilute polymeric fluids because in that setting the error in $\hat{\psi}$ is irrelevant; we are solely interested in the $\underset{\approx}{\tau}$ error.

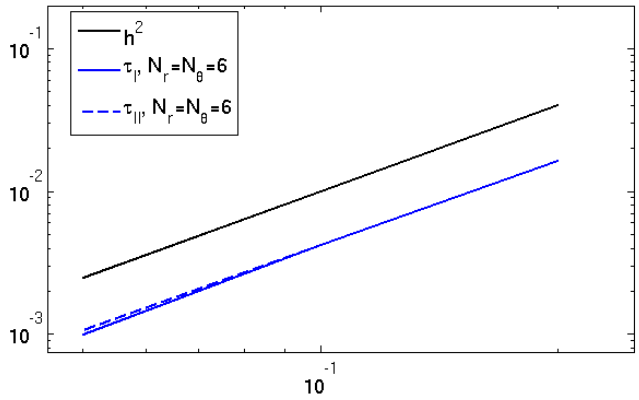


Figure 3.4: Plots of the $\tau_{\mathrm{I},11}$ and $\tau_{\mathrm{II},11}$ convergence data in Table 3.1. The black line shows the expected asymptotic decay rate, h^2, and the solid and dashed blue lines show, respectively, the $\tau_{\mathrm{I},11}$ and $\tau_{\mathrm{II},11}$ data for $(N_r, N_\theta) = (6, 6)$. The data for the other values of (N_r, N_θ) are not plotted since the τ_{11} convergence data in Table 3.1 is virtually unaffected by increasing the number of spectral basis functions.

Recall from the discussion in Section 3.8.1 that we expect method I to require significantly less computational work per time-step in the $\underset{\sim}{q}$-direction than method II. To demonstrate this in practice, we solved the same enclosed flow model problem using both method I and method II. We used a 20×20 uniform mesh $\mathcal{T}_h$ of square finite elements with $Q_\Omega = 3600$ and basis $\mathcal{B}$ with $(N_r, N_\theta) = (15, 15)$ so that $N_D = 465$. With $N_{\mathrm{proc}} = 4$, the total computation time per time-step for method I was 1.75 seconds, whereas for method II it was 3.42 seconds. This difference is due to the fact that method II took 2.37 seconds per time-step to perform the $\underset{\sim}{q}$-direction computations, whereas method I only took 0.70 seconds per time-step in the $\underset{\sim}{q}$-direction.

Nevertheless, for problems of physical interest, method II is often the preferred alternating-direction method. This is because the fully implicit temporal discretisation used by method II is more stable than the semi-implicit scheme in method I, especially for larger flow rates and Weissenberg numbers (*cf.* Section 2.6.2). Hence method I can require much smaller time-step sizes than method II, and this can often outweigh the reduced computational complexity per time-step of method I. Also, for large-scale

problems we generally prefer to satisfy only QH2 rather than QH1 since with QH2 we can obtain a smaller value of Q_Ω, which in turn reduces the computational work required in each time-step of the alternating-direction method.

We now move on to consider the scaling of the computation time as we increase the number of processors in the parallel implementation of the alternating-direction method. The enclosed-flow problem considered above provides a convenient test case with which we can quantify the parallel speedup for the alternating-direction method. We studied this speedup by, first of all, solving the enclosed flow problem on one node of the Lonestar parallel computer (each node contains 4 processors) to get the base computation time per time-step, which we denote $T(1)$. We then repeated the same computation, but using more computational nodes of the parallel computer and we recorded the computation time, $T(N)$, in each case, where N denotes the number of computational nodes that were used. We refer to the ratio $T(1)/T(N)$ as the *parallel speedup*.

The parameters that have the most significant effect on the computation time of the parallel alternating-direction scheme are N_D and Q_Ω, since these determine the number of $\underset{\sim}{x}$- and $\underset{\sim}{q}$-direction solves that need to be performed each time-step. Note that there are only two steps in the alternating-direction algorithm for which the computation time does not scale down proportionally to the number of processors being used: the matrix assembly for (3.43), which must be performed exactly once per time-step irrespective of N_{proc}, and also the dense matrix redistribution that precedes direction changes in the alternating-direction method. However, if the $\underset{\sim}{x}$- and $\underset{\sim}{q}$-direction solves dominate the overall computation time, then we can expect that the parallel speedup will scale linearly with the number of processors being used.

In order to examine the scaling of the parallel speedup in practice, we performed computations for two different discrete spaces, such that (i) $N_D = 120$ and $Q_\Omega = 3600$, and (ii) $N_D = 1800$ and $Q_\Omega = 8100$. We solved the enclosed flow problem for these spaces using a number of different choices of N_{proc}. We used method II with basis $\mathcal{B}$ to obtain the data below, but the parallel speedup behaviour is essentially the same whether we use methods I or II or bases $\mathcal{A}$ or $\mathcal{B}$. The base computation times were $T(1) = 0.53$ seconds for the $(N_D, Q_\Omega) = (120, 3600)$ computation, and $T(1) = 157.0$ seconds for the $(N_D, Q_\Omega) = (1800, 8100)$ case.

The parallel speedup of the alternating-direction method for the two discrete spaces discussed above is plotted in Figure 3.5. In the case that $(N_D, Q_\Omega) = (1800, 8100)$, we obtained a parallel speedup of 14.8 when $N = 15$ (*i.e.* $N_{\text{proc}} = 60$), whereas the speedup tailed off to less than 10 when $N = 15$ for the computation with $(N_D, Q_\Omega) =$

$(120, 3600)$. This difference in the scaling of the parallel speedup is primarily due to the fact that the overhead from the redistribution of D^n is much larger, as a proportion of the overall computation time, for the smaller problem. For example, for the $(N_D, Q_\Omega) = (120, 3600)$ problem, matrix redistribution took 8.66% of the overall computation time when $N = 1$, but when $N = 15$, it increased to 30.4%. By contrast, in the larger problem with $(N_D, Q_\Omega) = (1800, 8100)$, more time is spent on the $\underset{\sim}{q}$- and $\underset{\sim}{x}$-direction solves in each time-step, so that only 0.89% of the computation time was taken for the matrix redistribution when $N = 1$, which increased to 2.25% when $N = 15$. Since 2.25% is still only a small proportion of the overall computation time, the matrix redistribution overhead does not significantly detract from the near optimal scaling of the parallel speedup shown in Figure 3.5 for the $(N_D, Q_\Omega) = (1800, 8100)$ case. This indicates that as long as the values of N_D and Q_Ω are large enough, the alternating-direction method can scale efficiently to a very large number of processors.

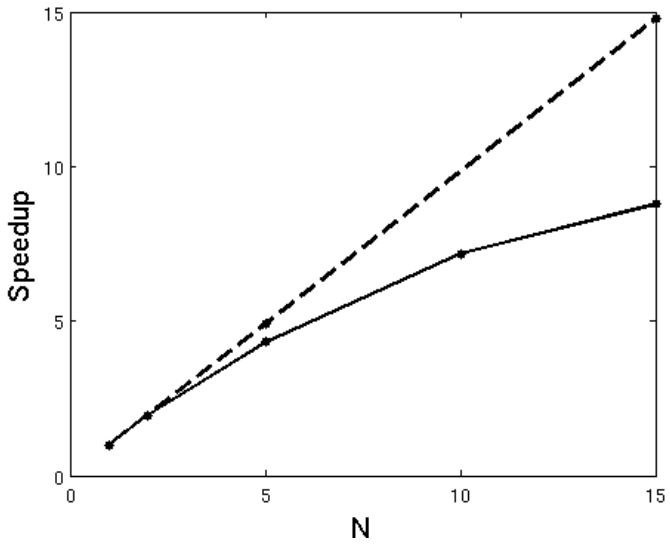


Figure 3.5: Plot of speedup, *i.e.* $T(1)/T(N)$, as the number of computational nodes is increased from 1 to 15. The speedup data for $(N_D, Q_\Omega) = (120, 3600)$ is plotted as a solid line and the dashed line shows the data for $(N_D, Q_\Omega) = (1800, 8100)$. For each computation we chose the number of nodes so that $N_{\text{proc}}(= 4N)$ was a common divisor of N_D and Q_Ω in order to ensure optimal load balancing in each case so that the comparisons of computation time are fair.

3.10 Conclusions

In this chapter we developed an alternating-direction method for the Fokker–Planck equation, which is a hybrid of a classical Douglas–Dupont-type Galerkin alternating-direction scheme, and a new quadrature based scheme. We were able to derive a range of theoretical results for this scheme, including stability results in Section 3.4 and convergence estimates in Section 3.7. Much of this theory built upon the analysis of the Fokker–Planck equation in D that was considered in Chapter 2.

We also put particular emphasis on practical computations in this chapter, and we discussed the implementation of the alternating-direction scheme in Section 3.8, and followed up in Section 3.9 by presenting a range of computational results for alternating-direction methods I and II applied to a model problem with a fixed velocity field, $\underset{\sim}{y}$. We demonstrated that the convergence rates observed in practice for this model problem are accurately described by the theoretical results in Section 3.7. Moreover, we showed that, just as in Chapter 2, the $\underset{\sim}{q}$-direction spectral method yields a more accurate solution for $\underset{\approx}{\tau}$ than it does for $\hat{\psi}$, which means that if we are solely interested in the accuracy of $\underset{\approx}{\tau}$ – as is the case when we consider the Navier–Stokes–Fokker–Planck system – then we can take fewer spectral basis functions than we would need if $\hat{\psi}$ were the quantity of primary interest. This leads to significant savings when we solve the Navier–Stokes–Fokker–Planck system, since the computational work required by the alternating-direction method for the Fokker–Planck equation depends strongly on N_D, the number of $\underset{\sim}{q}$-direction basis functions.

In the next chapter we combine the numerical methods developed in this chapter for the Fokker–Planck equation with a finite element scheme for solving the Navier–Stokes equations to obtain an algorithm for solving the full micro-macro model for dilute polymeric fluids.

Chapter 4

The Coupled Navier–Stokes–Fokker–Planck System

4.1 Introduction

In this chapter we develop an algorithm for solving the Navier–Stokes–Fokker–Planck system, (1.42)–(1.46), and we use this algorithm to obtain computational results for flow problems that are of physical interest. This chapter is relatively brief because the components of our algorithm are already well understood; we use a standard mixed finite element method for solving the Navier–Stokes equations and we couple this to the alternating-direction scheme for the Fokker–Planck equation that was considered in detail in Chapter 3. Theoretical analysis of the coupled algorithm is outside the scope of this dissertation; our focus in this chapter is on obtaining practical computational results. We expect, however, that a convergence analysis along the lines of those developed in the papers [7] and [8] could be pursued in the case of the numerical algorithm applied herein to the coupled Navier–Stokes–Fokker–Planck system.

The numerical method for the Navier–Stokes–Fokker–Planck system is discussed in Section 4.2, and we present numerical results in Section 4.3. Note that throughout this chapter we consider the FENE potential only but, once again, the methodology would be the same for any spring potential that satisfies Hypotheses A and B.

4.2 Numerical method for the micro-macro model

The algorithm we use to couple the numerical methods for the Navier–Stokes equations and the Fokker–Planck equation is essentially the same as those used by Chauvière &

Lozinski [23, 24, 60] and Helzel & Otto [38] for this purpose. We discuss this procedure below, but first we introduce numerical methods for the Navier–Stokes equations, and also for the Stokes equations.

Recall the non-dimensionalised Navier–Stokes equations from Chapter 1, in which $\underset{\sim}{\nabla}_x \cdot \underset{\approx}{\tau}$ arises as a forcing term:

$$\frac{\partial \underset{\sim}{u}}{\partial t} + \underset{\sim}{u} \cdot \underset{\sim}{\nabla}_x \underset{\sim}{u} + \underset{\sim}{\nabla}_x p = \frac{\gamma}{\mathrm{Re}} \Delta_x \underset{\sim}{u} + \frac{b+d+2}{b} \frac{1-\gamma}{\mathrm{Re}\,\mathrm{Wi}} \underset{\sim}{\nabla}_x \cdot \underset{\approx}{\tau}, \tag{4.1}$$

$$\underset{\sim}{\nabla}_x \cdot \underset{\sim}{u} = 0. \tag{4.2}$$

In this chapter we will also consider a Stokes–Fokker–Planck model, which is valid in the limit $\mathrm{Re} \to 0_+$. In the Stokes equations the incompressibility condition (4.2) is unchanged, but we use the following momentum equation (in dimensional form):

$$\underset{\sim}{\nabla}_x p = \nu_s \Delta_x \underset{\sim}{u} + \frac{1}{\rho} \underset{\sim}{\nabla}_x \cdot \underset{\approx}{\tau}, \tag{4.3}$$

instead of (1.11). We non-dimensionalise (4.3) by using (1.24) and the pressure rescaling $p = (\nu U_0 / L_0)\hat{p}$,[1] to obtain:

$$\underset{\sim}{\nabla}_x p = \gamma \Delta_x \underset{\sim}{u} + \frac{b+d+2}{b} \frac{1-\gamma}{\mathrm{Wi}} \underset{\sim}{\nabla}_x \cdot \underset{\approx}{\tau}. \tag{4.4}$$

Next, we introduce mixed finite element approximations of the incompressible Navier–Stokes and Stokes equations. The numerical analysis of these equations is well understood and therefore we discuss our approach only briefly; for further details see [31] or [34].

As in Chapter 3, let $\mathcal{T}_h$ denote a finite element triangulation of $\overline{\Omega}$, and let V_h be the corresponding finite element space with quadratic shape functions that we used for the alternating-direction method for $\hat{\psi}_{h,N}$ in Chapter 3. Also, let P_h denote the $\mathrm{H}^1(\Omega)$-conforming finite element space based on $\mathcal{T}_h$ that uses linear shape functions. Then V_h and P_h are the Taylor–Hood finite element spaces for the Navier–Stokes equations (*cf.* Chapter 5 of [31]); these spaces are known to satisfy the inf-sup stability condition (*cf.* Section 12.6 of [19]). As noted in Chapter 3, in general the Taylor–Hood scheme does not yield a pointwise divergence free velocity field. In the context of the coupled Navier–Stokes–Fokker–Planck system, this may lead to undesirable effects, for example, related to the integral conservation property identified for the Fokker–Planck equation in (3.10). We did not examine the behaviour of this integral property in our numerical experiments presented in Section 4.3, but this is a question of interest for future research.

[1] This pressure scaling is appropriate for creeping flow.

Using the discrete spaces introduced above, our numerical method for the Navier–Stokes system is defined as follows:

Suppose $\underset{\sim}{u}_h^0 \in (V_h)^d$, $p_h^0 \in P_h$ and $\underset{\approx}{\tau}_{h,N}^n \in (\mathrm{L}^2(\Omega))^{d\times d}$ for $n = 0, \ldots, N_T - 1$ are given. Then, for $n = 0, \ldots, N_T - 1$, find $\underset{\sim}{u}_h^{n+1} \in (V_h)^d$ and $p_h^{n+1} \in P_h$ satisfying:

$$\begin{aligned}
&\int_\Omega \frac{\underset{\sim}{u}_h^{n+1} - \underset{\sim}{u}_h^n}{\Delta t} \cdot \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} + \int_\Omega \underset{\sim}{u}_h^{n+1} \cdot \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^{n+1} \cdot \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} - \int_\Omega p_h^{n+1} \underset{\sim}{\nabla}_x \cdot \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} \\
&\quad + \frac{\gamma}{\mathrm{Re}} \int_\Omega \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^{n+1} : \underset{\sim}{\nabla}_x \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} + \frac{b+d+2}{b} \frac{1-\gamma}{\mathrm{Re}\,\mathrm{Wi}} \int_\Omega \underset{\approx}{\tau}_{h,N}^n : \underset{\sim}{\nabla}_x \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} \\
&\quad + \int_{\partial\Omega} \left(p_h^{n+1} \underset{\approx}{I} - \frac{\gamma}{\mathrm{Re}} \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^{n+1} - \frac{b+d+2}{b} \frac{1-\gamma}{\mathrm{Re}\,\mathrm{Wi}} \underset{\approx}{\tau}_{h,N}^n \right) \cdot \underset{\sim}{v} \cdot \underset{\sim}{n}\,\mathrm{d}s = 0 \quad \forall \underset{\sim}{v} \in (V_h)^d, \qquad (4.5) \\
&\int_\Omega q\, \underset{\sim}{\nabla}_x \cdot \underset{\sim}{u}_h^{n+1}\,\mathrm{d}\underset{\sim}{x} = 0 \qquad \forall q \in P_h. \qquad (4.6)
\end{aligned}$$

Note that for tensors $\underset{\approx}{A}$ and $\underset{\approx}{B}$, the colon notation used above is defined as $\underset{\approx}{A} : \underset{\approx}{B} := \sum a_{ij} b_{ij}$.

In this chapter we consider channel flow problems in which we have an inflow boundary, $\partial\Omega_{\mathrm{in}}$, an outflow boundary, $\partial\Omega_{\mathrm{out}}$ and channel wall boundaries $\partial\Omega_0$, such that $\partial\Omega = \partial\Omega_{\mathrm{in}} \cup \partial\Omega_{\mathrm{out}} \cup \partial\Omega_0$. We assume that the channel wall boundaries are stationary and we impose the no-slip boundary condition $\underset{\sim}{u}_h = \underset{\sim}{0}$ on $\partial\Omega_0$. Also, we impose $\underset{\sim}{u}_h = \underset{\sim}{u}_{\mathrm{in}}$ on $\partial\Omega_{\mathrm{in}}$, where $\underset{\sim}{u}_{\mathrm{in}}$ is an inflow velocity profile corresponding to a fully-developed flow. In Section 4.3, the maximum of $\underset{\sim}{u}_{\mathrm{in}}$ is denoted by $U_{\max}$. As a result of these Dirichlet boundary conditions, we have $\underset{\sim}{v} = \underset{\sim}{0}$ on $\partial\Omega_{\mathrm{in}} \cup \partial\Omega_0$. Also, on $\partial\Omega_{\mathrm{out}}$, we impose

$$\left(p_h^{n+1} \underset{\approx}{I} - \frac{\gamma}{\mathrm{Re}} \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^{n+1} - \frac{b+d+2}{b} \frac{1-\gamma}{\mathrm{Re}\,\mathrm{Wi}} \underset{\approx}{\tau}_{h,N}^n \right) = \underset{\approx}{0}.$$

Hence the boundary term in (4.5) vanishes on all of $\partial\Omega$. Note that the $\underset{\approx}{\tau}_{h,N}$ terms in (4.5) are at time-level n rather than $n+1$; we shall see below that this enables us to couple the Fokker–Planck and Navier–Stokes equations in a convenient manner.

The momentum equation, (4.5), is nonlinear due to the term $\int_\Omega \underset{\sim}{u}_h^{n+1} \cdot \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^{n+1} \cdot \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x}$. Hence, we use Newton's method to solve the nonlinear system of equations arising from (4.5) and (4.6) at each time-level.

We now turn our attention to the Stokes equations, which we discretise in a very similar manner. The difference is that we replace (4.5) with the following equation:

$$\begin{aligned}
&- \int_\Omega p_h^{n+1} \underset{\sim}{\nabla}_x \cdot \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} + \gamma \int_\Omega \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^{n+1} : \underset{\sim}{\nabla}_x \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} + \frac{b+d+2}{b} \frac{1-\gamma}{\mathrm{Wi}} \int_\Omega \underset{\approx}{\tau}_{h,N}^n : \underset{\sim}{\nabla}_x \underset{\sim}{v}\,\mathrm{d}\underset{\sim}{x} \\
&\quad + \int_{\partial\Omega} \left(p_h^{n+1} \underset{\approx}{I} - \gamma \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^{n+1} - \frac{b+d+2}{b} \frac{1-\gamma}{\mathrm{Wi}} \underset{\approx}{\tau}_{h,N}^n \right) \cdot \underset{\sim}{v} \cdot \underset{\sim}{n}\,\mathrm{d}s = 0 \qquad \forall \underset{\sim}{v} \in (V_h)^d. \quad (4.7)
\end{aligned}$$

We we apply the same boundary conditions as discussed above for the Navier–Stokes case, and therefore the boundary term in (4.7) vanishes also. Note that there is no time derivative in (4.4), and hence in this case the time dependence comes only through $\underset{\approx}{\tau}^n_{h,N}$ and the boundary data. The Stokes equations are linear and therefore we do not require a Newton scheme in this case.

The mixed finite element methods described above for the Navier–Stokes and Stokes equations were implemented in the finite element library `libMesh` [47]. In both cases, we solve the linear systems that arise from the finite element discretisations using GMRES with incomplete LU factorisation as a preconditioner. In order to obtain faster convergence rates for the iterative solver one could apply more advanced preconditioning techniques, such as the techniques discussed in [31] that take advantage of the structure of the linear systems arising from the discretisation of Stokes or Navier–Stokes problems. However, there is little incentive for us to accelerate the convergence of our Navier–Stokes or Stokes solvers in this way because the overall computation time for computations with the Navier–Stokes–Fokker–Planck system is dominated by solving the Fokker–Planck equation on $\Omega \times D$.

In Chapter 3, we restricted our attention to enclosed flows to simplify the analysis in that chapter, but we are now interested in problems that have inflow and outflow boundaries. Therefore, we need to define the boundary conditions for the Fokker–Planck equation on $\partial\Omega_{\text{in}}$ and $\partial\Omega_{\text{out}}$.

In fact, since the Fokker–Planck equation on Ω is a pure advection problem, we do not need to do anything different on $\partial\Omega_{\text{out}}$ since by definition we have $\underset{\sim}{u}_h \cdot \underset{\sim}{n} > 0$ there.[2] However, we do need to treat the inflow boundary differently. Suppose we set $\underset{\sim}{u}^n_h|_{\partial\Omega_{\text{in}}} = \underset{\sim}{u}^n_{\text{in}}$ for the Stokes/Navier–Stokes system for $n = 1, \ldots, N_T$. Then that boundary data also defines $\underset{\approx}{\kappa}^n_{\text{in}} = \underset{\sim}{\nabla}_x \underset{\sim}{u}^n_{\text{in}}$ on $\partial\Omega_{\text{in}}$,[3] and $\underset{\approx}{\kappa}_{\text{in}}$ in turn determines the inflow boundary data, $\hat{\psi}_{\text{in}}$, on $\partial\Omega_{\text{in}} \times D$ for the Fokker–Planck equation. That is, for $s \in \partial\Omega_{\text{in}}$, $\hat{\psi}^n_{\text{in}}(s,\cdot) : \underset{\sim}{q} \in D \mapsto \hat{\psi}^n_{\text{in}}(s,\underset{\sim}{q}) \in \mathbb{R}$ for $n = 1, \ldots, N_T$ is determined by solving the $\underset{\sim}{q}$-direction Fokker–Planck equation corresponding to $\underset{\approx}{\kappa}^n_{\text{in}}(s)$, so that $\hat{\psi}^n_{\text{in}}(s,\cdot) \in \mathcal{P}_N(D)$ for each n. Writing

$$\hat{\psi}_{\text{in}}(s,\underset{\sim}{q}) = \sum_{k=1}^{N_D} \hat{\psi}_{\text{in},k}(s) Y_k(q), \quad (s,\underset{\sim}{q}) \in \partial\Omega_{\text{in}} \times D,$$

it then follows from (3.31) that $\hat{\psi}_{\text{in},k}$ defines the inflow boundary data on $\partial\Omega_{\text{in}}$ for $\hat{\psi}_k$ in (3.43). In practice we only solve for $\hat{\psi}_{\text{in}}$ at the nodes of $\mathcal{T}_h$ on $\partial\Omega_{\text{in}}$ so that we

[2] $\underset{\sim}{n}$ is the outward unit normal to $\partial\Omega$.

[3] Since $\underset{\sim}{u}_{\text{in}}$ is a fully-developed flow, we assume that the velocity field upstream of $\partial\Omega_{\text{in}}$ has the same profile $\underset{\sim}{u}_{\text{in}}$; this ensures that $\underset{\sim}{\nabla}_x \underset{\sim}{u}_{\text{in}}$ is well-defined on the inflow boundary.

can impose the inflow boundary condition on the line function $\hat{\psi}_k$ in an interpolatory sense. Notice also that we can compute the inflow boundary data for $\hat{\psi}_{h,N}$ before we begin solving the Navier–Stokes–Fokker–Planck system, since $\underset{\sim}{u}_{\text{in}}$ and $\underset{\approx}{\kappa}_{\text{in}}$ are specified *a priori*.

We now define the algorithm for solving the Navier–Stokes–Fokker–Planck system. First of all, we initialise the system to the equilibrium state by setting $\underset{\sim}{u}_h^0 = \underset{\sim}{0}$ on Ω, and therefore $\underset{\approx}{\kappa}^0 = \underset{\sim}{\nabla}_x \underset{\sim}{u}_h^0 = \underset{\approx}{0}$ on Ω also. Putting $\underset{\approx}{\kappa} = \underset{\approx}{0}$ in (2.59), we can see that $\psi = M$ is the corresponding equilibrium steady-state solution, and hence we set $\hat{\psi}_{h,N}^0 = \sqrt{M} \in V_h \otimes \mathcal{P}_N(D)$ on $\Omega \times D$.[4] Also, for consistency with $\hat{\psi}_{h,N}^0$, we set $\underset{\approx}{\tau}_{h,N}^0 = \underset{\approx}{I}$ on Ω. Then, for $n = 0, \ldots, N_T - 1$, we perform the following steps:

1. Compute $\underset{\sim}{u}_h^{n+1} \in V_h$ and $p_h^{n+1} \in P_h$ using the mixed finite element method discussed above for either the Navier–Stokes or Stokes system. We use the tensor $\underset{\approx}{\tau}_{h,N}^n$ in (4.5) or (4.7).

2. Use method I or method II to compute $\hat{\psi}_{h,N}^{n+1} \in V_h \otimes \mathcal{P}_N(D)$ with $\underset{\approx}{\kappa}^n$ in (3.40) for method I or with $\underset{\approx}{\kappa}^{n+1}$ in (3.56) for method II, and $\underset{\sim}{u}_h^{n+1}$ in (3.43) for either method.

3. Using (1.45), compute $\underset{\approx}{\tau}_{h,N}^{n+1}$ on Ω based on $\hat{\psi}_{h,N}^{n+1} \in V_h \otimes \mathcal{P}_N(D)$.

4. Return to 1. and continue marching in time.

Note that the $\underset{\approx}{\tau}_{h,N}$ terms in the momentum equations (4.5) or (4.7) are explicit in time. This allows the Stokes/Navier–Stokes equations to be coupled to the Fokker–Planck equation in a simple manner, but the drawback is that the algorithm defined in steps 1. to 4. above is only conditionally stable. In Section 4.3 we use $\Delta t = 0.01$ and this time-step size is sufficiently small to yield a reliable numerical method for the micro-macro problems that we consider.

4.3 Numerical Results

In this section, we consider two distinct problems. The first is a planar contraction flow in the $d = 2$ case, which we discuss in Section 4.3.1, and the second is a flow around a sphere in the $d = 3$ case, considered in Section 4.3.2. For each of these two problems we present numerical results for one particular discrete space $V_h \otimes \mathcal{P}_N(D)$, but in each case we performed mesh refinement studies (*i.e.* we solved using a sequence

[4] We assume here that $\sqrt{M} \in \mathcal{P}_N(D)$, which is reasonable according to Remark 2.17.

of increasingly refined spaces) to ensure that the numerical results shown below are accurate.

4.3.1 4–1 planar contraction flow

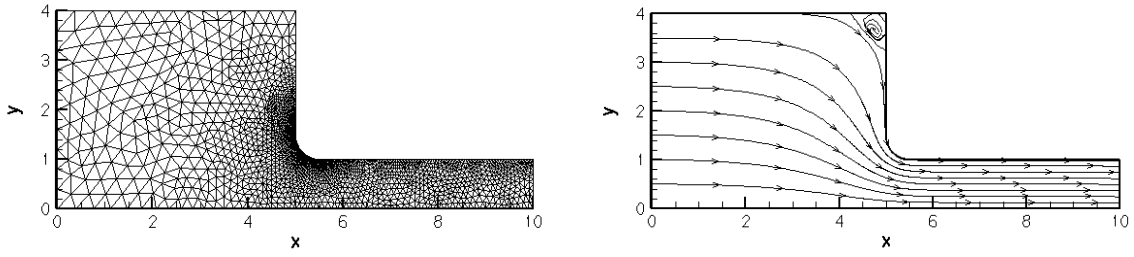
Contraction flows are standard benchmark problems in computational rheology because they are challenging from the numerical point of view and they also have practical relevance in industrial applications (for a detailed discussion of contraction flows see Chapter 8 of [69]). In this section we consider the coupled Navier–Stokes–Fokker–Planck model with $\mathrm{Re} = 1$ in a contracting domain, which is 10 units long, 4 units wide in the wider section and 1 unit wide in the narrow section. We set $\partial\Omega_{\mathrm{in}}$ and $\partial\Omega_{\mathrm{out}}$ to be the left-hand and right-hand boundaries of Ω, respectively, and we let the top edge boundary be $\partial\Omega_0$. In this case, to save computational work we also imposed a symmetry boundary condition on the bottom boundary by setting the y-component of $\underset{\sim}{y}_h$ to zero there. We set $\underset{\sim}{y}_{\mathrm{in}}$ to be a parabolic inflow profile, corresponding to steady Poiseuille flow in a channel, that vanishes at the top boundary and achieves its maximum value of $U_{\max} = 1$ at the symmetry boundary.

As specified in Chapter 3, we need $\underset{\approx}{\kappa} = \underset{\sim}{\nabla}_x \underset{\sim}{y}_h \in \mathrm{L}^\infty(\Omega)$ in order to use alternating-direction methods I or II. Clearly, for any finite element approximation, $\underset{\sim}{y}_h$, this condition will be satisfied. Nevertheless, for the moment, let us consider the weak solution, $\underset{\sim}{y} \in \mathrm{H}^k(\Omega)$ for some $k > 0$. In order to guarantee that $\underset{\sim}{\nabla}_x \underset{\sim}{y} \in \mathrm{L}^\infty(\Omega)$, we require the embedding $\mathrm{H}^{k-1}(\Omega) \subset \mathrm{L}^\infty(\Omega)$ to hold; a sufficient condition for this embedding is that $k > 2$. However, contraction flows of polymeric fluids are typically simulated using 'L-shaped' domains and it is well known that the Stokes and Navier–Stokes equations exhibit a corner singularity on domains of this type so that in general $\underset{\sim}{y} \notin \mathrm{H}^2(\Omega)$ (*cf.* Remark 5.10 in [31]). Therefore, $\underset{\sim}{\nabla}_x \underset{\sim}{y}$ will not, in general, belong to $\mathrm{L}^\infty(\Omega)$, and hence the sequence $\underset{\approx}{\kappa}_h = \underset{\sim}{\nabla}_x \underset{\sim}{y}_h$ will not be uniformly bounded in h as $h \to 0_+$. As a result, instead of an L-shaped domain, we use the physical space domain with a rounded corner shown in Figure 4.1. Also, in order to resolve the solution satisfactorily, the finite element mesh, $\mathcal{T}_h$, has been graded so that it is finer near the corner.

We applied the algorithm defined in Section 4.2 for the coupled Navier–Stokes–Fokker–Planck system to the contraction flow problem described above. We set $b = 12$, $\mathrm{Wi} = 0.8$, $\gamma = 0.59$ and took 500 time-steps with $\Delta t = 0.01$ so that $T = 5$. We used alternating-direction method II with basis $\mathcal{A}$ and the $p = 4$ quadrature rule on triangles for which $Q_{\hat{K}} = 6$ (*cf.* Section 3.8.3) so that QH2 was satisfied. The

mesh $\mathcal{T}_h$ contained 905 triangular finite elements and therefore $Q_\Omega = 5430$.[5] Also, we used $(N_r, N_\theta) = (20, 20)$ for the $\underset{\sim}{q}$-direction spectral method, so that $N_D = 820$. The macroscopic velocity field at $T = 5$ is plotted in Figure 4.1(b) and the corresponding components of $\underset{\approx}{\tau}_{h,N}$ are shown in Figure 4.2. The computation was performed using 40 processors of the Lonestar supercomputer at the Texas Advanced Computing Centre using the parallel implementation of the alternating-direction method described in Section 3.8.4, and each time-step took 1.16 seconds.

As shown in Table 2.4, the backward Euler temporal discretisation of the Fokker–Planck equation in the $\underset{\sim}{q}$-direction is more stable than the semi-implicit discretisation in the case that $\mathrm{Wi}\,\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(\Omega)} = 5$. Therefore, for the contraction flow problem considered here, in which $\mathrm{Wi}\,\|\underset{\approx}{\kappa}\|_{\mathrm{L}^\infty(\Omega)} \approx 10$ (the maximum $\underset{\approx}{\kappa}$ values occur near the corner), the stability advantage of method II outweighs method I's advantage of lower computational cost per time-step.



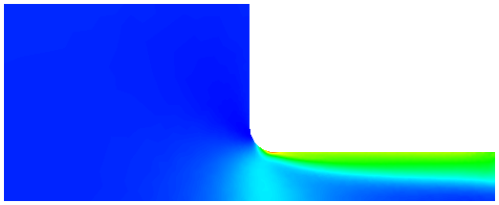
(a) (b)

Figure 4.1: (a) The finite element mesh $\mathcal{T}_h$ used for the contraction flow computations. $\mathcal{T}_h$ contains 905 triangular elements. (b) Streamlines for the macroscopic velocity field; this corresponds closely to the Figure 8.9 in [69], which shows computational results for planar contraction flows obtained using the fully macroscopic Oldroyd B model.

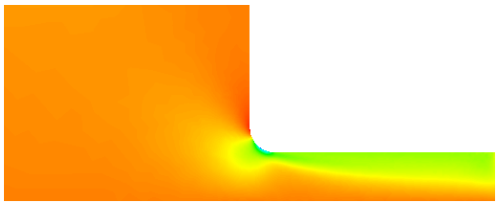
4.3.2 Flow around a sphere

The planar flow of a polymeric fluid around a cylindrical obstacle in a channel has also been a popular benchmark problem in the computational rheology literature (see Chapter 9 of [69]). In this section we consider a three-dimensional analogue in which we solve the micro-macro model for a suspension of FENE dumbbells for the flow around a sphere with radius 1 in a three-dimensional channel with 4×4 square cross-section.

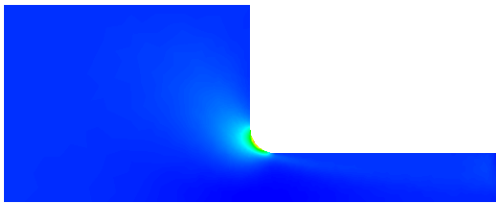
[5] 6335 quadrature points would have been required to satisfy QH1; hence we obtain a significant reduction in the number of $\underset{\sim}{q}$-direction solves per time-step by satisfying QH2 only.



τ_{11}



τ_{12}



τ_{22}

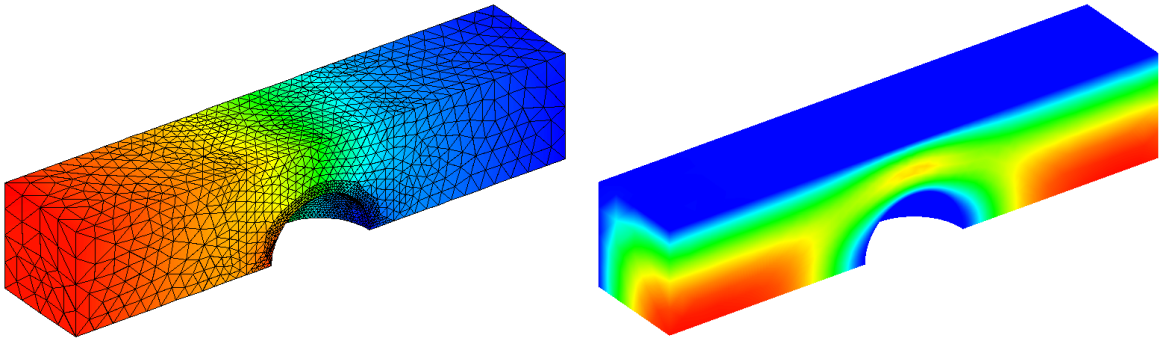
Figure 4.2: The components of $\underset{\sim}{\tau}_{h,N}$ at $T = 5$. In the τ_{11} plot, values range from 0.45 (blue) to 15.7 (red), in the τ_{12} ($= \tau_{21}$) plot we have -9.75 (blue) to 1.41 (red) and in the τ_{22} plot, 0.46 (blue) to 11.5 (red). The polymeric extra-stress is largest in the region near the rounded corner.

In this case $\Omega \subset \mathbb{R}^3$ and $\Omega \times D \subset \mathbb{R}^6$. We set $b = 12$, $\mathrm{Wi} = 1$, $\gamma = 0.59$ and we used the Stokes equations for the macroscopic velocity field.

The mesh $\mathcal{T}_h$ is shown in Figure 4.3. We set $\underset{\sim}{u}_{\mathrm{in}}$ to be the velocity profile corresponding to steady Stokes flow in a channel with square cross-section, with $U_{\max} = 1$. We also imposed a no-slip boundary condition condition on the channel walls and on the spherical obstacle, and we set two symmetry boundary conditions so that we only needed to simulate the flow in one quarter of the domain. We again used alternating-direction method II for this problem since $\mathrm{Wi}\,\|\underset{\approx}{K}\|_{\mathrm{L}^\infty(\Omega)} \approx 5$.

The mesh $\mathcal{T}_h$ contains 5150 tetrahedral elements. According to Section 3.8.3, we require $Q_{\hat{K}} = 14$ in order to satisfy either QH1 or QH2, and hence we have $Q_\Omega = 72100$. For the $\underset{\sim}{q}$-direction spectral method we used basis $\mathcal{C}$ with $(N_r, N_{\mathrm{sph}}) = (12, 12)$, so that $N_D = 1092$. Therefore, in each time-step, 72100 three-dimensional $\underset{\sim}{q}$-direction solves and 1092 three-dimensional $\underset{\sim}{x}$-direction solves were performed. We took 100 time-steps

with $\Delta t = 0.01$ to reach $T = 1$. Plots of the x-component of $\underset{\sim}{y}_h$ and of p_h at $T = 1$ are shown in Figure 4.3. Also, the components of the polymeric extra-stress tensor at $T = 1$ are shown in Figure 4.4. This computation was performed with $N_{\text{proc}} = 128$ and it took 38.7 seconds to evaluate each time-step of the coupled Stokes-Fokker–Planck system.



(a) (b)

Figure 4.3: (a) Plot of the pressure, $p_h \in P_h$, at $T = 1$, with values ranging from 0.5 (blue) to 14.4 (red). Also, this plot shows the mesh $\mathcal{T}_h$. Note that the mesh is very fine in the vicinity of the spherical obstacle in order to resolve the solution structure in that region. (b) The x-component of the macroscopic velocity field at $T = 1$; values range from 0 (blue) to 1 (red).

4.4 Conclusions

In this chapter we introduced a deterministic multiscale algorithm for the micro-macro model of dilute polymeric fluids. This algorithm couples the alternating-direction scheme from Chapter 3 to a finite element method (for Stokes or Navier–Stokes) for computing the macroscopic velocity field. We used this algorithm to simulate two channel flows; a 4–to–1 contraction (with a rounded reentrant corner to avoid a singularity in $\underset{\sim}{y}$) in Section 4.3.1, and a flow around a spherical obstacle in a channel with square cross-section in Section 4.3.2.

We made extensive use of parallel computation in order to obtain the computational results in Section 4.3. In particular, to the best of our knowledge the micro-macro model has not previously been used in the case that $\Omega \times D \in \mathbb{R}^6$ and this was only made feasible in Section 4.3.2 through the use of large-scale parallel computation.

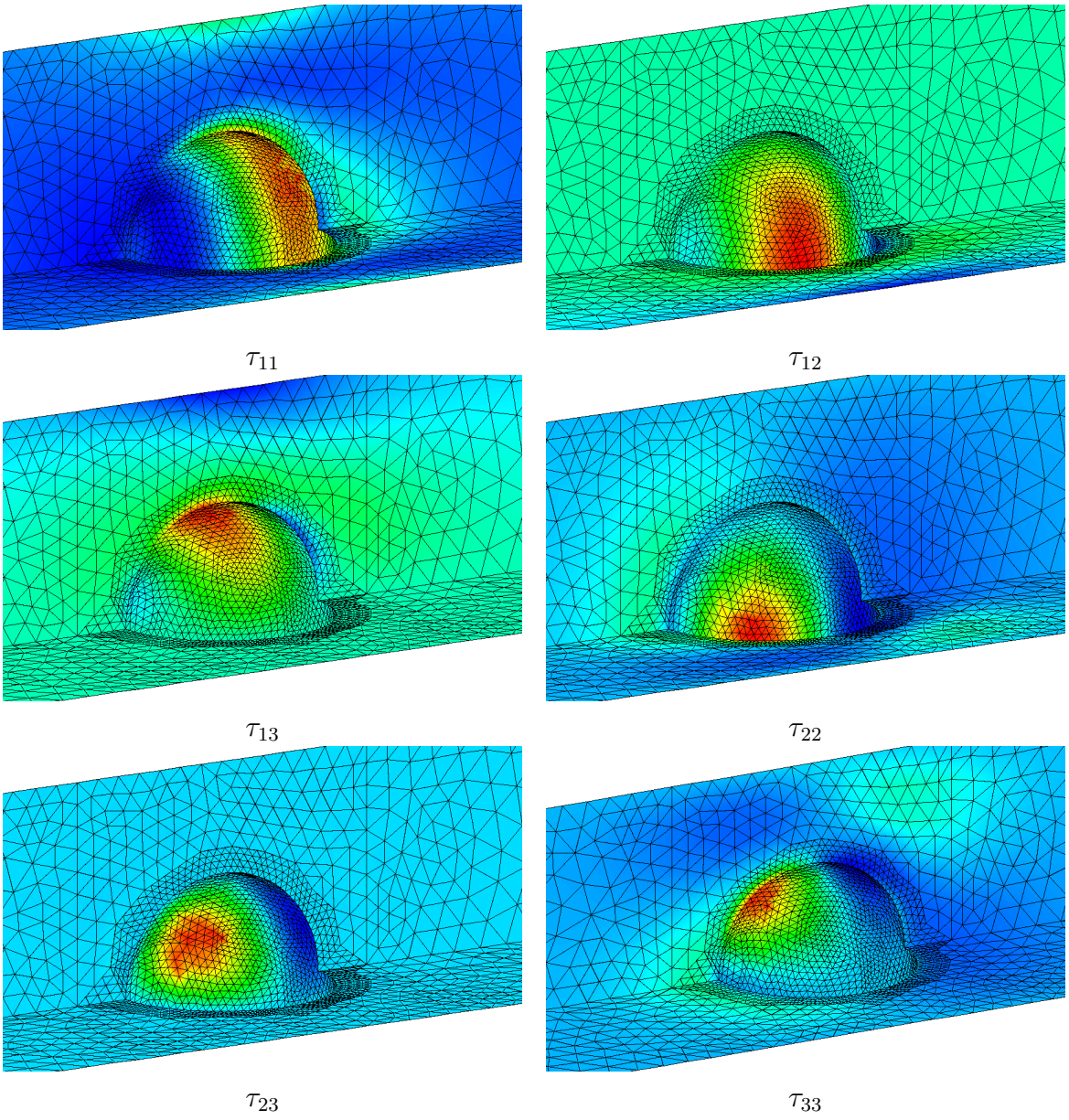


Figure 4.4: Plots of the components of the polymeric extra-stress tensor, $\underset{\approx}{\tau}_{h,N}$, at $T = 1$ for the channel flow around a spherical obstacle. The minimum (blue) and maximum (red) values in each plot are as follows; τ_{11}: 0.53 to 6.25, τ_{12}: -1.25 to 2.41, τ_{13}: -1.21 to 2.5, τ_{22}: 0.48 to 3.35, τ_{23}: -0.33 to 1.15 and τ_{33}: 0.47 to 3.46.

Chapter 5

Conclusions

In this dissertation we have considered the analysis and implementation of numerical methods for solving the multiscale Navier–Stokes–Fokker–Planck system, (1.42)–(1.46), which models the flow of dilute polymeric fluids. From both the theoretical and computational point of view, the most challenging component of this coupled model is the high-dimensional Fokker–Planck equation, (1.44), which is posed on the domain $\Omega \times D$ in $2d$ spatial dimensions. Hence, most of our attention was focused on the Fokker–Planck equation, and we developed a computational framework for this equation that is efficient in practice and is also underpinned by rigorous theoretical analysis.

First of all, in Chapter 2, we considered the Fokker–Planck equation on D only. We derived a range of analytical results for the weak solution of this equation, and we proved stability bounds and optimal order convergence estimates for a Galerkin spectral method for this problem. These results were obtained for the Maxwellian-transformed Fokker–Planck equation for any spring potential satisfying Hypotheses A and B. This transformation led to a convenient symmetrisation of the principal part of the differential operator, and consequently facilitated the derivation of theoretical results. We also considered an alternative transformation of the FENE Fokker–Planck equation due to Lozinski & Chauvière [24], in which $\hat{\psi} := \psi/M^{2s/b}$. We showed that the Chauvière–Lozinski-transformed Fokker–Planck equation is well-posed as long as $b \geq 4s^2/(2s-1)$ and $s > 1/2$, and in a series of remarks in Chapter 2, we indicated how one could extend the results that were derived for the Maxwellian-transformed formulation to the Chauvière–Lozinski formulation.

In Section 2.6, we presented a range of computational results for the spectral method on D based on the Maxwellian-transformed formulation. We demonstrated that this spectral method exhibits the rapid spatial convergence characterised in the convergence estimates in Section 2.5. We also compared this method to the method of Chauvière & Lozinski based on numerical results reported in [24], and we showed that, at least for

moderate values of b, the two methods converge at a comparable rate. In the context of the Navier–Stokes–Fokker–Planck system, the convergence of the extra-stress tensor $\underset{\approx}{\tau}$ is, in fact, more important than the convergence rate of ψ and we demonstrated using an argument based on Parseval's identity that for our Galerkin spectral method, the error in $\underset{\approx}{\tau}$ will typically be much smaller than the error in ψ. We demonstrated that in practical computations this manifests itself as superconvergence of $\underset{\approx}{\tau}$.

In Chapter 3 we introduced an alternating-direction framework for solving the Fokker–Planck equation on $\Omega \times D \in \mathbb{R}^{2d}$. Approaches of this type have been used successfully for this problem already, *e.g.* see the work of Chauvière & Lozinski [23, 24, 60] or Helzel & Otto [38]. However, these authors did not consider the behaviour of their alternating-direction schemes in detail from a theoretical point of view, whereas the numerical analysis of our alternating-direction scheme is a priority in Chapter 3. We proposed a hybrid alternating-direction method that combines a Douglas–Dupont-type Galerkin alternating-direction method in the $\underset{\sim}{x}$-direction with a new quadrature-based scheme in the $\underset{\sim}{q}$-direction. We were able to establish a number of theoretical results for this algorithm and, in particular, we proved an *a priori* convergence estimate for method I in the case that Quadrature Hypothesis 1 is satisfied.

We tested our computational approach on an enclosed flow problem with a fixed velocity field. The results from these computations are shown in Section 3.9 and we showed that the convergence rates that we observe in practice conform to the theoretical estimates obtained in Section 3.7. We obtained these results using a parallel implementation of the numerical method, and we also used this enclosed flow model problem to study the parallel speedup obtained when the number of processors, N_{proc}, is increased. We showed that the parallel implementation of the alternating-direction method can scale well to a large number of processors.

Finally, in Chapter 4, we coupled the alternating-direction method from Chapter 3 to a mixed finite element method for the Navier–Stokes or Stokes equations to obtain an algorithm for the full micro-macro model for dilute polymeric fluids. We used this algorithm to obtain computational results for a 4–to–1 contraction flow in the $d = 2$ case, and also for a flow around a spherical obstacle in a channel with square cross-section in the $d = 3$ case. We used parallel computation again in order to significantly reduce the computation time that was required for these problems. To the best of our knowledge the flow around a sphere problem considered in Section 4.3.2 is the first time the micro-macro model has been solved in a case where $\Omega \times D \subset \mathbb{R}^6$.

There has already been a lot of impressive work on the development of practical deterministic multiscale methods for simulating dilute polymeric fluids, most notably by

Lozinski, Chauvière and collaborators [23, 24, 59, 60, 61]. These authors demonstrated the feasibility of the deterministic multiscale approach for some classes of problems and showed that it has a number of advantages over more well-established approaches such as stochastic or fully macroscopic methods. However, the deterministic multiscale method has not previously been the subject of detailed numerical analysis. As discussed above, the primary contribution of this thesis has been to advance the theoretical understanding of numerical methods for the Fokker–Planck equation, which, for deterministic multiscale numerical methods, is the pivotal component of the micro-macro model. Also, we developed practical numerical methods that conformed to the hypotheses of our analytical results to ensure that these methods are based on rigorous mathematical foundations.

5.1 Future directions

The work in this thesis could be extended in a number of ways.

First of all, we did not consider the numerical analysis of the algorithm for the coupled Navier–Stokes–Fokker–Planck equation in Chapter 4. It would be interesting and useful to develop theoretical results for this scheme, especially in order to quantify the time-step limitation introduced by the conditionally stable coupling scheme, and perhaps to consider other schemes that avoid such a restriction. Numerical analysis of the Navier–Stokes–Fokker–Planck system has been considered in the papers [7, 8] by Barrett & Süli. In [7], the authors showed convergence for a general family of Galerkin-type methods in the corotational case and this is extended, for a finite element discretisation, to the general noncorotational case in [8].

Another possibly fruitful direction would be to develop a numerical framework for simulating dilute polymeric fluids in which the polymer molecules are modelled as bead-spring chains. As described in Chapter 1, this would lead to a higher-dimensional configuration space, and is consequently much more challenging than the dumbbell case from a computational point of view. This topic has already received a lot of attention in the literature, although so far the emphasis has been on the Fokker–Planck equation for homogeneous flows, *i.e.* in which there is no $\underset{\sim}{x}$-dependence. Therefore, it is conceivable that a good method for simulating a suspension of bead-spring chains would be to use the alternating-direction framework developed in Chapter 3, except with a $\underset{\sim}{q}$-direction numerical method that is appropriate for the Fokker–Planck equation in a high-dimensional configuration space (*e.g.* a sparse grid or reduced basis method, see Section 1.4), rather than the $\underset{\sim}{q}$-direction spectral method that we used in this work.

Another interesting direction of future research would be to apply the alternating-direction methodologies that we have developed for the Fokker–Planck equation to other equations that may also be well suited to such methods. For example, the Vlasov–Fokker–Planck equation for modelling electrostatic plasmas [37,83] has a similar structure to the Fokker–Planck equation for polymeric fluids in that it is also posed on a domain that is the cartesian product of a physical space and a configuration space domain. Indeed, operator splitting methods have already been applied to the Vlasov–Fokker–Planck equation in [37].

Finally, we made extensive use of parallel computation in this dissertation, and we showed that our alternating-direction schemes are very well suited to implementation on parallel architectures. It would be interesting to use the alternating-direction methods developed in this thesis on much larger numbers of processors than we have considered here, for example, with N_{proc} being on the order of 10^3 or 10^4. This would presumably enable much larger problems than even the flow around the sphere considered in Section 4.3.2 to be solved. It is also conceivable, however, that for very large-scale problems, each $\underset{\sim}{q}$-direction or $\underset{\sim}{x}$-direction subproblem may be 'too large' to be solved on a single processor, as we proposed in Section 3.8.4. If we sought to solve each subproblem using multiple processors, then interesting issues related to load balancing and mesh partitioning may arise; for example, we could presumably couple the partitioning strategy outlined in Section 3.8.4 with standard algorithms for solving PDEs on parallel architectures, *e.g.* mesh partitioning or domain decomposition strategies.

References

[1] D. J. Acheson. *Elementary Fluid Dynamics*. Oxford University Press, 1990.

[2] A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings. A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modeling of complex fluids. *J. Non-Newtonian Fluid Mech.*, 139:153–176, 2006.

[3] A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings. A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modelling of complex fluids. part ii: Transient simulation using space-time separated representations. *J. Non-Newtonian Fluid Mech.*, 144:98–121, 2007.

[4] F. G. Avkhadiev and K.-J. Wirths. Unified Poincaré and Hardy inequalities with sharp constants for convex domains. *ZAMM Z. Angew. Math. Mech.*, 87(8-9):632–642, 2007.

[5] S. Balay, K. Buschelman, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 2.1.5, Argonne National Laboratory, 2004.

[6] J. W. Barrett, Ch. Schwab, and E. Süli. Existence of global weak solutions for some polymeric flow models. *Math. Models and Methods in Applied Sciences*, 15(3):939–983, 2005.

[7] J. W. Barrett and E. Süli. Numerical approximation of corotational dumbbell models for dilute polymers. *IMA Journal of Numerical Analysis*. Accepted for publication, 19 March 2008.

[8] J. W. Barrett and E. Süli. Numerical approximation of kinetic dilute polymer models with microscopic cut-off. In preparation, 2008.

[9] J. W. Barrett and E. Süli. Existence of global weak solutions to dumbbell models for dilute polymers with microscopic cut-off. *M3AS: Mathematical Models and Methods in Applied Sciences*, 18(6):935–971, 2008.

[10] J. W. Barrett and Endre Süli. Existence of global weak solutions to some regularized kinetic models for dilute polymers. *Multiscale Model. Simul.*, 6(2):506–546 (electronic), 2007.

[11] G. K. Batchelor. *An Introduction to Fluid Dynamics*. Cambridge University Press, 1967.

[12] C. Bernardi and Y. Maday. Spectral methods. In P.G. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis*, volume V. Elsevier, 1997.

[13] O. V. Besov, Ja. Kadlec, and A. Kufner. Certain properties of weight classes. *Dokl. Akad. Nauk SSSR*, 171:514–516, 1966.

[14] O. V. Besov and A. Kufner. The density of smooth functions in weight spaces. *Czechoslovak Math. J.*, 18 (93):178–188, 1968.

[15] A. V. Bhave, R. C. Armstrong, and R. A. Brown. Kinetic theory and rheology of dilute, nonhomogeneous polymer structures. *J. Chem. Phys.*, 95:2988–3000, 1991.

[16] B. Bialecki and R. Fernandes. An orthogonal spline collocation alternating direction implicit crank-nicolson method for linear parabolic problems on rectangles. *SIAM J. Numer. Anal.*, 36(5):1414–1434, 1999.

[17] R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager. *Dynamics of Polymeric Liquids, Volume 1, Fluid Mechanics*. John Wiley and Sons, second edition, 1987.

[18] R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager. *Dynamics of Polymeric Liquids, Volume 2, Kinetic Theory*. John Wiley and Sons, second edition, 1987.

[19] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, second edition, 2002.

[20] C. Canuto, A. Quarteroni, M. Y. Hussaini, and T. A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer, 2006.

[21] M. Celia and G. Pinder. An analysis of alternating-direction methods for parabolic equations. *Numerical Methods for Partial Differential Equations*, (1):57–70, 1985.

[22] M. Celia and G. Pinder. Generalized alternating-direction collocation methods for parabolic equations. i. spatially varying coefficients. *Numerical Methods for Partial Differential Equations*, (3):193–214, 1990.

[23] C. Chauvière and A. Lozinski. Simulation of complex viscoelastic flows using Fokker–Planck equation: 3D FENE model. *J. Non-Newtonian Fluid Mech.*, 122:201–214, 2004.

[24] C. Chauvière and A. Lozinski. Simulation of dilute polymer solutions using a Fokker–Planck equation. *Computers and Fluids*, 33:687–696, 2004.

[25] P. Clément. Approximation by finite element functions using local regularization. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. RAIRO Analyse Numérique*, 9(R-2):77–84, 1975.

[26] W. T. Coffey, Y. P. Kalmykov, and J. T. Waldron. *The Langevin Equation: With Applications in Physics, Chemistry and Electrical Engineering.* World Scientific, 1996.

[27] P. Delaunay, A. Lozinski, and R. G. Owens. Sparse tensor-product Fokker-Planck-based methods for nonlinear bead-spring chain models of dilute polymer solutions. *CRM Proceedings and Lecture Notes*, 41:73 – 89, 2007.

[28] J. Douglas and T. DuPont. Alternating-direction galerkin methods on rectangles. *Numerical Solution of Partial Differential Equations, II (SYNSPADE 1970)*, pages 133–214, 1971.

[29] Q. Du, C. Liu, and P. Yu. FENE dumbbell model and its several linear and nonlinear closure approximations. *Multiscale Model. Simul.*, 4(3):709–731, 2005.

[30] H. Eisen, W. Heinrichs, and K. Witsch. Spectral collocation methods and polar coordinate singularities. *J. Comput. Phys.*, 96(2):241–257, 1991.

[31] H. Elman, D. Silvester, and A. Wathen. *Finite elements and fast iterative solvers.* Oxford Science Publications, 2005.

[32] X. J. Fan. Molecular models and flow calculations: II. simulation of steady planar flow. *Acta Mechanica Sinica*, 5:216–226, 1989.

[33] P. J. Flory. *Statistical Mechanics of Chain Molecules*. Wiley-Interscience, 1969.

[34] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*. Springer, 1986.

[35] M. Grosso, P. L. Maffettone, P. Halin, R. Keunings, and V. Legat. Flow of nematic polymers in eccentric cylinder geometry: influence of closure approximations. *J. Non-Newtonian Fluid Mech.*, 94:119–134, 2000.

[36] P. Halin, G. Lielens, R. Keunings, and V. Legat. The Lagranian particle method for macroscopic and micro-macro viscoelastic flow computations. *J. Non-Newtonian Fluid Mech.*, 79:387–403, 1998.

[37] K. J. Havlak and H. D. Victory, Jr. On deterministic particle methods for solving Vlasov-Poisson-Fokker-Planck systems. *SIAM J. Numer. Anal.*, 35(4):1473–1519 (electronic), 1998.

[38] C. Helzel and F. Otto. Multiscale simulations of suspensions of rod-like molecules. *J. Comp. Phys.*, 216:52–75, 2006.

[39] W. Huang and B. Guo. Fully discrete Jacobi-spherical harmonic spectral method for Navier-Stokes equations. *Appl. Math. Mech. (English Ed.)*, 29(4):453–476, 2008.

[40] M. A. Hulsen, A. P. G. van Heel, and B. H. A. A. van den Brule. Simulation of viscoelastic flows using Brownian configuration fields. *J. Non-Newtonian Fluid Mech.*, 70:79–101, 1997.

[41] B. Jourdain and T. Lelièvre. Mathematical analysis of a stochastic differential equation arising in the micro-macro modelling of polymeric fluids. *Probabilistic Methods in Fluids*, pages 205–223, 2003.

[42] B. Jourdain, T. Lelièvre, and C. Le Bris. Numerical analysis of micro-macro simulations of polymeric fluid flows: A simple case. *Math. Models Methods Appl. Sci.*, 12:1205–1243, 2002.

[43] B. Jourdain, T. Lelièvre, and C. Le Bris. Existence of solution for a micro-macro model of polymeric fluid: the FENE model. *J. Funct. Anal.*, 209(1):162–193, 2004.

[44] P. Keast. Moderate-degree tetrahedral quadrature formulas. *Comput. Methods Appl. Mech. Engrg.*, 55(3):339–348, 1986.

[45] R. Keunings. On the Peterlin approximation for finitely extensible dumbbells. *J. Non-Newtonian Fluid Mech.*, 68:85–100, 1997.

[46] R. Keunings. Micro–macro methods for the multiscale simulation of viscoelastic flow using molecular models of kinetic theory. *Rheology Review*, pages 67–98, 2004.

[47] B. S. Kirk, J. W. Peterson, R. M. Stogner, and Carey G. F. libmesh: A C++ library for parallel adaptive mesh refinement/coarsening simulations. *Engineering with Computers*, 23(3–4):237–254, 2006.

[48] J. G. Kirkwood. *Macromolecules*. Gordon and Breach, 1967.

[49] D. J. Knezevic and E. Süli. Spectral galerkin approximation of Fokker–Planck equations with unbounded drift. *Submitted to M2AN, January 2008*.

[50] A. N. Kolmogorov. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math. Ann.*, 104, 1931.

[51] H. A. Kramers. The viscosity of macromolecules in a streaming fluid. *Physica*, 11(1), 1944.

[52] A. Kufner. *Weighted Sobolev Spaces*. Teubner-Texte zur Mathematik. Teubner, 1980.

[53] B. Lapeyre, É. Pardoux, and R. Sentis. *Introduction to Monte-Carlo Methods for Transport and Diffusion Equations*. Oxford University Press, 2003.

[54] M. Laso and H. C. Öttinger. Calculation of viscoelatic flow using molecular models: the CONNFFESSIT approach. *J. Non-Newtonian Fluid Mech.*, 47:1–20, 1993.

[55] T. Li and P. Zhang. Mathematical analysis of multi-scale models of complex fluids. *Commun. Math. Sci.*, 5(1):1–51, 2007.

[56] G. Lielens, P. Halin, I. Jaumain, R. Keunings, and V. Legat. New closure approximations for the kinetic theory of finitely extensible dumbbells. *J. Non-Newtonian Fluid Mech.*, 76:249–279, 1998.

[57] P.-L. Lions and N. Masmoudi. Global solutions for some Oldroyd models of non-Newtonian flows. *Chinese Ann. Math. Ser. B*, 21:131–146, 2000.

[58] P.-L. Lions and N. Masmoudi. Global existence of weak solutions to some micro-macro models. *C. R. Math. Acad. Sci. Paris*, 345:15–20, 2007.

[59] A. Lozinski. *Spectral methods for kinetic theory models of viscoelastic fluids*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2003.

[60] A. Lozinski and C. Chauvière. A fast solver for Fokker–Planck equation applied to viscoelastic flows calculation: 2D FENE model. *Journal of Computational Physics*, 189:607–625, 2003.

[61] A. Lozinski, C. Chauvière, J. Fang, and R. G. Owens. Fokker–Planck simulations of fast flows of melts and concentrated polymer solutions in complex geometries. *J. Rheology*, 47:535–561, 2003.

[62] J. N. Lyness and D. Jespersen. Moderate degree symmetric quadrature rules for the triangle. *J. Inst. Math. Appl.*, 15:19–32, 1975.

[63] M. Marcus, V. J. Mizel, and Y. Pinchover. On the best constant for Hardy's inequality in $\mathbf{R}^n$. *Trans. Amer. Math. Soc.*, 350(8):3237–3255, 1998.

[64] T. Matsushima and P. S. Marcus. A spectral method for polar coordinates. *J. Comput. Phys.*, 120:365–374, 1995.

[65] D. A. McQuarrie. *Statistical Mechanics*. University Science Books.

[66] R. Nayak. *Molecular simulation of liquid crystal polymer flow: a wavelet-finite element analysis*. PhD thesis, MIT, 1998.

[67] J. G. Oldroyd. On the formulation of rheological equations of state. *Proc. Roy. Soc. London*, pages 523–541, 1950.

[68] H. C. Öttinger. *Stochastic Processes in Polymeric Fluids*. Springer, 1996.

[69] R. G. Owens and T. N. Phillips. *Computational Rheology*. Imperial College Press, 2002.

[70] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 2 edition, 1984.

[71] P. E. Rouse. A theory of the linear viscoelastic properties of dilute solutions of coiling polymers. *J. Chem. Phys.*, 21:1272–1280, 1953.

[72] J. D. Schieber and H. C. Öttinger. The effects of bead inertia on the rouse model. *J. Chem. Phys.*, 89(11), 1988.

[73] C. Schwab, E. Süli, and R.-A. Todor. Sparse finite element approximation of high-dimensional transport-dominated diffusion problems. *M2AN: Mathematical Modelling and Numerical Analysis*. (Accepted for publication, 3 April 2008).

[74] L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.

[75] J. Shen. Efficient spectral galerkin methods III: Polar and cylindrical geometries. *SIAM J. Sci. Comput.*, 18(6):1583–1604, 1997.

[76] W. E. Stewart and J. P. Sørensen. Hydrodynamic interaction effects in rigid dumbbell suspensions. II. computations for steady shear flow. *Journal of Rheology*, 16(1):1–13, 1972.

[77] E. Süli and D. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, 2003.

[78] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. Second edition. Joh. Ambrosius Barth Publ., 1995.

[79] W. T. M. Verkley. A spectral model for two-dimensional incompressible fluid flow in a circular basin I. Mathematical formulation. *J. Comput. Phys.*, 136(1):100–114, 1997.

[80] T. von Petersdorff and C. Schwab. Numerical solution of parabolic equations in high dimensions. *M2AN Math. Model. Numer. Anal.*, 38(1):93–127, 2004.

[81] N. J. Walkington. Quadrature on simplices of arbitrary dimension. `http://www.math.cmu.edu/~nw0z/publications/00-CNA-023/023abs/`.

[82] H. R. Warner. Kinetic theory and rheology of dilute suspensions of finitely extendible dumbbells. *Ind. Eng. Chem. Fundamentals*, pages 379–387, 1972.

[83] S. Wollman and E. Ozizmir. A deterministic particle method for the Vlasov-Fokker-Planck equation in one dimension. *J. Comput. Appl. Math.*, 213(2):316–365, 2008.

[84] Q. Zhou and A. Akhavan. A comparison of FENE and FENE-P dumbbell and chain models in turbulent flow. *J. Non-Newtonian Fluid Mech.*, 109:115–155, 2003.

[85] O. C. Zienkiewicz, R. L. Taylor, and Zhu J. Z. *The Finite Element Method: Its basis and fundamentals*. Butterworth-Heinemann, 2005.

[86] B. H. Zimm. Dynamics of polymer molecules in dilute solution: viscoelasticity, flow birefringence and dielectric loss. *J. Chem. Phys.*, 24:269–278, 1956.