The Pennsylvania State University

The Graduate School

Department of Mathematics

# MODELLING AND SIMULATIONS OF NON-NEWTONIAN FLUID FLOWS

A Thesis in

Mathematics

by

Young-Ju Lee



Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2004

The thesis of Young-Ju Lee was reviewed and approved* by the following:

Jinchao Xu
Professor of Mathematics
Thesis Adviser
Chair of Committee

Chun Liu
Associate Professor of Mathematics

Andrew Belmonte
Associate Professor of Mathematics

James Vrentas
Dow Professor of Chemical Engineering

Nigel Higson
Distinguished Professor of Mathematics
Head of the Department of Mathematics

*Signatures are on file in the Graduate School.

# Abstract

We observe that various rate-type non-Newtonian constitutive equations can be recast into the well-known symmetric Riccati equations. From the careful study on the Riccati form of various models, some robust and stable discretizations of the rate-type models are constructed. Discrete analogue of energy estimates have been derived and confirm the stability of our new schemes. As a consequence of discrete energy estimates, a global existence and uniqueness of an approximate solution is established regardless of the size of the "Weissenberg number".

We discuss how to solve the resulting discrete problem based upon the preconditioned MINRES (Minimum Residual) method and fast and efficient solver has been constructed. Moreover, based on the fact that the efficient preconditioner should be constructed for the Laplace equation with pure Neumann boundary condition, a certain framework on the convergence analysis of the method of successive subspace corrections for singular problems has been provided in a Hilbert space setting.

Extensive numerical studies on a falling sphere through the Johnson-Segalman fluids are performed. The main motivation behind our numerical studies is from some belief that a range of parameters exhibiting a non-monotonic relation between the shear rate and the strain rate displayed for the steady shear flow of the Johnson-Segalman model might produces a continual and sustained oscillation of a falling sphere in a worm-like micellar fluid. In contrast to such a belief, our numerical experiments did not show a sustaining oscillation of a falling sphere, rather a transient oscillation. We then conclude

that a property of model like a non-monotonic shear stress-strain rate can not alone be used for explaining a continual oscillation of a falling sphere in a worm-like micellar fluid. Finally, we report some intriguing numerical experiments that show how the slip parameter "$a$" affects the pattern of oscillations of sphere and also a negative wake.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgments

I am most grateful and indebted to my thesis advisor, Prof. Jinchao Xu, for his advice, patience, encouragement and his continuous support during my time here at Penn State. I, especially, acknowledge the summer research support from Prof. Jinchao Xu in 2004 for preparation of this thesis.

I am also grateful to Prof. Liu for his insightful discussions, Prof. Belmonte for his guidance to the challenging but fascinating project and also Prof. Vrentas for his kindness to agree to serve on my thesis committee and for helpful discussions.

My special thanks go to Prof. Zikatanov and Dr. Sun for their valuable help and friendship. Thanks are also due to all CCMA (Center for Computational Mathematics and Applications) family for allowing me to run my computational works without any restrictions.

Last but not least, I wish to express my most sincere gratitude for Eun-Ju, Sung-Min and Hannah for their patience, continuous encouragement and smiles for me which led me to look up on sunny sides in every respect.

I dedicate this thesis to my late mother-in law, mother of Eun-Ju who blessed us but did not make it to my concluding day.

# Chapter 1

# Introduction

## 1.1 Background and Motivations

Fluids comprised of large macromolecules, known as viscoelastic fluids, can produce a great variety of new phenomena. The better known examples of this richness, among many others, include the rod climbing Weissenberg effect [23], die swell [22], extrusion instabilities [11] and the oscillation of falling sphere in a worm-like micellar fluid, [45]. Especially, the oscillation of a falling sphere in a worm-like micellar fluids is shown to be sustained and irregular and it is one of relatively recent and compelling examples of flow instability provided by worm-like micellar fluids, consisting of elongated flexible micro-structures formed in concentrated surfactant solutions [31]. Although careful rheological characterization of the test fluids studied and simulated have led to a greater understanding of a micellar fluid, there has not been much understanding on proper mathematical models that are responsible for such an oscillating phenomenon of a falling sphere.

Mathematical models governing viscoelastic fluids are in fact much more complex than those of traditional Newtonian fluid dynamics and it is widely acknowledged that numerical simulations play a crucial role in mathematical modelling for such experimental results and the role of numerical simulation has increased tremendously over the past

two decades. There have also been rapid advances in the development of numerical algorithms for simulating viscoelastic flows. However, the developments are hampered by various difficulties. One of most notable examples of such difficulties is the breakdown in convergence of the algorithms at critical values of the Weissenberg number, which was first observed in the late 1970's [66]. Since then it has been called the high Weissenberg number problem. Despite extensive efforts in the search for suitable methods, the high Weissenberg number problem still remains elusive. We would like to remark that it has been widely believed that such a problem is attributed to the lack of positivity preserving property of the so-called conformation tensor $\mathcal{C}$ on the discrete level. The conformation tensor $\mathcal{C}$, from the molecular theories, denotes the ensemble average of the dyadic product of end-to-end vector $\mathbf{Q}$ of the (Hookean) dumbbell (see e.g. [6]). The positive-definite (non-negative) character of $\mathcal{C}$ is then necessary for making meaningful microscopic interpretation and so it is known to give rise to a criteria for the physical admissibility of the models (see [6], p.281). Indeed, the positivity of the conformation tensor has been proved to be valid for various models, including UCM (upper convected Maxwell model) and the Oldroyd-B model. These models are known to be mathematically stable in the classical sense of Hadamard, namely the solution for such models depends continuously on the initial data (see e.g. [40], [46], [47] and [48]). Moreover, the stability analysis depends crucially on the fact that the conformation tensor $\mathcal{C}$ remains positive definite while it evolves in time.

The importance of keeping the positivity of the conformation tensor is further supported by another observation that the trace of the conformation tensor $\mathcal{C}$ can be

considered to be an elastic energy. Indeed, from the positive-definiteness of the conformation tensor, one can derive an energy law (see [59] for Oldroyd-B model and also chapter 3 in this thesis, where the discrete analogue of energy estimate has also been driven). Such a property is also valid for other models including Giesekus, Phan-Thien and Tanner, Leonov and Larson models (see [6] and [44] for detailed descriptions) for a wide range of parameters and the local loss of this property on the discrete level causes the onset of oscillation of solutions and leads to disastrous effects (see [6], [26], [43], [48] and [66]).

Clearly, it is significant to develop numerical schemes that preserve the positive definiteness of the conformation tensor. It is in fact believed that the aforementioned numerical difficulty can be overcome in time-dependent calculations with positivity preserving schemes (see Owens and Phillips (2002), p.197, [66] and a loss of the positivity has long been known to be due to the fact that the constitutive equation has not been discretized properly (see also [6], [26], [43] and [48]). Such improper discretization may result from a lack of understanding on the mathematical models. It is well-known that mathematical analysis for models describing complex fluids is quite challenging. We would like to refer readers to the following articles [20], [56], [58] and [57] for the state of the art mathematical analysis on certain viscoelastic models.

Apparently, finding positivity preserving discretization scheme is not an easy task. In a recent (2002) book "*Computational Rheology*" written by Owens and Phillips (see [66] p.59), authors noted that "although the continuous system possesses the property that $\mathcal{C}$ is positive definite, this may *not* be carried over to the corresponding discrete problem".

The only attempt known to the author for obtaining positivity preserving scheme can be found in a very recent (2003) work [59]. Their main idea was to set $\mathcal{C} = AA^T$ and then try to write down equations for $A$ approximately on the discrete level. Hence, the positivity of $\mathcal{C}$ is forced with such an approach. One may argue that this is an unnatural approach since, judging from the integral expression of $\mathcal{C}$, the artificially introduced $A$ has no apparent physical meaning. Moreover, this approach seems to be restricted to Oldroyd-B models and its extension to other models does not seem to be obvious and also the scheme is only first order accurate.

The additional difficulties that often arise from the numerical modelling viscoelastic fluids are encountered by the fact that the time-dependent models should be simulated as is the current situation where we are interested in the transient motion of a solid sphere in a worm-like micellar fluid. Explicit schemes are difficult to develop due to the incompressibility condition and even if it is possible, such schemes are in general not efficient due to the restrictive time step size. Hence, the development of fast solvers are crucial for efficient numerical modelling of complex fluids.

In the light of all the comments above, the following tasks are raised and discussed in my thesis work.

- Find the mathematical model that is responsible for the oscillation of falling sphere in a worm-like micellar fluid.
- Develop robust and stable discretizations that preserve the positive-definiteness of the conformation tensor.
- Construct fast and efficient algorithms to solve the resulting discrete system.

In this thesis, we accomplished the last two tasks completely and based on our numerical framework, we took an initial attempt to accomplish the first task. Following some belief, [4, 45] that a rheological property seen in the simple steady shear flow of micellar fluid, namely, non-monotonic shear stress-strain rate relation might be responsible for a continual and irregular oscillation of a falling sphere, we simulate a falling sphere in a fluid governed by a representative mathematical model, so-called the Johnson-Segalman model, [49] for which such a rheological property is well-known to be apparent. However, we could not obtain a sustained and continual oscillations from the Johnson-Segalman model and based on our extensive numerical experiments, we then conclude that a property of the Johnson-Segalman model showing a non-monotonic shear stress and the rate of strain relation for the steady shear flow can not alone be used to explain a continual oscillation observed in a worm-like micellar fluid.

## 1.2 Organization of the thesis

The thesis begins with the discussion of viscoelastic models. Especially, the chapter 2 is devoted to understand various viscoelastic models that are strongly coupled non-linear systems in a unified framework. The chapter 3 then discusses how to discretize these systems in robust and stable manners, presents the discrete analogue of energy estimates and proves the global existence and uniqueness of the discrete solution no matter how the size of "Weissenberg number" is big. By this, we resolve so-called the high Weissenberg number problem which has long been elusive among the non-Newtonian numerical practitioners.

The development of robust and efficient algorithms to solve the resulting discrete systems appears in the chapter 4. Starting with the observation that the robust discretization reduces the full non-linear system to Stokes-like system, which is symmetric. the preconditioned MINRES method is proposed to solve the resulting symmetric discrete system. Finally, an appropriate preconditioner is developed and rigorously analyzed.

In the chapter 5, extensive numerical reports on the falling sphere in the Johnson-Segalman fluid are presented based on the numerical machinery developed through chapters 3-4.

## 1.3 Principal results

The development of the unified and efficient numerical framework that can be used for numerical modelling of complex fluid flows has been proposed in this thesis.

In this section, we shall elaborate our principal results obtained in this thesis work in more details.

### 1.3.1 Robust discretization for the rate-type non-Newtonian models

The mathematical models governing complex fluids are much more complicated than the traditional Newtonian models. The models are strongly coupled non-linear PDEs. For example, the Johnson-Segalman model can be written as follows. The equations of motion and continuity for unsteady incompressible flow are

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) = \mathbf{b} - \nabla p + \mu_s \nabla \mathbf{u} + \mathrm{div}\tau, \tag{1.3.1}$$

$$\mathbf{divu} = 0, \tag{1.3.2}$$

where $\mathbf{b}$ is a forcing term and the Johnson-Segalman constitutive relation is given as follows

$$\tau + \lambda \frac{\delta_E \tau}{\delta_E t} = \mu_p (\nabla \mathbf{u} + \nabla \mathbf{u}^T), \tag{1.3.3}$$

where $\lambda$ is the characteristic relaxation time, $\mu_s$ and $\mu_p$ are the newtonian viscosity and the polymeric viscosity respectively and

$$\frac{\delta_E \tau}{\delta_E t} := \frac{D\tau}{Dt} - \left( \frac{a+1}{2} \nabla \mathbf{u} + \frac{a-1}{2} \nabla \mathbf{u}^T \right) \tau - \tau \left( \frac{a+1}{2} \nabla \mathbf{u} + \frac{a-1}{2} \nabla \mathbf{u}^T \right). \tag{1.3.4}$$

Here the parameter "$a$" is called the slip parameter and it will be of our main concern in the chapter 5.

The quantity $\dfrac{\delta_E}{\delta_E t}$ in (1.3.4) acts upon a symmetric tensor $\tau$ and is oftentimes called the Gordon-Schowalter derivative, [32]. One crucial problem for the aforementioned PDEs is in numerical approximations for the constitutive equation (1.3.3). Constitutive equations for most viscoelastic fluids including (1.3.3) are difficult to discretize and there is no known stable schemes. We believe that such a difficulty is due mainly to the fact that there has not been a guideline or right understanding on the models. Naive approaches such as an upwinding scheme may not lead to robust discretization schemes.

One distinct and core research result we obtained in the thesis is new discretization schemes which preserve some important property of the conformation tensors. Our guidance to discretize systems like (1.3.3) is based on the new observation that most rate-type models can be recast into the symmetric Riccati differential equations. This guideline can be used for a unified numerical algorithmic framework that can be used for simulating most existing constitutive equations in a way that the positivity of the

conformation tensor in continuous level can be naturally realized for its discrete counterpart. Also our numerical solutions do have the same energy estimates with its continuous counterparts. It is agreed that discrete realizations of properties of continuous solutions is quite challenging as mentioned by R. Keunings, [51] by the following : "Numerical methods should be faithful to the original mathematical model. I would thus find it crucial that mathematicians continue to explore the properties of the exact (and most probably forever unknown) solutions to the governing equations, a very difficult task indeed!".

The main observation on the similarity between the rate-type models and the Riccati equations can be elaborated by rewriting the rate-type models in terms of certain Lie derivative, into an "ordinary" differential equation. The Riccati equation arises in many fields of applied mathematics, engineering and economic sciences, especially in the domains such as, just to cite a few, linear optimal control and filtering problems with quadratic cost functionals, differential geometry and singular perturbation theory. Moreover, there are well-developed theory of symmetric (or Hermitian) Riccati equation. For basic theory of the matrix Riccati differential equations, we refer interested readers to the monograph of Reid (1972), [70]. For the state of the art of the theory of symmetric matrix Riccati equations, readers refer to the recent monograph by Abou-Kandil et al, [1].

Using a semi-Lagrangian approach and finite element method, we obtain a class of positivity preserving discretization schemes whose accuracies are of up to second order both in space and in time under some mild conditions.

The stability of our schemes is then demonstrated by the discrete analogue of energy estimates. We would like to remark that the importance of the positivity preserving has not been quantitatively shown so far. Moreover, based on the discrete energy estimates, we show the global existence and uniqueness of the discrete solutions for all models having energy estimate. This includes especially, the Oldroyd-B model. By this, we confirm and extend the common belief that there will not be the high Weissenberg number problem for such a model.

### 1.3.2 Efficient iterative techniques: multigrid method and preconditioned MINRES method

The robust discretizations described briefly in the previous section lead us to solve the following form of discrete problem (a non-dimensional form):

$$\begin{pmatrix} \frac{\mathrm{Re}}{k} I - \eta_s \Delta_h & \nabla \\ -\mathbf{div} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_h \\ p_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ \mathbf{g}_h \end{pmatrix}, \tag{1.3.5}$$

where Re is the Reynolds number, $k$ is the time step size and $\mathbf{f}_h$ and $\mathbf{g}_h$ are source terms respectively. $\mathbf{g}_h$ is in general zero due to the incompressibility condition, $\mathbf{divu} = 0$.

One can immediately notice that the system (1.3.5) is symmetric and some Krylov space method like MINRES (Minimum Residual Method) can be used. But a plain application of such a method is not so efficient since the convergence rate is badly dependent on the various parameters such as $k$, Re and the mesh size $h$.

To obtain fast and efficient techniques to solve the problem like (1.3.5), it is crucial to apply a robust preconditioner. In the chapter 4, we introduce a preconditioned MINRES method and its algorithmic details. We then study how to construct a robust preconditioner for the system of equations (1.3.5) to obtain efficient iterative techniques. Especially, we observe that one needs to precondition the Laplace operator with pure Neumann boundary condition. Motivated by this observation, we analyze the convergence rate of the method of successive subspace corrections including the multigrid method applied for general singular problems and obtained the sharpest possible convergence rate in a Hilbert space setting.

By the aforementioned results, we resolve non-trivial issues arising in simulating non-Newtonian fluid simulations. These advanced techniques are believed to allow us to attack various models without much efforts. In other words, we have developed some numerical framework that can be used to model various phenomena in the viscoelastic fluids.

### 1.3.3 Numerical studies of the falling sphere through a cylinder

The thesis focuses on finding methods for handling complex problems which is readily applicable in reality. The constructed machinery can be applied to find the right models showing some interesting phenomenon like the falling sphere experiments in worm-like micellar fluids.

We apply our powerful numerical framework against the falling sphere simulations in an attempt to identify the right model for a continual oscillation. The test model is the Johnson-Segalman model. In our attempts, we extract various parameters in the

model that show the non-monotonic shear stress-strain rate for the steady shear flow of the model. Using such parameters, we simulate a falling sphere in the Johnson-Segalman model. From the extensive numerical experiments, we conclude that such a non-monotonic property that the Johnson-Segalman model possesses may not alone be used to explain the continual and sustained oscillation of a falling sphere in a worm-like micellar fluid.

We also obtained various new intriguing numerical results, most of which are related to the slip parameter "$a$". Namely, we see that some relation between the pattern of oscillation of a falling sphere or the formation of negative wake and the slip parameter "$a$".

## 1.4 Some remarks and Future works

Several interesting unresolved problems are in order.

- Development of Numerical Packages
- Adaptivity
- Numerical Error Analysis

The current setting of numerical experiments are two space dimensional. Especially, exploiting the fact that the flow is symmetric, we use axisymmetric formulations of our models (see Appendix A). However, our numerical framework has been constructed for both two and three space dimensions. Such an extension in the side of implementations is not much difficult. We believe such simulations shall dramatically distinguish our simulations from other available works and lead us to look fully three dimensional

phenomenon in the fluid flows. Moreover, our new frameworks are applicable to a wide range of macroscopic models, which seems to cast a light into some sort of unifying theory of macroscopic models. It is the goal to extend our restricted numerical framework to fully generalized situations.

We shall also further our search for right models responsible for the continual oscillation of a falling sphere. There are several models available in literatures which are believed to predict the oscillations of a falling sphere. These include White-Metzner model [87] and FENE-PM models [86]. Some relevant studies are provided especially for the FENE-PM model in the Appendix C. The aforementioned numerical tools are believed to provide systematic and efficient way of modelling of complex flows.

Our schemes are mainly based upon the semi-Lagrangian framework and in our point of view, what is crucial to the success of a semi-Lagrangian scheme is to use proper grid adaptation techniques. The basic reason is because the semi-Lagrangian scheme is based on discretizing the total derivative $\frac{D}{Dt} = \frac{\partial}{\partial t} + (\mathbf{u} \cdot \nabla)$ along the particle trajectory, whose characteristic feet is not in general located on the grid point, so the numerical solution is not available to use. The introduction of interpolation errors is then unavoidable. Another reason to use the grid adaptation is to capture commonly observed stress boundary layer in non-Newtonian flow simulations. To reduce the interpolation error and better capture sharp boundary layers, we are in the process of developing and implementing some special grid adaptation techniques in our simulations. Along this line, we have succeeded 4 to 1 contraction flows simulation using the adaptivity techniques currently although it is not reported in the thesis.

Finally, we shall take a challenging step for the numerical analysis of our new schemes. Especially, the error analysis should be performed to prove our scheme is indeed superior to other existing schemes. We hope that both energy estimates in chapter 3 and the newly observed relation between the Riccati equations and the rate-type models in chapter 2 will guide us to accomplish this task.

# Chapter 2

# Viscoelastic models

## 2.1 Introduction

The simplest constitutive equations suitable for modelling the behavior of dilute polymeric solutions under general flow conditions are those of the Oldroyd type. In particular, the Oldroyd B model, derived by Oldroyd, [65] in 1950, is an empirical expression generalizing the linear viscoelastic equation by writing the stress and strain relation in tensorial form and satisfying certain admissibility criteria, [9, 65]. This model is so crude that it can not capture many features of real rheological fluids. More sophisticated model are then developed such as FENE dumbbell model and their approximations FENE-P and FENE-CR models. There are also other various differential macroscopic models, [84, 54, 46, 9].

This chapter is devoted to reformulate the aforementioned various differential models in terms of the conformation tensor $\mathcal{C}$ and also provide an analytical solution for the conformation $\mathcal{C}$ tensor in terms of velocity gradients. Especially, the reformulation of the various models with respect to the conformation tensor $\mathcal{C}$ shall present itself as the symmetric matrix Riccati equations. The main ingredient for such a reformulation is to search for the conformation tensor for a given model. Constitutive equations for the viscoelastic models involve the time derivative because the stress is determined by the entire deformation history of the fluid, [66]. Namely, the fluid has a memory and

the stress is not instantaneously related to the rate of strain. There are various time derivatives such as the material time derivative, the upper convected time derivative and the Gordon-Schowalter derivative. Detailed description of such derivatives shall be presented in this chapter, see (2.1.1), (2.1.5) and (2.4.1). Moreover, we shall see that these derivatives can be viewed in general as the Lie derivative and their particular forms shall be determined by the choice of deformation tensors.

Especially, in this chapter, we shall take the Gordon-Schowalter derivative (2.4.1) as the most general objective time derivative. This includes the upper convected time derivative and the lower time derivative in its special form.

A simple but important observation to search for a conformation tensor for any given model is that a model with the upper convected time derivative has the conformation tensor $\mathcal{C} = \tau + \dfrac{\mu_p}{\mathrm{We}} I$ and in general a model with the Gordon-Schowalter derivative has the conformation tensor $\mathcal{C} = \tau + \dfrac{\mu_p}{a\mathrm{We}} I$. The reformulation shall be performed in terms of such a conformation tensor corresponding to the given objective derivative in the model. Following [66], we shall denote $\mathcal{C}$, the conformation tensor by $\tau_A$.

As mentioned in the chapter 1, the positive definiteness of the conformation tensor is for the physical admissibility [6]. There are also many literatures (see [66] and references cited therein) emphasizing the importance of keeping the positive definiteness of the conformation tensor in the discrete sense. We shall also show that the reformulated model can be solved analytically for the conformation tensor and prove that the conformation tensor is indeed positive definite in time evolution. Moreover, the analytic

expression of the conformation tensor shall clarify the relation between the integral models and the differential models. We would like to remark that some differential models are not known to possesses their integral counterparts, [46].

The main tool to solve the model for the conformation tensor is to rewrite the objective derivative in terms of the Lie derivative and make the equation as a simple ordinary differential equation.

We believe that this chapter sheds a light into a unified way of viewing various rate-type models.

We begin this section with a basic review on Kinematics of bodies.

### 2.1.1 Deformation tensor

To describe a particle moving in a fluid, we introduce two configurations, say $\Omega_t$ and $\Omega_s$. The motion of a particle is then manifested by the following mapping :

$$y : \Omega_s \mapsto \Omega_t.$$

Let us denote $y(X, t, s)$ the position of particle $X$ at time $s$ and $y(X, t, t) = X$.

We shall now introduce a set of notation for ease of our presentation throughout this paper. For any given tensor $\sigma$, $\sigma(y(X, t, s), s)$ shall be denoted by $\sigma(t, s)$ or $\sigma_d(s)$ and $\sigma(X, t)$ by $\sigma(t)$. The same notation also apply to any vector $\mathbf{u}$.

The velocity of the particle is given by $\mathbf{u} = \dot{y}$, where the dot indicates the partial derivative with respect to $s$ with $t$ fixed. In the Eulerian description $(y, s)$, the chain

rule gives the familiar material derivative defined as follows :

$$\frac{Dg}{Ds} = \frac{\partial g}{\partial s} = g_s + (\mathbf{u} \cdot \nabla)g, \tag{2.1.1}$$

where $\nabla$ is the gradient for the $y$ variables. Classical mechanics assumes that $y : \Omega_s \mapsto \Omega_t$ is a diffeomorphism and the relative deformation gradient is the matrix defined by (see [58], [66] or [71])

$$F_{ij}(t,s) = \frac{\partial y_i(X,t,s)}{\partial X_j}.$$

We note that an application of the chain rule gives an Eulerian description,

$$\begin{aligned} \frac{DF(t,s)}{Ds} &= \dot{F}(t,s) = \left( \frac{\partial}{\partial s} \frac{\partial y_i}{\partial X_j}(X,t,s) \right) \\ &= \left( \frac{\partial u_i}{\partial X_j}(X,t,s) \right) = \left( \frac{\partial u_i}{\partial y_k} \frac{\partial y_k}{\partial X_j}(X,t,s) \right) = \nabla \mathbf{u}(t,s) F(t,s). \end{aligned}$$

Our convention for the gradient of a vector $\mathbf{u}$ is that the $(i,j)$ component of $\nabla \mathbf{u}(t,s)$ is $\partial u_i / \partial y_j$, where $\mathbf{u} = (u_i)_{i=1}^d$ with $d = 2$ or 3. The inverse of $F$ is often called the displacement gradient tensor. From the relation that $FF^{-1} = I$, we obtain

$$\frac{DF^{-1}(t,s)}{Ds} = -F^{-1}(t,s) \nabla \mathbf{u}(t,s). \tag{2.1.2}$$

We note that the following relations hold true:

$$F^{-1}(t,s) = F(s,t) \tag{2.1.3}$$

and

$$\frac{DF(s,t)}{Dt} = \nabla \mathbf{u}(t) F(s,t). \tag{2.1.4}$$

### 2.1.2 Two identities on upper convective derivatives

In this subsection, we shall introduce two main identities related to the upper convective time derivative of tensors. The upper convected derivative denoted by $\frac{\delta_F}{\delta_F t}$ for the tensor $\zeta$ is defined through :

$$\frac{\delta_F \zeta}{\delta_F t} = \frac{\partial \zeta}{\partial t} + (\mathbf{u} \cdot \nabla)\zeta - \nabla \mathbf{u} \zeta - \zeta \nabla \mathbf{u}^T. \tag{2.1.5}$$

We introduce first identity for the upper convected derivative as follows: for any $(X,t) \in \Omega \times (0,\infty)$,

$$\frac{\delta_F \zeta}{\delta_F t}(X,t) = \lim_{s \to t} F(t,s) \frac{D\left(F(s,t)\zeta(t,s)F(s,t)^T\right)}{Ds} F(t,s)^T. \tag{2.1.6}$$

This can be shown by a straightforward calculation based on the facts (2.1.2) and (2.1.3). This is often called the Lie derivative of $\zeta$ and also known as the Truesdell stress rate (see Simo and Hughes (1998),p.254, [80]).

The second identity is on the upper convective derivative of the identity tensor (see e.g. [65]), namely

$$\frac{\delta_F I}{\delta_F t} = -2\mathcal{D}(\mathbf{u}). \tag{2.1.7}$$

This identity can be obtained by simply setting $\zeta = I$ in (2.1.5) and is crucial when we reformulate the constitutive equation in terms of the conformation tensor. It can also be viewed as a tool for approximating the rate of strain $\mathcal{D}(\mathbf{u})$.

The identity (2.1.6) has been used for developing objective time-stepping algorithm commonly called *incrementally objective discretization*, a nomenclature first introduced in Hughes and Winget (1980), [42]. Since then it has also been used for simulating some non-Newtonian models by Baaijens (1993), [3]. In their works, the direct discretization of the upper convected time derivative (2.1.5) has been performed along the particle trajectory based on the Lagrangian framework and the approximation for the rate of strain has been made based on the following identity:

$$\lim_{s \to t} F(s,t)^T \frac{DC(t,s)}{Ds} F(s,t) = 2\mathcal{D}(\mathbf{u})(t) \tag{2.1.8}$$

where $C$ is called the *Cauchy* strain tensor defined through:

$$C(y(X,t,s),s) = F^T(t,s)F(t,s)$$

Our approach is subtle but fundamentally different from their algorithms from the following two aspects. Namely, we shall use the so-called semi-Lagrangian framework (see e.g. [68]) for the time discretization and shall use (2.1.7) to approximate the rate of strain. Indeed, the use of (2.1.7) is instrumental and necessary in that it allows us to view various constitutive models (e.g. the Oldroyd-B model) as the Riccati differential equation and leads to the positivity preserving scheme. More detailed time discretization

scheme shall be described in the chapter 3 and we will see that the use of (2.1.8) may not lead to the positivity preserving discretization.

In the following section, to illustrate the basic form of viscoelastic models, the Oldroyd-B model is introduced. We then present its reformulated version by using the identity (2.1.7).

## 2.2 An illustration: the Oldroyd-B model

In this section, we shall give a brief description of the Oldroyd-B model [65] as an illustrative example for viscoelastic models.

### 2.2.1 The Oldroyd-B model

Let us consider the flow of the Oldroyd-B fluid occupying a bounded domain $\Omega \subset \mathbb{R}^d$. The equations of motion for unsteady incompressible flows are

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) = -\nabla p + \mathrm{div}\sigma,$$

$$\mathbf{div}\mathbf{u} = 0,$$

respectively, where $\mathbf{u}$ is the velocity field, $p$ is the isotropic pressure, $\sigma$ is the extra-stress tensor and $\rho$ is the density of the fluid. The extra-stress tensor is related to the rate-of-strain tensor $\dot{\gamma} = \nabla\mathbf{u} + (\nabla\mathbf{u})^T$ by the following constitutive equation

$$\sigma + \lambda_1 \frac{\delta_F \sigma}{\delta_F t} = \mu\left(\dot{\gamma} + \lambda_2 \frac{\delta_F \dot{\gamma}}{\delta_F t}\right) \tag{2.2.1}$$

where $\lambda_1$ and $\lambda_2$ are characteristic relaxation and retardation times of the fluid, respectively, $\mu$ is the constant shear viscosity. Furthermore, if the extra-stress tensor is expressed in terms of its solvent and polymeric contributions, $\tau$

$$\sigma = \mu_s \dot{\gamma} + \tau,$$

then the polymeric contribution of stress, $\tau$ satisfies

$$\tau + \lambda_1 \frac{\delta_F \tau}{\delta_F t} = \mu_p \mathcal{D}(\mathbf{u}).$$

The constants, $\mu_s$ and $\mu_p$ are the solvent and polymeric viscosities, respectively, where $\mu = \mu_s + \mu_p$, and

$$\eta_s = \frac{\lambda_2}{\lambda_1}\mu \quad \text{and} \quad \eta_p = \left(1 - \frac{\lambda_2}{\lambda_1}\right)\mu.$$

In dimensionless form, the governing equations can be given by

$$\mathrm{Re}\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) = -\nabla p + \mathrm{div}\tau + \eta_s \mathrm{div}\mathcal{D}(\mathbf{u}) \tag{2.2.2}$$

$$\mathbf{divu} = 0 \tag{2.2.3}$$

$$\tau + \mathrm{We}\frac{\delta_F \tau}{\delta_F t} \quad = \quad 2(1-\eta_s)\mathcal{D}(\mathbf{u}) = 2\eta_p \mathcal{D}(\mathbf{u}), \tag{2.2.4}$$

where $\mathcal{D}(\mathbf{u}) = (\nabla\mathbf{u} + \nabla\mathbf{u}^T)/2$ and

$$\beta = \frac{\lambda_1}{\lambda_2}, \quad \mathrm{Re} = \frac{\rho U L}{\mu} \quad \text{and} \qquad \mathrm{We} = \frac{\lambda_1 U}{L}.$$

In case $\eta_s = 0$, UCM (the upper convected Maxwell) model results and it is well-known that the model (2.2.1) or (2.2.4) is the simplest non-linear extensions of Maxwell's idea of formulating a system of ordinary differential equations which determines the stress in terms of the velocity gradient. As mentioned earlier, the equations (2.2.2),(4.4.1) and (2.2.4) are well-known to be stable in the sense of Hadamard. (see e.g. [66])

### 2.2.2 A Reformulation of the Oldroyd-B model using the conformation tensor

One crucial technique used in this paper is based on the reformulation of the models in terms of the conformation tensor. The aforementioned model, the Oldroyd-B model has a property that the following tensor is positive-definite:

$$\tau_A(X,t) = \tau(X,t) + \frac{\eta_p}{\mathrm{We}} I.$$

Thanks to (2.1.7), the constitutive equation (2.2.4) can be rewritten as follows:

$$\tau_A + \mathrm{We}\frac{\delta_F \tau_A}{\delta_F t} = \frac{\eta_p}{\mathrm{We}} I. \tag{2.2.5}$$

We also note that, for the Oldroyd-B model, it is well-known that the conformation tensor $\tau_A$ can be written as the following integral form.

$$\tau_A(X,t) = \int_{-\infty}^{t} \frac{\eta_p}{\mathrm{We}^2} \exp\left(\frac{-(t-s)}{\mathrm{We}}\right) F(s,t)F(s,t)^T ds, \tag{2.2.6}$$

From the integral expression for $\tau_A$, it is immediate to see that $\tau_A$ is positive definite while the rate type model (2.2.5) does not show such a property immediately.

The positive-definiteness of $\tau_A$ appeared to be first established by Hulsen (see [44]) directly from the models such as (2.2.5) under certain conditions. This shall be revisited in §2.4 in terms of our new framework and shall be proved in a simpler way. Even though the model (2.2.6) is known to be equivalent to the rate-type model (2.2.5), it does not seem to have been fully clarified. Namely, the equivalent relation has been claimed by showing that the rate-type model (2.2.5) can be deduced from the integral model (2.2.6) without showing the other direction. So, if the integral model is not available *a priori*, the equivalent integral model for the corresponding rate-type model is not considered to be known (see e.g. Joseph (1990, p.15), [46]). Indeed, to obtain the integral model from its rate-type model, one needs to find out the analytic solution to the rate-type model. It also seems to have been missing (see Renardy (2000, p.18),[71]). We shall elaborate this issue and derive various integral models from their rate-type models in §2.4.

## 2.3 Generalized Riccati equations in terms of Lie derivatives

In this section, we shall prepare for our new framework to construct positivity preserving schemes by introducing a generalized Riccati differential equation in terms of a general Lie derivative.

The classic symmetric Riccati differential equation that we are interested in is of the following form

$$\frac{d\mathcal{C}(t)}{dt} = A(t)\mathcal{C} + \mathcal{C}A(t)^T - \mathcal{C}B(t)\mathcal{C} + U(t), \tag{2.3.1}$$

with a symmetric positive semidefinite initial condition $\mathcal{C}(0) = \mathcal{C}_0$.

This type of equations have been well studied in the literatures. Among others, there are two important properties of Riccati differential equation that are interesting to us in the current work. First of all, this equation has a certain closed-form solution, from which the solution $\mathcal{C}$ can be proved to be symmetric positive definite under certain conditions (see the Proposition (2.3.1) below). Secondly, the positivity preserving schemes for such equations are easily devised, especially in time (see [24] and §3.2.3 in the chapter 3 below).

### 2.3.1 A general Lie derivative

The Riccati equation (2.3.1) will be generalized by replacing the ordinary derivative $\frac{d}{dt}$ by a general Lie derivative defined as follows:

**Definition 2.3.1.**

$$\frac{\delta_L \tau}{\delta_L t} = \lim_{s \to t} L(t,s) \frac{D(L(s,t)\tau(t,s)L(s,t)^T)}{Ds} L(t,s)^T, \tag{2.3.2}$$

*where $L(s,t)$ is a smooth tensor.*

The tensor $L$ can be viewed as a transformation rule and in general satisfy the following ordinary differential equation.

$$\frac{DL(s,t)}{Dt} = R(t)L(s,t), \quad L(s,s) = I. \tag{2.3.3}$$

See for example (2.1.4), the case for the upper convected derivative, where $L = F$ and $R(t) = \nabla \mathbf{u}(t)$.

The tensor $L$ determined by (2.3.3) is called *the transition matrix or (evolution matrix)* in the community concerning the Riccati differential equation (see [1], p.2). However, it is not known to the authors that the Riccati differential equation has ever been studied in terms of the Lie derivative.

Note that the choice of $L$ shall change the rate of stress. The possible choices for $L$ useful for our exposition in this paper are listed as follows:

$$L(t,s) = \begin{cases} I, & R(t) = 0 & \text{(material)}, \\ F(t,s), & R(t) = \nabla \mathbf{u}(t) & \text{(upper convected)}, \\ F(s,t), & R(t) = -\nabla \mathbf{u}^T(t) & \text{(lower convected)}, \\ E(t,s), & R(t) = \dfrac{a+1}{2}\nabla \mathbf{u}(t) + \dfrac{a-1}{2}\nabla \mathbf{u}^T(t) & \text{(Gordon-Schowalter)}. \end{cases}$$

A rate defined by the following relation is known to be objective (see e.g. [3] p. 1119)

$$\frac{\delta_L \tau}{\delta_L t} = \frac{D\tau}{Dt} - (\omega + H)\tau - \tau(\omega + H)^T, \tag{2.3.4}$$

where $\omega = \dfrac{\nabla \mathbf{u} - \nabla \mathbf{u}^T}{2}$, $H$ is some objective tensor (see e.g. [41] for the definition of the objectivity). The rate defined by (2.3.4) can be cast into the Lie derivative with the transition tensor $L(s,t)$ satisfying the following ODE

$$\frac{DL(s,t)}{Dt} = (\omega + H)L(s,t), \quad L(s,s) = I.$$

With some appropriate choice of $H$, we can show that three cases except for $L = I$ are objective rates. We can also consider some other objective stress rates available in any literatures. Indeed, it has been addressed especially in Hughes (1984), [41], Simo and Hughes (1998) that any possible objective stress rate is a particular case of a fundamental geometric object known as the Lie derivative and moreover under the tensor $L$, a transformation rule, a somewhat complicated expression e.g. (2.1.5) becomes a rather simple time derivative (2.1.6).

### 2.3.2 A generalized Riccati equation

A generalized Riccati differential equations in terms of the above general Lie derivative is as follows:

$$\frac{\delta_L \mathcal{C}(t)}{\delta_L t} = A(t)\mathcal{C}(t) + \mathcal{C}(t)A(t)^T - \mathcal{C}(t)B(t)\mathcal{C}(t) + U(t). \tag{2.3.5}$$

In this formulation, we shall assume that the coefficient matrices are bounded and piecewise continuous and that the matrices $B$ and $U$ are symmetric and positive semidefinite.

Let us first consider a special case of the equation (2.3.5).

**Lemma 2.3.1.** *The solution to the following ordinary differential equation written in terms of a general Lie derivative*

$$\frac{\delta_L \mathcal{C}(t)}{\delta_L t} = U(t) \tag{2.3.6}$$

*can be given in the following closed-form :*

$$\mathcal{C}(t) = L(s,t)\mathcal{C}(s)L(s,t)^T + \int_s^t L(\nu,t)U(\nu)L(\nu,t)^T d\nu.$$

*Proof.* The ordinary differential equation (2.3.6) can be rewritten as followings:

$$\lim_{s\to t} L(t,s)\frac{D(L(s,t)\mathcal{C}(s)L(s,t)^T)}{Ds}L(t,s)^T = U(t).$$

This expression can then be cast into the following form :

$$\lim_{s\to t} \frac{D(L(s,t)\mathcal{C}(s)L(s,t)^T)}{Ds} = \lim_{s\to t} L(s,t)U(s)L(s,t).$$

To distinguish the fixed variable in the above expression, we shall denote $t$ by $\hat{t}$ for now. We then have the following ordinary differential equation for $L(t,\hat{t})U(t)L(t,\hat{t})$

$$\frac{D(L(t,\hat{t})\mathcal{C}(t)L(t,\hat{t})^T)}{Dt} = L(t,\hat{t})U(t)L(t,\hat{t})^T.$$

Taking integration and changing $\hat{t}$ back to $t$, we obtain the desired results. □

**Proposition 2.3.1.** *The solution of* (2.3.5) *exists and it is symmetric and nonnegative for all* $t \geq 0$. *Further, if* $\mathcal{C}(s)$ *or* $U(s)$ *is positive for some* $s \geq 0$, *then* $\mathcal{C}(t)$ *is positive for all* $t > s$.

*Proof.* We begin with a reformulation of the general Riccati equation (2.3.5). Notice

$$\frac{\delta_L \mathcal{C}(t)}{\delta_L t} = \frac{D\mathcal{C}(t)}{Dt} - R(t)\mathcal{C}(t) - \mathcal{C}(t)R(t)^T$$

Now by a simple reformulation, we obtain that

$$\begin{aligned} \frac{D\mathcal{C}(t)}{Dt} &= \left(A(t) + R(t) - \frac{1}{2}\mathcal{C}(t)B(t)\right)\mathcal{C}(t) \\ &+ \mathcal{C}(t)\left(A(t) + R(t) - \frac{1}{2}\mathcal{C}(t)B(t)\right)^T + U(t). \end{aligned} \tag{2.3.7}$$

Let us set

$$G(t) = A(t) + R(t) - \frac{1}{2}\mathcal{C}(t)B(t)$$

and rewrite (2.3.7) as followings:

$$\frac{\delta_\Phi \mathcal{C}(t)}{\delta_\Phi t} = U(t),$$

where

$$\frac{D\Phi(s,t)}{Dt} = G(t)\Phi(s,t), \quad \Phi(\nu,\nu) = I. \tag{2.3.8}$$

By the Lemma 2.3.1, we obtain that

$$\mathcal{C}(t) = \Phi(s,t)\mathcal{C}(s)\Phi(s,t)^T + \int_s^t \Phi(\nu,t)U(\nu)\Phi(\nu,t)^T d\nu. \tag{2.3.9}$$

Since $\Phi(s,t)$ is non-singular for all $s$ and $t$, the statement of the proposition follows as long as the solution exists. However, from the fact that $B(t)$ is semi positive-definite, we can deduce that with $\|\mathcal{C}(t)\| = \sup_{\|\chi\|=1} \chi^T \mathcal{C}(t)\chi$,

$$\|\mathcal{C}(t)\| \leq \|\mathcal{C}(0)\| + \int_0^t \left(2\|A(\nu) + R(\nu)\|\|\mathcal{C}(\nu)\| + \|U(\nu)\|\right) d\nu,$$

from which it follows by the Gronwall's inequality that $\|\mathcal{C}(t)\|$ is finite and hence $\mathcal{C}(t)$ exists for all $t > 0$. This completes the proof. □

## 2.4 Reformulating constitutive equations as generalized Riccati equations

In this section, we shall show that various interesting constitutive equations can be reformulated into the generalized Riccati equations (2.3.5) based on the Lagrangian frame.

Note that the reformulation will be made in terms of the conformation tensor denoted by $\tau_A$ (with abuse of notation) and $\tau_A$ is determined by the rate used in the model. Especially, the objective rate of our interest here is

$$\frac{\delta_E \tau_A}{\delta_E t} = \frac{D\tau_A}{Dt} - R(t)\tau_A - \tau_A R(t)^T, \tag{2.4.1}$$

where

$$R(t) = \frac{a+1}{2}\nabla \mathbf{u}(t) + \frac{a-1}{2}\nabla \mathbf{u}(t)^T. \tag{2.4.2}$$

In this case, the tensor $\tau_A$ will be of the following form:

$$\tau_A = \tau + \frac{\eta_p}{a\mathrm{We}} I. \tag{2.4.3}$$

For models with the upper convected derivative, the conformation tensor $\tau_A$ is given with $a = 1$ and with $a = -1$ in case the lower convected derivative is used.

The objective rate (2.4.1) is often called the Gordon-Schowalter derivative [32]. This can be written as a Lie derivative (2.3.2) with the transition matrix $E(s,t)$ satisfying the following ordinary differential equation:

$$\frac{DE(s,t)}{Dt} = R(t)E(s,t), \quad E(s,s) = I. \tag{2.4.4}$$

The tensor $E(s,t)$ obeying (2.4.4) has been first introduced by Johnson and Segalman (1977), [49] as a deformation tensor for viscoelastic fluids having non-affine histories.

In this section, first we shall study some interesting properties of the following Riccati equation:

$$\frac{\delta_E \tau_A}{\delta_E t} = -\alpha \tau_A + \beta I \tag{2.4.5}$$

where $\alpha, \beta$ may depend on $\tau_A$. Second, we shall illustrate how various rate-type models including multi-mode FENE-PM model can be cast into the aforementioned special Riccati equations (2.4.5) (see Table 2.4.1).

In the end of this section, we shall extend this idea to some general single variable models with the property of the positive definiteness and present their reformulations into another type of Riccati equations.

Let us begin with an explicit solution of the Riccati equation (2.4.5).

**Lemma 2.4.1.** *The solution to* (2.4.5) *satisfies*

$$\begin{aligned} \tau_A(t) &= \exp\left(-\int_s^t \alpha(\varsigma)d\varsigma\right) E(s,t)\tau_A(t,s)E(s,t)^T \\ &+ \int_s^t \exp\left(-\int_\nu^t \alpha(\varsigma)d\varsigma\right)\beta(\nu)E(\nu,t)E(\nu,t)^T d\nu \end{aligned} \tag{2.4.6}$$

*Proof.* By a argument similar to what was used in the proposition (2.3.1), we can solve the equation (2.4.5) to obtain

$$\tau_A(x,t) = \Phi(s,t)\tau_A(t,s)\Phi(s,t)^T + \int_s^t \beta(\nu)\Phi(\nu,t)\Phi(\nu,t)^T d\nu, \tag{2.4.7}$$

where

$$\frac{D\Phi(s,t)}{Dt} = \left(\frac{a+1}{2}\nabla\mathbf{u}(t) + \frac{a-1}{2}\nabla\mathbf{u}(t)^T - \frac{\alpha(t)}{2}I\right)\Phi(s,t), \quad \Phi(s,s) = I.$$

Now we shall show that $\Phi(s,t)$ can be expressed by the following form :

$$\Phi(s,t) = \exp\left(-\int_s^t \frac{\alpha(\nu)}{2}d\nu\right)E(s,t). \tag{2.4.8}$$

To see this, note that $\Phi_1(s,t) = E(s,t)$ is the solution to the following ordinary differential equation :

$$\frac{D\Phi_1(s,t)}{Dt} = \left(\frac{a+1}{2}\nabla\mathbf{u}(t) + \frac{a-1}{2}\nabla\mathbf{u}(t)^T\right)\Phi_1(s,t)$$

and the solution to the equation

$$\frac{D\Phi_2(s,t)}{Dt} = -\frac{\alpha(t)}{2}\Phi_2(s,t)$$

is given by

$$\Phi_2(s,t) = \exp\left(-\int_s^t \frac{\alpha(\nu)}{2}d\nu\right) I. \tag{2.4.9}$$

A simple observation that $\Phi(s,t) = \Phi_1(s,t)\Phi_2(s,t)$ completes the proof. □

We shall now derive $\alpha$ and $\beta$ corresponding to some interesting models such as the Oldroyd-B, the Johnson-Segamal, the Phan-Thien and Tanner, and the FENE-PM respectively. Let us first summarize the result in the following table.

Table 2.4.1. Lists of Models

| Model | $\alpha(t)$ | $\beta(t)$ |
|---|---|---|
| Oldroyd-B ($a = 1$) | $1/\mathrm{We}$ | $\eta_p/\mathrm{We}^2$ |
| Johnson-Segalman | $1/\mathrm{We}$ | $\eta_p/a\mathrm{We}^2$ |
| Phan-Thien and Tanner | $u/\mathrm{We}$ | $\eta_p/a\mathrm{We}^2$ |
| FENE-PM | $\dfrac{g}{\mathrm{We}_j} - \dfrac{D\ln g}{Dt}$ | $\dfrac{\eta_p}{\mathrm{We}_j^2}g$ |

Here $u$ and $g$ are scalar functions given as (2.4.12) and (2.4.13) below respectively. Some interesting results that can be deduced from the result (2.4.6) of the Lemma 2.4.1 and the Table 2.4.1 are in order. First, if $\alpha \geq c$ for some positive constant $c$, under the assumption that the transition matrix $E(s,t)$ is bounded for $s \leq t$, we obtain formally the following integral models by taking $s \to -\infty$,

$$\tau_A(t) = \int_{-\infty}^{t} \exp\left(-\int_{\nu}^{t} \frac{\alpha(\nu)}{2} d\nu\right) \beta(\nu) E(\nu,t) E(\nu,t)^T d\nu. \tag{2.4.10}$$

Especially, this includes the Johnson-Segalman integral model, which does not seem to be known before (see Joseph (1990), p.15, [46]). Second, the expression for $\Phi(s,t)$ in (2.4.8) is quite useful in the computational viewpoint. Namely, combining an approximation for $E(s,t)$ that is ubiquitous in viscoelastic models with an approximation of (2.4.9), we can handle various non-linear models without much extra efforts.

Now, we shall illustrate the procedure of reformulations by taking several interesting models. Some of constitutive equations that are interesting to us are originally given as follows:

$$u\tau + \mathrm{We}\frac{\delta_E \tau}{\delta_E t} = 2\eta_p \mathcal{D}. \tag{2.4.11}$$

Here the function $u$ is defined through

$$u = \exp\left(\frac{\varepsilon \mathrm{We}}{\eta_p} tr(\tau)\right), \tag{2.4.12}$$

where $\varepsilon$ is a parameter. This corresponds to so-called the Phan-Thien and Tanner model, [84]. If $\varepsilon = 0$ or $u = 1$, then the Jonson-Segalman model, [49] results and if further $E = F$ and $a = 1$, the Oldroyd-B model (2.2.4) results.

Note that the equation (2.4.11) relates the stress and the rate of strain. The reformulation shall hide this relation. Observing the relation (2.4.3) and simple change of variables, we obtain the following equations for $\tau_A$ :

$$\mathrm{We}\frac{\delta_E \tau_A}{\delta_E t} = -u\tau_A + \frac{\eta_p}{a\mathrm{We}} uI.$$

We then obtain the expressions listed in the Table (2.4.1) except for the FENE-PM model.

Let us now consider the multi-mode FENE-PM model, [86]. This model is also formulated in a way that it relates the stress and the rate of strain, however, similarly to what was done before, by introducing the conformation tensor in each mode and non-dimensionalizing appropriately, we can obtain the following equations:

$$\begin{aligned} \tau &= \sum_{j=1}^{N-1} \tau_j, \\ \frac{\eta_p}{\mathrm{We}_j} g &= \left( g - \mathrm{We}_j \frac{D \ln g}{Dt} \right) \tau_{A,j} + \mathrm{We}_j \frac{\delta_F \tau_{A,j}}{\delta_F t}, \end{aligned}$$

where

$$g = 1 + (3/b) \left\{ 1 + \frac{\mathrm{We}_j}{\eta_p} \left( \frac{\mathrm{tr}(\tau)}{3(N-1)} \right) \right\}, \tag{2.4.13}$$

and $b$ is the so-called FENE parameter, $\mathrm{We}_j$ is a positive constant and $\tau_{A,j} = \tau_j + (\eta_p/\mathrm{We}_j)I$. Hence, in this case, $\alpha$ and $\beta$ are given as follows:

$$\alpha = \frac{D \ln g}{Dt} - \frac{g}{\mathrm{We}_j} \quad \text{and} \quad \beta = \frac{\eta_p}{\mathrm{We}_j^2} g.$$

The general single variable models introduced in Hulsen (1990) [44] and Beris and Edward (1994) [6] can be given in terms of the conformation tensor $\tau_A$ as follows,

$$\frac{D\tau_A(t)}{Dt} = A(t)\tau_A(t) + \tau_A(t)A(t) + g_1(t)I + g_2(t)\tau_A + g_3(t)\tau_A^2, \qquad (2.4.14)$$

where $g_i$, $(i = 1, 2, 3)$ may be functions of $\tau_A$. As mentioned before, Hulsen (1990), [44] provided a sufficient condition that $g_1(\tau_A) > 0$ for which the conformation tensor $\tau_A$ for models of the form (2.4.14) remains positive definite. His arguments were based on the investigation of the rate of change of the determinant of $\tau_A$ along the trajectory, from which he showed that an initially positive tensor $\tau_A$ can not attain non-negative eigenvalues under the assumption that $g_1(\tau_A) > 0$. Our new framework cast (2.4.14) into the general Riccati equation

$$\frac{D\tau_A(t)}{Dt} = \tilde{A}(t)\tau_A(t) + \tau_A(t)\tilde{A}(t)^T - \tau_A(t)B(t)\tau_A(t) + U(t),$$

where

$$\tilde{A}(t) = A(t) + \frac{g_2(\tau_A)}{2}I, \quad B(t) = -g_3(\tau_A)I \quad \text{and} \quad U(t) = g_1(\tau_A)I.$$

or

$$\frac{\delta_\Phi \tau_A(t)}{\delta_\Phi t} = -\tau_A(t)B(t)\tau_A(t) + U(t),$$

where

$$\frac{D\Phi(s,t)}{Dt} = \tilde{A}(t)\Phi(s,t), \quad \Phi(s,s) = I.$$

Under the assumption that $\tau_A$ is symmetric positive semi-definite initially, the simple application of the proposition (2.3.1) immediately implies that the conformation tensor $\tau_A$ evolves in time with the property of the positivity if $g_1(\tau_A) > 0$. Hence, our general framework recovers his observation in a very transparent manner.

# Chapter 3

# New Numerical Algorithms, Discrete Energy Estimates and Global Existence of Solutions

## 3.1 Introduction

From the onset attempts at numerically simulating flow of viscoelastic fluids have been hampered by a breakdown in convergence of the algorithms employed at critical values of the Weissenberg numbers. The first manifestation of the so-called high Weissenberg number problem was in the late 1970s for calculations of viscoelastic flow using finite difference methods and Galerkin finite element methods. Due to these difficulties, the successful computation of highly elastic flows exhibiting interesting experimentally observed phenomena remains elusive.

There is general and broad agreement that numerical approximation errors are primarily to blame for the loss of convergence of iterative algorithms at limiting values of the Weissenberg number. The most significant source of errors is known to be the loss of positive definiteness of the conformation tensor in the discrete level due to the discretization errors, [66, 26].

This chapter is devoted to construct stable and robust discretization schemes that preserve the positive-definiteness of the conformation tensor in the discrete level. Moreover, the importance of keeping the positive definiteness shall be quantified by the stability analysis. Namely, we shall derive the discrete analogue of energy estimates.

This discrete energy estimate shall then be further used to show the global existence and uniqueness of the discrete solution.

We believe that this type of analysis shall provide some guidance how to defeat the high Weissenberg number problem and open the gate to simulate highly elastic fluid flows.

This chapter begins with a brief review on some reformulated form of viscoelastic models in terms of the conformation tensor discussed in the chapter 2.

## 3.2 The Lagrange-Galerkin approach that preserves positivity

In this section, we shall propose numerical approximations to the following systems of viscoelastic flow equations that preserve the positivity of the conformation tensor $\tau_A$ in both time and space discrete senses regardless of the time step size and the mesh size :

$$\mathrm{Re}\frac{D\mathbf{u}}{Dt} = -\nabla p + \mathrm{div}\tau_A + \eta_s \mathrm{div}\mathcal{D}(\mathbf{u}), \tag{3.2.1}$$

$$\mathbf{divu} = 0, \tag{3.2.2}$$

and

$$\frac{\delta_E \tau_A}{\delta_E t} = -\alpha\tau_A + \beta I. \tag{3.2.3}$$

Here $\alpha$ and $\beta$ are positive and may depend on $\tau_A$. For simplicity, we shall assume that the velocity has no-slip boundary condition, namely $\mathbf{u} = 0$ on $\partial\Omega$.

Throughout this section, let us denote the current time $t = t^{n+1}$, the previous time $t^n$, the time step size $t^{n+1} - t^n = k$, $\mathbf{u}_h^{n+1} = \mathbf{u}_h(X, t^{n+1})$, $p_h^{n+1} = p_h(X, t^{n+1})$, $\tau_{A,h}^{n+1} = \tau_{A,h}(X, t^{n+1})$ and $y^n = y(X, t^{n+1}, t^n)$.

As discussed in the previous chapter 2, 3.2.3 represents a large class of constitutive equations. Recasting these constitutive equations in such a special form plays a crucial role in obtaining positivity preserving schemes. If a discretization is performed in their original formulations such as (2.2.4) and (2.4.11) or in the reformulation based on the identity (2.1.8) used in some literatures including [3] and [80] rather than (2.1.7) used here, the positivity is difficult to be preserved.

Let us illustrate this fact using the Oldroyd-B model (2.2.4). Using the identity (2.1.8), an application of the implicit Euler method on the Lagrangian frame leads to the following discrete system:

$$\begin{aligned} \tau_A^{n+1} &= \frac{\mu_p(2\mathrm{We}+k)}{\mathrm{We}(\mathrm{We}+k)} I + \frac{\mathrm{We}}{\mathrm{We}+k} F(t^n, t^{n+1}) \tau_A(t^{n+1}, t^n) F^T(t^n, t^{n+1}) \\ &- \frac{\mu_p}{\mathrm{We}+k} \left( F(t^n, t^{n+1}) F^T(t^n, t^{n+1}) + F^T(t^{n+1}, t^n) F(t^{n+1}, t^n) \right). \end{aligned}$$

It is immediate to see that the positivity of tensor $\tau_A^{n+1}$ may not be preserved unless $k$ is sufficiently small and, even for sufficiently small $k$, the positivity property may still not be preserved after sufficiently many time steps.

In the rest of this section, we shall introduce the full numerical approximations that preserve the positivity of tensor $\tau_A$ based on the so-called Lagrange-Galerkin method. Before considering temporal discretization, let us first consider discretizations on spatial variables.

### 3.2.1 Spatial discretization

In this section, we shall take a spatial discretization based on the finite element method and introduce a property that the approximation for the stress are required to possess in achieving the positivity. We assume that the domain $\Omega \subset \mathbb{R}^d$ has been partitioned into elements $\Im_h = \{K\}$ and that the partitions $\Im_h$ satisfies

$$\overline{\Omega} = \bigcup_{K \in \Im_h} \overline{K}.$$

Based on this partitions $\Im_h$, we shall choose appropriate approximation spaces $\mathbf{V}_h, W_h$ and $\mathbf{S}_h$ for the primitive variables, $\mathbf{u}$, $p$ and $\tau_A$ at any instant time $t$, respectively. Let us denote $\Pi_h^{\mathbf{V}}$, $\Pi_h^W$ and $\Pi_h^{\mathbf{S}}$ by the standard interpolation operators determined by $\mathbf{V}_h$, $W_h$ and $\mathbf{S}_h$ respectively.

The semi-discrete weak formulation of the system of equations (3.2.1), (3.2.2) and (3.2.3) based on the aforementioned finite elements shall then be formulated as follows:

Find $(\mathbf{u}_h(\cdot,t), p_h(\cdot,t), \tau_{A,h}(\cdot,t)) \in \mathbf{V}_h \times W_h \times \mathbf{S}_h$ such that $\forall (\mathbf{v}_h, q_h, \sigma_h) \in \mathbf{V}_h \times W_h \times \mathbf{S}_h$

$$\begin{aligned} \mathrm{Re}\left(\frac{D\mathbf{u}_h}{Dt}, \mathbf{v}_h\right) + (p_h, \mathbf{div}\mathbf{v}_h) &+ \eta_s\left(\mathcal{D}(\mathbf{u}_h), \mathcal{D}(\mathbf{v}_h)\right) = \left(\tau_{A,h}, \mathcal{D}(\mathbf{v}_h)\right), \\ (\mathbf{div}\mathbf{u}_h, q_h) &= 0, \\ \left(\frac{\delta_E \tau_{A,h}}{\delta_E t}, \sigma_h\right) &= -\left(\alpha \tau_{A,h}, \sigma_h\right) + \left(\beta I, \sigma_h\right), \end{aligned} \tag{3.2.4}$$

where $(\cdot,\cdot)$ denotes the usual inner $L^2(\Omega)$ product.

Let $\mathbf{V}_h^*, W_h^*$ and $\mathbf{S}_h^*$ denote the dual spaces for $\mathbf{V}_h, W_h$ and $\mathbf{S}_h$ respectively. For ease of our presentations, we shall define the following operators $A_h : \mathbf{V}_h \mapsto \mathbf{V}_h^*$, $\nabla_h : W_h \mapsto \mathbf{V}_h^*$ and $\mathrm{div}_h : \mathbf{S}_h \mapsto \mathbf{V}_h^*$ by

$$\begin{aligned}
\langle A_h \mathbf{u}_h, \mathbf{v}_h \rangle &= (\mathcal{D}(\mathbf{u}_h), \mathcal{D}(\mathbf{v}_h)), \quad \forall \mathbf{v}_h \in \mathbf{V}_h, \\
\langle \nabla_h p_h, \mathbf{v}_h \rangle &= -(p_h, \mathbf{div} \mathbf{v}_h), \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \\
\left\langle \mathrm{div}_h \tau_{A,h}, \mathbf{v}_h \right\rangle &= \Big(\tau_{A,h}, \mathcal{D}(\mathbf{v}_h)\Big), \quad \forall \mathbf{v}_h \in \mathbf{V}_h,
\end{aligned}$$

where the bracket $\langle \cdot, \cdot \rangle$ denotes the dual pairing.

The weak formulation (3.2.4) can then be written as follows:

$$\begin{aligned}
\mathrm{Re} \frac{D\mathbf{u}_h}{Dt} + \nabla_h p_h + \eta_s A_h \mathbf{u}_h &= \mathrm{div}_h \tau_{A,h} \quad \text{in } \mathbf{V}_h^* && (3.2.5) \\
\mathbf{div} \mathbf{u}_h &= 0 \quad \text{in } W_h^* && (3.2.6) \\
\frac{\delta_E \tau_{A,h}}{\delta_E t} &= -\alpha \tau_{A,h} + \beta I \quad \text{in } \mathbf{S}_h^*. && (3.2.7)
\end{aligned}$$

The actual choice of finite element spaces $\mathbf{V}_h, W_h$ and $\mathbf{S}_h$ depends on many considerations including the stability and approximation property. For example, $\mathbf{V}_h$ and $W_h$ can be chosen among the stable pairs (which satisfy certain inf-sup conditions) for the Navier-Stokes equations (see [16] or [30]). The choice of $\mathbf{S}_h$ requires special caution as it is the crucial space that leads to positivity preserving schemes.

#### 3.2.1.1 The choice of $\mathbf{S}_h$ and positivity preserving interpolant

In this subsection, we shall address details about the choice of the approximation space $\mathbf{S}_h$. To keep the positivity of $\tau_A$, there should exist a linear operator $\Pi_h^{\mathbf{S}}$ whose range is $\mathbf{S}_h$ and that preserves the positivity in the following sense :

$$\sigma > 0 \quad \Rightarrow \quad \Pi_h^{\mathbf{S}}(\sigma) > 0. \tag{3.2.8}$$

Here $\sigma$ is any $d \times d$ symmetric tensor and $\sigma > 0$ means that the tensor $\sigma$ is symmetric and positive definite. See e.g. (3.2.16), (3.2.20) and remarks that follow.

Such an operator can be in particular generated by the following linear and positive operator $\Pi_h$ defined on scalar functions

$$u > 0 \quad \Rightarrow \quad \Pi_h(u) > 0 \tag{3.2.9}$$

where $u$ is any scalar function.

Note that the inequality " $>$ " is meant to be " $>$ " almost everywhere since functions may not be defined pointwise. To see $\Pi_h^{\mathbf{S}}$ can be constructed by $\Pi_h$, take any $\sigma = (\sigma_{ij})_{i,j=1,\cdots,d}$, $d \times d$ symmetric tensor and consider any nonzero vector $\xi = (\xi_i)_{i=1,\cdots,d}$ and observe that

$$0 < \xi^T \sigma \xi \;\; = \;\; \sum_{i,j=1}^{d} \xi_i \sigma_{ij} \xi_j \quad \Rightarrow \quad 0 < \Pi_h \left( \xi^T \sigma \xi \right) \quad \text{by (3.2.9)}$$

Now, by the linearity of $\Pi_h$, we see that

$$\Pi_h\left(\xi^T\sigma\xi\right) = \sum_{i,j=1}^{d} \xi_i \Pi_h(\sigma_{ij})\xi_j \quad \in S(\Omega;h)$$

Hence $\Pi_h^{\mathbf{S}}$ is defined through

$$\Pi_h^{\mathbf{S}}(\sigma) = (\Pi_h(\sigma_{ij}))_{ij=1,\cdots,d} \quad \in S^{d\times d}(\Omega;h).$$

The most natural candidate for $S(\Omega;h)$ is the space of Lagrange finite elements of total polynomial degree $\leq k$ with $k \geq 0$. For $k = 0$, we have a finite element space of piecewise constant. In this case, the existence of the positive preserving $\Pi_h$ is obvious. For example, we can take, on each element $K$,

$$\Pi_h(u)(x) = \frac{1}{|K|}\int_K u\,dx \quad \forall x \in K. \tag{3.2.10}$$

Of course, this choice of $S(\Omega;h)$ only leads to the first order approximation for the conformation tensor.

For $k = 1$, there are two possibilities. The first one is to choose globally continuous piecewise linear finite element function. In this case, the standard pointwise nodal value interpolant certainly would be positivity preserving. In case $u$ is rough, point values of $u$ are not well-defined, we may define the nodal value of $\Pi_h(u)(x_i)$ as the local mean-value as follows

$$\Pi_h(u)(x_i) := \frac{1}{|B_i|}\int_{B_i} u\,dx, \tag{3.2.11}$$

where $B_i = B(x_i, r_i(x_i))$, the ball centered at $x_i$ with radius $r_i(x_i)$ with $r_i(x_i)$ chosen small enough so that $B_i$ is contained in the union of closed elements containing $x_i$. The above construction (3.2.11) was proposed recently by Nochetto and Wahlbin (2001) [64] and it is easy to see that such an operator $\Pi_h$ preserves linear functions and has a second order accuracy.

Another possibility for $k = 1$ is the discontinuous piecewise linear finite element space. In this case, the construction of positivity preserving operator $\Pi_h$ for the above continuous piecewise linear element case can obviously be applied here.

For $k \geq 2$, however, it is known that it is impossible to construct a positivity preserving interpolant that has more than second order accuracy. For details, we refer to [64].

In summary, we can choose $\mathbf{S}_h$ as either piecewise constant or piecewise linear finite element spaces. The approximation accuracy for such choices is either first order or second order. It is in general not possible to construct a positivity preserving scheme for the conformation tensor whose approximation accuracy is more than second order.

### 3.2.2 Temporal discretizations

In this section, we shall discretize the equations (3.2.5) and (3.2.7) in time. We shall use the particle following approach in order to exploit the connection between the Riccati equation and the constitutive equation (3.2.3). In this work, we shall take the semi-Lagrangian methodology, (see [67] and [68]) rather than the Lagrangian method since in the Lagrangian framework, the mesh moves with the particle and in case of large deformation, the mesh can be severely distorted and re-meshing is inevitable introducing

additional numerical errors (see e.g. [3] or [67] and the references cited therein). To implement the semi-Lagrangian method, given any material particle, $X$ at the current time $t^{n+1}$, the particle path should be determined by

$$\frac{dy(X,t^{n+1},s)}{ds} = \mathbf{u}(y(X,t^{n+1},s),s), \quad y(X,t^{n+1},t^{n+1}) = X. \tag{3.2.12}$$

The Lie derivative $\dfrac{\delta_L \xi}{\delta_L t}$ at time $t^{n+1}$ can then be approximated on the time interval $[t^n, t^{n+1}]$, for example, by the following first order time discretization :

$$\frac{\delta_L \xi}{\delta_L t}(t^{n+1}) \approx \frac{\xi(t^{n+1}) - L(t^n,t^{n+1})\xi(t^{n+1},t^n)L(t^n,t^{n+1})}{k}, \tag{3.2.13}$$

where $\xi$ is either a vector or a tensor and correspondingly $L$ is either $I$ or $E$ and $\xi(t^{n+1},t^n) = \xi(t^n)\circ y^n$.

Throughout this section, we shall assume that $\mathbf{S}_h$ has been chosen so that it is the range of $\Pi_h^{\mathbf{S}}$, which satisfies the property (3.2.8).

The main goal of this section is to develop the positivity preserving time discretization of (3.2.7). The following two approaches will be taken. The first approach is to discretize the material derivative. It will be first order time accurate. The second approach is to use the analytic solution (2.4.6) introduced in the previous chapter 2. Most of our presentations here is based on the work by Dieci and Eirola (1994), [24].

Let us begin by recalling that

$$\frac{\delta_E \tau_A}{\delta_E t} = \frac{D\tau_A}{Dt} - R(t)\tau_A - \tau_A R(t)^T. \tag{3.2.14}$$

Regarding the time derivative as the material derivative, taking an implicit Euler scheme to (3.2.7), we obtain

$$\begin{aligned} \alpha^{n+1}\tau_{A,h}^{n+1} &+ \left( \frac{\tau_{A,h}^{n+1} - \Pi_h^{\mathbf{S}}(\tau_{A,h}^n \circ y^n)}{k} - \Pi_h^{\mathbf{S}}(R_h^{n+1}\tau_{A,h}^{n+1}) - \Pi_h^{\mathbf{S}}(\tau_{A,h}^{n+1}(R_h^{n+1})^T) \right) \\ &= \beta^{n+1} I. \end{aligned} \tag{3.2.15}$$

An interesting fact is that this equation can be recast into the well-known algebraic Riccati differential equation as follows:

$$\begin{aligned} &\left( \frac{\alpha^{n+1}k+1}{2k} - R_h^{n+1} \right) \tau_{A,h}^{n+1} + \tau_{A,h}^{n+1} \left( \frac{\alpha^{n+1}k+1}{2k} - R_h^{n+1} \right)^T \\ = &\frac{\tau_{A,h}^n \circ y^n}{k} + \beta^{n+1} I. \end{aligned} \tag{3.2.16}$$

Note that the equivalence between two equations (3.2.15) and (3.2.16) should be asserted with taking $\Pi_h^{\mathbf{S}}$ for both sides of equation (3.2.16). We simply did not write it to clarify (3.2.16) is an algebraic Riccati equation. We also note that it is necessary to consider the range of $\Pi_h^{\mathbf{S}}$ with the property (3.2.8) as the approximation space $\tau_A$ to make sure $\tau_{A,h}^n \circ y^n > 0$ and under this condition, it can be also shown that the equation (3.2.16) has a unique positive definite solution $\tau_{A,h}^{n+1}$ (see e.g. [24] or [53]).

The numerical schemes based on the analytic solution (2.4.6) to (3.2.7) require approximations of (2.4.9), $E(s,t)$ and the integral expression in (2.4.6). Approximations of the first two quantities can be made in any ways since these do not affect the positivity property. However the integral expression in (2.4.6) should be approximated, for example by using numerical quadrature with positive weights in order to keep the positivity.

We shall introduce three different approximations for each of them. For every cases, the first scheme is a first order implicit scheme, the second scheme is a second order single step implicit scheme and the third scheme is a second order two step explicit scheme.

- Approximations of $\exp\left(-\int_s^t \alpha(\nu)d\nu\right)$ :

$$\begin{aligned}
\alpha_{t^n,t^{n+1}} &:= 1 - \alpha^{n+1}k, \\
\widetilde{\alpha}_{t^n,t^{n+1}} &:= \exp\left(-\frac{k}{2}\left[\alpha(t^n) + \alpha(t^{n+1})\right]\right), \\
\widehat{\alpha}_{t^{n-1},t^{n+1}} &:= \exp\left(-k\alpha(t^n)\right).
\end{aligned}$$

- Approximations of $E(s,t)$

$$\begin{aligned}
E_h(t^n,t^{n+1}) &:= I + kR_h(t^n)\circ y^n \quad \text{or} \quad \left(I - kR_h(t^{n+1})\right)^{-1}, \\
\widetilde{E_h}(t^n,t^{n+1}) &:= \left(I - \frac{k}{2}\left(R_h^{n+1} + R_h^n \circ y^n\right)\right)^{-1}\left(I + \frac{k}{2}\left(R_h^{n+1} + R_h^n \circ y^n\right)\right), \\
\widehat{E_h}(t^{n-1},t^{n+1}) &:= \left(I - kR_h^n \circ y^n\right)^{-1}\left(I + kR_h^n \circ y^n\right),
\end{aligned}$$

where

$$R_h(t) = \frac{a+1}{2}\nabla_h \mathbf{u}_h(t) + \frac{a-1}{2}\nabla_h \mathbf{u}_h(t)^T. \tag{3.2.17}$$

- Approximations of $\int_s^t \beta(t)\Phi(\nu,t)\Phi(\nu,t)^T d\nu$

$$
\begin{aligned}
I_{t^n,t^{n+1}} &:= k\beta^{n+1} \\
\widetilde{I}_{t^n,t^{n+1}} &:= \frac{k}{2}\left(\beta^n \widetilde{\Phi_h}(t^n,t^{n+1})\widetilde{\Phi_h}(t^n,t^{n+1})^T + \beta^{n+1}\right) \\
\widehat{I}_{t^{n-1},t^{n+1}} &:= k\beta^n \left(\frac{I+\widehat{\Phi_h}(t^{n-1},t^{n+1})}{2}\right)\left(\frac{I+\widehat{\Phi_h}(t^{n-1},t^{n+1})}{2}\right)^T,
\end{aligned}
$$

where

$$
\begin{aligned}
\widetilde{\Phi_h}(t^n,t^{n+1}) &:= \widetilde{\alpha}_{t^n,t^{n+1}}\widetilde{E_h}(t^n,t^{n+1}) \\
\widehat{\Phi_h}(t^{n-1},t^{n+1}) &:= \widehat{\alpha}_{t^{n-1},t^{n+1}}\widehat{E_h}(t^{n-1},t^{n+1})
\end{aligned}
$$

Based on the aforementioned various approximations, we can devise three different approximations for (3.2.7).

First, based on the first order approximations $\alpha_{t^n,t^{n+1}}$ and $I_{t^n,t^{n+1}}$, we have

$$
\begin{aligned}
&\frac{\tau_{A,h}^{n+1} - \Pi_h^{\mathbf{S}}\left(E_h(t^n,t^{n+1})(\tau_{A,h}^n \circ y^n)E_h(t^n,t^{n+1})^T\right)}{k} \\
&\qquad = -\alpha^{n+1}\tau_{A,h}^{n+1} + \beta^{n+1} I,
\end{aligned}
\tag{3.2.18}
$$

It is interesting to note that (3.2.18) can also be obtained by a direct application of implicit Euler method to (3.2.3) based on the time discretization (3.2.13).

The numerical approximation based on the implicit second order single step scheme, namely $\widetilde{I}_{t^n,t^{n+1}}$ can be given as follows.

$$\begin{aligned}\tau_{A,h}^{n+1} &= \Pi_h^{\mathbf{S}}\left(\widetilde{\Phi_h}(t^n,t^{n+1})\left(\tau_{A,h}^n\circ y^n+\frac{k}{2}\beta^n\right)\widetilde{\Phi_h}(t^n,t^{n+1})^T\right) \\ &+ \frac{k}{2}\beta^{n+1}.\end{aligned} \tag{3.2.19}$$

Finally, two step explicit scheme $\widehat{I}_{t^{n-1},t^{n+1}}$ produces the following formula :

$$\begin{aligned}\tau_{A,h}^{n+1} &= \Pi_h^{\mathbf{S}}\left(\widehat{\Phi_h}(t^{n-1},t^{n+1})\left(\tau_{A,h}^{n-1}\circ y^{n-1}\right)\widehat{\Phi_h}(t^{n-1},t^{n+1})^T\right) \\ &+ \frac{k}{4}\Pi_h^{\mathbf{S}}\left(\beta^n\left(I+\widehat{\Phi_h}(t^{n-1},t^{n+1})\right)\left(I+\widehat{\Phi_h}(t^{n-1},t^{n+1})\right)^T\right)\end{aligned} \tag{3.2.20}$$

Recall that the time discrete equation (3.2.7) is positivity preserving under the assumption that $\tau_{A,h}^n\circ y^n$ is positive definite and this holds for any spaces $\mathbf{S}_h$ which is the range of $\Pi_h^{\mathbf{S}}$ with the property (3.2.8).

### 3.2.3 Full spatial and temporal discretizations

In this section, we shall complete our numerical approximations combining discretizations of the momentum equation (3.2.5) with the aforementioned discretizations of the constitutive equations (3.2.7).

Taking first order approximation implicit Euler for (3.2.5) together with (3.2.18) for the constitutive equation, we obtain :

First Order Scheme

$$\mathrm{Re}\frac{\mathbf{u}_h^{n+1} - \Pi_h^{\mathbf{V}}\left(\mathbf{u}_h^n \circ y^n\right)}{k} + \nabla_h p_h^{n+1} + \eta_s A_h \mathbf{u}_h^{n+1} = \mathrm{div}_h \tau_{A,h}^{n+1} \tag{3.2.21}$$

$$\mathbf{divu}_h^{n+1} = 0 \tag{3.2.22}$$

$$\begin{aligned}\frac{\tau_{A,h}^{n+1} - \Pi_h^{\mathbf{S}}\left(E_h(t^n, t^{n+1})(\tau_{A,h}(t^n) \circ y^n)E_h(t^n, t^{n+1})^T\right)}{k} \\ = -\alpha^{n+1}\tau_{A,h}^{n+1} + \beta^{n+1} I,\end{aligned} \tag{3.2.23}$$

The Crank-Nicolson scheme for (3.2.5) combined with the implicit second order scheme (3.2.19) for (3.2.7) results in

Second Order Single Step Scheme

$$\begin{aligned}\mathrm{Re}\frac{\mathbf{u}_h^{n+1} - \Pi_h^{\mathbf{V}}\left(\mathbf{u}_h^n \circ y^n\right)}{k} &+ \frac{1}{2}\left(\nabla_h p_h^{n+1} + \nabla_h \Pi_h^W\left(p_h^n \circ y^n\right)\right) \\ +\frac{\eta_s}{2}\left(A_h \mathbf{u}_h^{n+1} + A_h \Pi_h^{\mathbf{V}}(\mathbf{u}_h^n \circ y^n)\right) &= \frac{1}{2}\left(\mathrm{div}_h \tau_{A,h}^{n+1} + \mathrm{div}_h \Pi_h^{\mathbf{S}}\left(\tau_{A,h}^n \circ y^n\right)\right)\end{aligned} \tag{3.2.24}$$

$$\mathbf{divu}_h^{n+1} = 0$$

$$\begin{aligned}\tau_{A,h}^{n+1} &= \Pi_h^{\mathbf{S}}\left(\widetilde{\Phi_h}(t^n,t^{n+1})\left(\tau_{A,h}^n\circ y^n+\frac{k}{2}\beta^n\right)\widetilde{\Phi_h}(t^n,t^{n+1})^T\right)\\ &+ \frac{k}{2}\beta^{n+1}.\end{aligned}$$

Finally, the combination of the stiffly stable second order BDF, [88] for (3.2.5) and second order explicit two step scheme (3.2.20) for (3.2.7) produces another second order schemes as follows:

Second Order Two Step Scheme

$$\begin{aligned}\mathrm{Re}\frac{\frac{3}{2}\mathbf{u}_h^{n+1}-\Pi_h^{\mathbf{V}}\left(2\mathbf{u}_h^n\circ y^n+\frac{1}{2}\mathbf{u}_h^{n-1}\circ y^{n-1}\right)}{k} &- \\ \nabla_h p_h^{n+1}+\eta_s A_h\mathbf{u}_h^{n+1} &= \mathrm{div}_h\tau_{A,h}^{n+1} \qquad (3.2.25)\end{aligned}$$

$$\mathbf{divu}_h^{n+1}=0$$

$$\begin{aligned}\tau_{A,h}^{n+1} &= \Pi_h^{\mathbf{S}}\left(\widehat{\Phi_h}(t^{n-1},t^{n+1})\left(\tau_{A,h}^{n-1}\circ y^{n-1}\right)\widehat{\Phi_h}(t^{n-1},t^{n+1})^T\right)\\ &+ \frac{k}{4}\Pi_h^{\mathbf{S}}\left(\beta^n\left(I+\widehat{\Phi_h}(t^{n-1},t^{n+1})\right)\left(I+\widehat{\Phi_h}(t^{n-1},t^{n+1})\right)^T\right)\end{aligned}$$

We note that the stiffly stable scheme (3.2.25) is preferred especially for the long time computations than the Crank-Nicolson scheme (3.2.24) since it is known that the presence of the explicit part of the pressure in such a formulation may incur some numerical instability (see e.g. [88]). We would like to remark that for such a two step scheme,

at least two previous step solutions should be available to proceed to the current step solution and so it may not be applied to obtaining the first step solution. However, to get the first step solution, a single step second order scheme can be used instead. This allows us to keep the overall temporal accuracy.

## 3.3 Stability Analysis : Continuous and Discrete energy estimates

The purpose of this section is to show the importance of keeping the positivity in the discrete sense. We shall first drive the energy estimates of the general form of viscoelastic models (3.2.1), (3.2.2) and (3.2.3) in the continuous level. We then take a specific algorithm (3.2.21), (3.2.22) and (3.2.23) presented in §3.2.3 to show the discrete analogue of energy law. Namely, we shall show several *a priori* estimate of the numerical solutions. By this, it means both the stability and the robustness of the algorithm. Our analysis crucially relies on the positivity of the conformation tensor and energy estimate holds no matter how large the Weissenberg number is. The analysis indeed includes the limiting case when $\mathrm{We} = \infty$.

For any positive definite tensor $\sigma$, the $L^1(\Omega)$ norm for $\sigma$ shall be defined as follows:

$$\|\sigma\|_{L^1} := \int_\Omega \mathrm{tr}(\sigma)\, dx. \tag{3.3.1}$$

In the following discussion, we shall further adopt the standard notation for Sobolev spaces $H_0^k(\Omega)$, $H^k(\Omega)$ with norms denoted by $\|\cdot\|_k$ and also $L^p(0,t;H^1(\Omega))$ and $L^p(0,t;L^q(\Omega))$.

Especially, by $f(x,t) \in L^p(0,t;H^1(\Omega))$, we mean that

$$\left(\int_0^t \|f(\cdot,\nu)\|_1^p \, d\nu\right)^{1/p} < \infty. \tag{3.3.2}$$

As usual, $\|\cdot\|_0$ shall denote the $L^2$ norm.

### 3.3.1 Continuous Energy Estimates

In this section, we shall drive the energy law of models (3.2.1), (3.2.2) and (3.2.3) and then we further show the energy estimates in the continuous level. In doing so, we shall make the following assumptions :

A1: $\exists$ a constant $c > 0$ such that $\alpha \geq c > 0$ and $\alpha(t) \in L^\infty(\Omega)$ for all $t \geq 0$.

A2: $\beta \geq 0$ and $\beta(t) \in L^1(\Omega)$ for all $t \geq 0$.

A3: $\tau_A(x,0)$ is semi-positive definite.

The assumption $A1$ is related to the so-called the damping. This holds for all the model in the Table 5.10 except for the FENE-PM model for which the definiteness of $\alpha$ seems to be difficult to determine. $A2$ and $A3$ are necessary for the positivity of the conformation tensor in time evolutions and especially $A2$ is also for the boundedness of the energy.

The following energy law is easy but important for the energy estimates.

**Lemma 3.3.1.** *Assume that $\alpha \geq 0$, A2 and A3, we have the following energy law for the models,* (3.2.1), (3.2.2) *and* (3.2.3) *:*

$$\begin{aligned} \mathrm{Re}\frac{d}{dt}\|\mathbf{u}(\cdot,t)\|_0^2 &+ \frac{1}{2a}\frac{d}{dt}\|\tau_A(\cdot,t)\|_{L^1} + \eta_s\|\mathcal{D}(\mathbf{u}(\cdot,t))\|^2 \\ &= -\frac{1}{2a}\|\alpha\tau_A(\cdot,t)\|_{L^1} + \frac{d}{2a}\int_\Omega \beta(\cdot,t)\,dx. \end{aligned} \tag{3.3.3}$$

*Proof.* We first take the trace on the equation (3.2.3). Now take the integration and use the fact that $\mathbf{divu} = 0$ to obtain :

$$\begin{aligned} \|\alpha\,\tau_A(\cdot,t)\|_{L^1} &+ \frac{d}{dt}\|\tau_A(\cdot,t)\|_{L^1} \\ -\int_\Omega \mathrm{tr}\left(R(t)\tau_A(\cdot,t)\right)\,dx &- \int_\Omega \mathrm{tr}\left(\tau_A(\cdot,t)R(t)^T\right)\,dx = d\int_\Omega \beta(\cdot,t)\,dx. \end{aligned} \tag{3.3.4}$$

We notice the following simple but important relation :

$$\begin{aligned} (\tau_A(\cdot,t) : \mathcal{D}(\mathbf{u}(\cdot,t))) &= \frac{1}{2a}\int_\Omega \mathrm{tr}\left(R(t)\tau_A\right)dx \\ &= \frac{1}{2a}\int_\Omega \mathrm{tr}\left(\tau_A R(t)^T\right)\,dx. \end{aligned} \tag{3.3.5}$$

Let us now consider the momentum equation. Multiplying $\mathbf{u}$ to (3.2.1) and taking integration, we obtain that

$$\mathrm{Re}\frac{d}{dt}\|\mathbf{u}(\cdot,t)\|_0^2 + \eta_s\|\mathcal{D}(\mathbf{u}(\cdot,t))\|_0^2 dx = -\left(\tau_A(\cdot,t) : \mathcal{D}(\mathbf{u}(\cdot,t))\right) \tag{3.3.6}$$

From (3.3.4) and (3.3.5), we obtain

$$\begin{aligned} \mathrm{Re}\frac{d}{dt}\|\mathbf{u}(\cdot,t)\|_0^2 &+ \frac{1}{2a}\frac{d}{dt}\|\tau_A(\cdot,t)\|_{L^1} + \eta_s\|\mathcal{D}(\mathbf{u}(\cdot,t))\|^2 \\ &= -\frac{1}{2a}\|\alpha\tau_A(\cdot,t)\|_{L^1} + \frac{d}{2a}\int_\Omega \beta(\cdot,t)\,dx. \end{aligned} \tag{3.3.7}$$

This completes the proof. □

Note that physically, the trace of the conformation tensor can be thought of as the length from the tail to the head of the macromolecule. As fluid flows, the molecules can be stretched, which means that the molecules store an energy. The more stretched, the more energy they store. One may then view $\|\tau_A\|_{L^1}$ as a total elastic energy due to the interaction between macromolecules and fluids. This is an elaborate argument on the importance of keeping the positivity of the conformation tensor in the physical terms.

**Theorem 3.3.1.** *Assume A1, A2 and A3. Then for any $t > 0$, the following estimates hold true :*

$$\begin{aligned} \mathrm{Re}\|\mathbf{u}(\cdot,t)\|_0^2 &+ \frac{1}{2a}\|\tau_A(\cdot,t)\|_{L^1} \le \exp\left(-C_1 t\right)\left(\mathrm{Re}\|\mathbf{u}(\cdot,0)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,0)\|_{L^1}\right) \\ &+ \frac{C_2(t)}{C_1}\left(1-\exp(-C_1 t)\right) \end{aligned} \tag{3.3.8}$$

*and*

$$\eta_s\int_0^t \|\nabla\mathbf{u}(\cdot,\nu)\|_0^2\,d\nu \le \left(\mathrm{Re}\|\mathbf{u}(\cdot,0)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,0)\|_{L^1} + C_2(t)\,t\right), \tag{3.3.9}$$

*where*

$$C_1 = \min\left(\frac{c_\Omega \eta_s}{\mathrm{Re}}, c\right) \quad \textit{and} \quad C_2(t) = \frac{d}{2a} \sup_{0\le s\le t} \|\beta(\cdot,s)\|_{L^1}. \tag{3.3.10}$$

*Here $c_\Omega$ is a positive constant depending only on $\Omega$ from the Korn's inequality.*

*Proof.* To obtain the estimates (3.3.8) and (3.3.9), we shall start with the energy law (3.3.3) from the Lemma 3.3.1. Using the Korn's inequality and the assumption $c = \min_\Omega \alpha > 0$, we obtain the following inequality from (3.3.3) :

$$\begin{aligned} \frac{d}{dt}\left(\mathrm{Re}\|\mathbf{u}(\cdot,t)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,t)\|_{L^1}\right) &\le -\eta_s c_\Omega \|\mathbf{u}(\cdot,t)\|_0^2 \\ &- \frac{c}{2a}\|\tau_A(\cdot,t)\|_{L^1} + \frac{d}{2a}\int_\Omega \beta(\cdot,t)\,dx, \end{aligned} \tag{3.3.11}$$

where $c_\Omega$ is a generic constant depending only on $\Omega$ from the Korn's inequality. Let us denote $C_1 = \min\left(\frac{c_\Omega \eta_s}{\mathrm{Re}}, c\right)$. With $C_1$, we then obtain the following inequality :

$$\begin{aligned} \frac{d}{dt}\left(\mathrm{Re}\|\mathbf{u}(\cdot,t)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,t)\|_{L^1}\right) &\le -C_1\left(\mathrm{Re}\|\mathbf{u}(\cdot,t)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,t)\|_{L^1}\right) \\ &+ \frac{d}{2a}\int_\Omega \beta(\cdot,t)\,dx. \end{aligned} \tag{3.3.12}$$

It is easy to drive the following inequality from (3.3.12) :

$$\begin{aligned} \mathrm{Re}\|\mathbf{u}(\cdot,t)\|_0^2 &+ \frac{1}{2a}\|\tau_A(\cdot,t)\|_{L^1} \\ &\le e^{-C_1 t}\left(\mathrm{Re}\|\mathbf{u}(\cdot,0)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,0)\|_{L^1}\right) + \frac{C_2(t)}{C_1}\left(1-\exp(-C_1 t)\right), \end{aligned} \tag{3.3.13}$$

where $C_2 = \dfrac{d}{2a} \sup\limits_{0\le s\le t} \|\beta(\cdot,s)\|_{L^1}$. So the estimate (3.3.8) follows. To obtain the estimate (3.3.9), we take the integration of (3.3.3) to get

$$\begin{aligned} \mathrm{Re}\|\mathbf{u}(\cdot,t)\|_0^2 &+ \frac{1}{2a}\|\tau_A(\cdot,t)\|_{L^1} + \eta_s \int_0^t \|\mathbf{u}(\cdot,\nu)\|_1^2 \, d\nu \qquad (3.3.14)\\ \le \mathrm{Re}\|\mathbf{u}(\cdot,0)\|_0^2 &+ \frac{1}{2a}\|\tau_A(\cdot,0)\|_{L^1} - \int_0^t \frac{1}{2a}\|\alpha\tau_A(\cdot,\nu)\|_{L^1} \, d\nu + C_2(t)t \\ \le \mathrm{Re}\|\mathbf{u}(\cdot,0)\|_0^2 &+ \frac{1}{2a}\|\tau_A(\cdot,0)\|_{L^1} + C_2(t)t. \end{aligned}$$

This completes the proof. □

Next, we shall consider the limiting case when $\mathrm{We} = \infty$. In this case, for most of models presented in the previous sections, $\alpha = \beta = 0$. So, we shall discuss the limiting case restricted to such case :

**Corollary 3.3.1.** *Assume that* $\mathrm{We} = \infty$ *and* $\alpha = \beta = 0$. *Then we have the following estimates :*

$$\mathrm{Re}\|\mathbf{u}(\cdot,t)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,t)\|_{L^1} \le \mathrm{Re}\|\mathbf{u}(\cdot,0)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,0)\|_{L^1}. \qquad (3.3.15)$$

*and*

$$\eta_s \int_0^t \|\nabla\mathbf{u}(\cdot,\nu)\|_0^2 \le \mathrm{Re}\|\mathbf{u}(\cdot,0)\|_0^2 + \frac{1}{2a}\|\tau_A(\cdot,0)\|_{L^1} \qquad (3.3.16)$$

*Proof.* Both estimates (3.3.15) and (3.3.16) immediately follow from (3.3.8) and (3.3.9) due to the fact that $\alpha = \beta = 0$ imply $C_1 = C_2(t) = 0$. This completes the proof. □

A priori estimates like what are obtained in this section are often essential ingredients in establishing the global (in time) well-posedness of non-linear partial differential

equations. It is still an open problem if any of the non-Newtonian models discussed in this paper is globally well-posed (in weak sense). This is a topic of active research, see e.g. [20], [56], [58] and [57].

### 3.3.2 Discrete Energy Estimates

In this section, we shall drive the discrete analogue of the energy estimate and shall demonstrate the stability and robustness of our new algorithms. Especially, we shall take a specific algorithm (3.2.21), (3.2.22) and (3.2.23) and show how the positivity play a role in the stability analysis. Let us begin this section by a modified Galerkin finite element method for discretizations (3.2.21), (3.2.22) and (3.2.23). Namely, given $(\mathbf{u}_h^n, \tau_{A,h}^n)$, find $(\mathbf{u}_h^{n+1}, \tau_{A,h}^{n+1}) \in \mathbf{V}_h \times \mathbf{S}_h$ such that for all $\mathbf{v}_h \in \mathbf{V}_h$,

$$\begin{aligned} \mathrm{Re}\left(\frac{\mathbf{u}_h^{n+1} - \Pi_h^{\mathbf{V}}\left(\mathbf{u}_h^n \circ y^n\right)}{k}, \mathbf{v}_h\right) &+ \eta_s\left(\mathcal{D}(\mathbf{u}_h^{n+1}) : \mathcal{D}(\mathbf{v}_h)\right) \\ &= -\left(\tau_{A,h}^{n+1} : \mathcal{D}(\mathbf{v}_h)\right) \end{aligned} \tag{3.3.17}$$

$$\begin{aligned} \frac{\tau_{A,h}^{n+1} - \Pi_h^{\mathbf{S}}\left(E_h(t^n, t^{n+1})(\tau_{A,h}^n \circ y^n)E_h(t^n, t^{n+1})^T\right)}{k} \\ = -\alpha^{n+1}\tau_{A,h}^{n+1} + \beta^{n+1} I, \end{aligned} \tag{3.3.18}$$

where

$$E_h(t^n, t^{n+1}) := \left(I - kR_h^{n+1}\right)^{-1}. \tag{3.3.19}$$

As indicated in the above formulation, we shall assume that $\mathbf{V}_h$ and $W_h$ are chosen in a way that

$$\mathbf{div}\mathbf{u}_h \in W_h, \quad \forall\, \mathbf{u}_h \in \mathbf{V}_h. \tag{3.3.20}$$

Observe further that the constitutive equation is formulated in the strong sense. This indeed results in a quite efficient numerical scheme in actual computations and the stability analysis shall be performed for this type of formulation.

Throughout this section, we shall further assume that he interpolation operator $\Pi_h^{\mathbf{S}}$ for the stress field is given by

$$\Pi_h^{\mathbf{S}}(\sigma) = \Big(\Pi_h(\sigma_{ij})\Big)_{ij=1,\cdots,d}, \tag{3.3.21}$$

where $\Pi_h$ is defined through (see (3.2.10) in §3.2.3)

$$\Pi_h(u)(y) := \sum_{l=1}^{N} \left(\frac{1}{|K_l|}\int_{K_l} u\, dx\right)\phi_l(y). \tag{3.3.22}$$

Here $\Im_h = \{K_l\}_{l=1}^N$ and $\phi_l(y)$ is a characteristic function which is one for $y \in \overline{K}_l$ and zero elsewhere. Furthermore, we apply the following approach to update $\tau_{A,h}^{n+1}$ from (3.3.18). More precisely, the interpolation $\Pi_h^{\mathbf{S}}$ operator shall be taken by the following procedures

$$\begin{aligned} &\Pi_h^{\mathbf{S}}\left(E_h(t^n,t^{n+1})(\tau_{A,h}^n\circ y^n)E_h(t^n,t^{n+1})\right) \\ =\;& \left(I-k(R_h^{n+1})_h\right)^{-1}(\tau_{A,h}^n\circ y^n)\left(I-k(R_h^{n+1})_h\right)^{-T}, \end{aligned} \tag{3.3.23}$$

where $(R_h^{n+1})_h := \Pi_h^{\mathbf{S}}(R_h^{n+1})$. We shall also assume that the volume preserving numerical approximation for the characteristic feet has been used.

We would like to remark that a stable pair of spaces $\mathbf{V}_h$ and $W_h$ satisfying the relation (3.3.20) can indeed be made especially with continuous piecewise polynomial of degree $k$ and $k-1$ for $\mathbf{V}_h$ and $W_h$ respectively with $k \geq 4$ (see [14] or [77]) and with such a choice of approximation spaces, volume preserving schemes for the characteristic feet can be devised for both two and three space dimensions (see e.g. K. Feng (1995) [28]).

Indeed, as the volume preserving scheme for the characteristic feet is crucial for the following stability analysis, it will be rigorously discussed in the section §3.3.3.

Finally, we shall make the following discrete analogue of the assumptions $A1$, $A2$ and $A3$ :

$A1_h$: $\exists$ a constant $c > 0$ such that $\alpha^n \geq c > 0$ and $\alpha^n \in L^\infty(\Omega)$ for all $n \geq 0$.

$A2_h$: $\beta^n \geq 0$ and $\beta^n \in L^1(\Omega)$ for all $n \geq 0$.

$A3_h$: $\tau_{A,h}^0$ is semi-positive definite.

The following simple lemma is instrumental to drive a discrete analogue of energy estimate.

**Lemma 3.3.2.** *Assume $A$ and $B$ are matrix-valued functions, $A_h = \Pi_h^{\mathbf{S}}(A)$ and $B_h = \Pi_h^{\mathbf{S}}(B)$. Then the following holds true :*

$$(A_h : B) = (A_h : B_h) = (A : B_h)\,. \tag{3.3.24}$$

*Proof.* Note that

$$(A : B) = \int_\Omega \mathrm{tr}(AB)\, dx. \tag{3.3.25}$$

and the trace operator is linear. It is then enough to show that for any scalar functions $f$ and $g$, the following holds true :

$$\int_\Omega f\, g_h\, dx = \int_\Omega f_h\, g_h\, dx = \int_\Omega f_h\, g\, dx, \tag{3.3.26}$$

where $f_h = \Pi_h(f)$ and $g_h = \Pi_h(g)$. We observe that

$$\begin{aligned}
\int_\Omega f\, g_h\, dx &= \int_\Omega f \sum_{l=1}^{N} \left( \frac{1}{|K_l|} \int_{K_l} g\, dx \right) \phi_l\, dy \\
&= \sum_{l=1}^{N} \left( \int_\Omega f\, \phi_l\, dy \right) \left( \frac{1}{|K_l|} \int_{K_l} g\, dx \right) \\
&= \sum_{l=1}^{N} \int_\Omega \left( \frac{1}{|K_l|} \int_{K_l} f\, dy \right) \phi_l\, dz \left( \frac{1}{|K_l|} \int_{K_l} g\, dx \right) \\
&= \int_\Omega \sum_{l=1}^{N} \left( \frac{1}{|K_l|} \int_{K_l} f\, dy \right) \phi_l \left( \frac{1}{|K_l|} \int_{K_l} g\, dx \right) \phi_l\, dz = \int_\Omega f_h\, g_h\, dx.
\end{aligned}$$

The other equality $\int_\Omega f_h\, g_h\, dx = \int_\Omega f_h\, g\, dx$ follows in a straightforward manner from the similar argument. This completes the proof. □

We shall now state and prove our main theorem in this section.

**Theorem 3.3.2.** *Assume that $A1_h$, $A2_h$ and $A3_h$. Then the Lagrange-Galerkin formulation* (3.3.17), (3.3.18) *together with* (3.3.19) *provide solutions $\mathbf{u}_h^n$ and $\tau_{A,h}^n$ for $n \geq 1$*

*satisfying the following estimate :*

$$\begin{aligned} \mathrm{Re}\|\mathbf{u}_h^n\|_0^2 &+ \frac{1}{2a}\|\tau_{A,h}^n\|_{L^1} \\ &\leq c_1 \exp\left(-C_1 t^n\right)\left(\mathrm{Re}\|\mathbf{u}_h^0\|_0^2 + \frac{1}{2a}\|\tau_{A,h}^0\|_{L^1}\right) + c_2 C_2^n \end{aligned} \tag{3.3.27}$$

*and*

$$2\eta_s \sum_{l=0}^{n} k\|\mathbf{u}_h^l\|_{H^1}^2 \leq c_1\left(\mathrm{Re}\|\mathbf{u}_h^0\|_0^2 + \frac{1}{2a}\|\tau_{A,h}^0\|_{L^1}\right) + 2C_2^n t^n, \tag{3.3.28}$$

*where* $C_2^n = \dfrac{d}{2a}\max_{0\leq l\leq n}\|\beta^l\|_{L^1}$ *with* $c_1$ *and* $c_1$ *being generic constants.*

*Proof.* From (3.3.18) and (3.3.23), the following relation holds :

$$\begin{aligned} \left(\frac{1}{k} + \alpha^{n+1}\right)\tau_{A,h}^{n+1} &= \frac{1}{k}\left(I - k(R_h^{n+1})_h\right)^{-1}\left(\tau_{A,h}^n \circ y^n\right)\left(I - k(R_h^{n+1})_h\right)^{-T} \\ &+ \beta^{n+1} I. \end{aligned} \tag{3.3.29}$$

To obtain the discrete energy estimate, we first multiply $\left(I - k(R_h^{n+1})_h\right)$ on the left and $\left(I - k(R_h^{n+1})_h^T\right)$ on the right of the equation (3.3.29) respectively and rearrange various terms appropriately. Finally, by taking the trace operator followed by taking the integration, we obtain that

$$\begin{aligned} &(1 + k\alpha^{n+1})\int_\Omega \mathrm{tr}\left((R_h^{n+1})_h \tau_{A,h}^{n+1} + \tau_{A,h}^{n+1}(R_h^{n+1})_h^T\right) dx \\ =\ &\left(\frac{1}{k} + \alpha^{n+1}\right)\|\tau_{A,h}^{n+1}\|_{L^1} - \frac{1}{k}\|\tau_{A,h}^n \circ y^n\|_{L^1} \\ +\ &k(1 + k\alpha^{n+1})\int_\Omega \mathrm{tr}\left((R_h^{n+1})_h\left(\tau_{A,h}^{n+1} - \frac{k}{1 + k\alpha^{n+1}}\beta^{n+1} I\right)(R_h^{n+1})_h^T\right) dx \\ -\ &d\int_\Omega \beta^{n+1}\, dx + k\int_\Omega \beta^{n+1}\mathrm{tr}\left((R_h^{n+1})_h + (R_h^{n+1})_h^T\right) dx. \end{aligned} \tag{3.3.30}$$

We note that from (3.3.29), $\tau_{A,h}^{n+1}$ has the following lower bounds :

$$\tau_{A,h}^{n+1} \geq \frac{k}{1+k\alpha^{n+1}}\beta^{n+1} I \quad \forall n \geq 1. \tag{3.3.31}$$

Further, by the Lemma 3.3.2,

$$\begin{aligned} \int_\Omega \beta^{n+1}\mathrm{tr}\left((R_h^{n+1})_h + (R_h^{n+1})_h^T\right) dx &= \int_\Omega \beta^{n+1}\mathrm{tr}\left(R_h^{n+1} + (R_h^{n+1})^T\right) dx \\ &= 2a\int_\Omega \beta^{n+1}\mathbf{div}\mathbf{u}_h^{n+1}\, dx = 0 \end{aligned} \tag{3.3.32}$$

and

$$\begin{aligned} \int_\Omega \mathrm{tr}\left((R_h^{n+1})_h \tau_{A,h}^{n+1}\right) dx &= \int_\Omega \mathrm{tr}\left(\tau_{A,h}^{n+1}(R_h^{n+1})_h^T\right) dx \\ &= a\left(\tau_{A,h}^{n+1} : (\mathcal{D}_h(\mathbf{u}_h^{n+1}))_h\right) \\ &= a\left(\tau_{A,h}^{n+1} : \mathcal{D}(\mathbf{u}_h^{n+1})\right). \end{aligned} \tag{3.3.33}$$

Finally, based on the volume preserving property of $y^n$, we have

$$\|\tau_{A,h}^n \circ y^n\|_{L^1} = \|\tau_{A,h}^n\|_{L^1}. \tag{3.3.34}$$

Taking into account (3.3.31), (3.3.32), (3.3.33) and (3.3.34), from (3.3.30), we obtain the following inequality :

$$\begin{aligned} \left(\tau_{A,h}^{n+1} : \mathcal{D}(\mathbf{u}_h^{n+1})\right) &\geq \gamma\left(\frac{1}{k} + c\right)\|\tau_{A,h}^{n+1}\|_{L^1} \\ &- \frac{\gamma}{k}\|\tau_{A,h}^n\|_{L^1} - \gamma d\int_\Omega \beta^{n+1}\, dx, \end{aligned} \tag{3.3.35}$$

where $\gamma = \dfrac{1}{2a(1+kc)}$ with $c = \min\limits_{n\geq 0, x\in\Omega} \alpha^n(x)$. We now consider the momentum equation (3.3.17) together with (3.3.35) to obtain that

$$\begin{aligned}
\frac{\mathrm{Re}}{k}\|\mathbf{u}_h^{n+1}\|_0^2 + \eta_s\|\mathbf{u}_h^{n+1}\|_1^2 &= \frac{\mathrm{Re}}{k}\left(\Pi_h^{\mathbf{V}}(\mathbf{u}_h^n \circ y^n), \mathbf{u}_h^{n+1}\right) - \left(\tau_{A,h}^{n+1} : \mathcal{D}(\mathbf{u}_h^{n+1})\right) \\
&\leq \frac{\mathrm{Re}}{k}\left(\mathbf{u}_h^n \circ y^n, \mathbf{u}_h^{n+1}\right) - \gamma\left(\frac{1}{k} + c\right)\|\tau_{A,h}^{n+1}\|_{L^1} \\
&+ \frac{\gamma}{k}\|\tau_{A,h}^n\|_{L^1} + \gamma d\int_\Omega \beta^{n+1}\,dx.
\end{aligned}$$

Applying the Cauchy Schwarz inequality and the standard kick-back argument, we obtain the following relation :

$$\begin{aligned}
\frac{\mathrm{Re}}{2k}\|\mathbf{u}_h^{n+1}\|_{L^2}^2 &+ \eta_s\|\mathbf{u}_h^{n+1}\|_{H^1}^2 + \gamma\left(\frac{1}{k} + c\right)\|\tau_{A,h}^{n+1}\|_{L^1} \\
&\leq \frac{\mathrm{Re}}{2k}\|\mathbf{u}_h^n\|_{L^2}^2 + \frac{\gamma}{k}\|\tau_{A,h}^n\|_{L^1} + \gamma d\int_\Omega \beta^{n+1}\,dx.
\end{aligned} \tag{3.3.36}$$

We shall now show the first estimate (3.3.27). Multiplying $k$ by both sides of (3.3.36) and using the Korn's inequality, we obtain that

$$\begin{aligned}
\kappa_1\|\mathbf{u}_h^{n+1}\|_{L^2}^2 &+ \kappa_2\|\tau_{A,h}^{n+1}\|_{L^1} \\
&\leq \mathrm{Re}\|\mathbf{u}_h^n\|_{L^2}^2 + \gamma\|\tau_{A,h}^n\|_{L^1} + k\gamma d\int_\Omega \beta^{n+1}\,dx \\
&\leq \exp\left(-C_1 k\right)\left(\kappa_1\|\mathbf{u}_h^n\|_{L^2}^2 + \kappa_2\|\tau_{A,h}^n\|_{L^1}\right) \\
&+ k\gamma d\int_\Omega \beta^{n+1}\,dx,
\end{aligned} \tag{3.3.37}$$

where $\kappa_1 = \mathrm{Re} + k\eta_s c_\Omega$, $\kappa_2 = \gamma(1+ck)$, $c_\Omega$ is a positive constant depending only on $\Omega$ and $C_1 > 0$ is a constant given by

$$\max\left(\frac{\mathrm{Re}}{\mathrm{Re}+k\eta_s c_\Omega}, \frac{1}{1+ck}\right) \leq \exp\left(-C_1 k\right), 0 \leq k \leq 1. \tag{3.3.38}$$

Now, we use the induction argument to obtain :

$$\begin{aligned}
\kappa_1\|\mathbf{u}_h^n\|_0^2 &+ \kappa_2\|\tau_{A,h}^n\|_{L^1} \\
&\leq \exp(-C_1 t^n)\left(\kappa_1\|\mathbf{u}_h^0\|_0^2 + \kappa_2\|\tau_{A,h}^0\|_{L^1}\right) \qquad (3.3.39)\\
&+ kd\gamma\int_\Omega \beta^{n+1}\,dx\sum_{l=0}^{n}\exp(-C_1 t^l) \\
&\leq \exp(-C_1 t^n)\left(\kappa_1\|\mathbf{u}_h^0\|_{L^2}^2 + \kappa_2\|\tau_{A,h}^0\|_{L^1}\right) + \tilde{C}_2^n, \qquad (3.3.40)
\end{aligned}$$

where

$$\tilde{C}_2^n = k\frac{d}{2a}\int_\Omega \beta^n\,dx\left(\frac{1-\exp(-C_1 t^n)}{1-\exp(-C_1 k)}\right). \tag{3.3.41}$$

It is easy to see that we can choose $c_1$ which depends only on Re and $\gamma$ and a generic constant $c_2$ such that

$$\kappa_1\|\mathbf{u}_h^0\|_0^2 + \kappa_2\|\tau_{A,h}^0\|_{L^1} \leq c_1\left(\mathrm{Re}\|\mathbf{u}_h^0\|^2 + \frac{1}{2a}\|\tau_{A,h}^0\|_{L^1}\right) \quad \text{and} \quad \tilde{C}_2^n \leq c_2 C_2^n, \tag{3.3.42}$$

where $C_2^n = \dfrac{d}{2a}\max\limits_{0\leq l\leq n}\|\beta(\cdot,t^l)\|_{L^1}$. We then obtain the desired result (3.3.27). We now drive the other estimate (3.3.28). First we multiply $2k$ to both sides of (3.3.36) and take

summation from $l = 1$ to $l = n$ for both sides to obtain :

$$\begin{aligned} 2\eta_s \sum_{l=1}^{n} k\|\mathbf{u}_h^l\|_{H^1}^2 &\leq c_1 \left( \mathrm{Re}\|\mathbf{u}_h^0\|_{L^2}^2 + \frac{1}{2a}\|\tau_{A,h}^0\|_{L^1} \right) + \sum_{l=1}^{n} 2k\frac{d}{2a}\int_\Omega \beta^l \, dx, \\ &\leq c_1 \left( \mathrm{Re}\|\mathbf{u}_h^0\|_{L^2}^2 + \frac{1}{2a}\|\tau_{A,h}^0\|_{L^1} \right) + 2C_2^n t^n. \end{aligned}$$

This completes the proof. □

We shall now consider the limiting case when $\mathrm{We} = \infty$. Especially for models $\alpha = \beta = 0$ in such a case.

**Corollary 3.3.2.** *Assume that* $\mathrm{We} = \infty$ *and* $\alpha = \beta = 0$. *Then the following estimates hold true :*

$$\mathrm{Re}\|\mathbf{u}_h^n\|_{L^2}^2 + \|\tau_h^n\|_{L^1} \leq \mathrm{Re}\|\mathbf{u}_h^0\|_{L^2}^2 + \|\tau_h^0\|_{L^1} \quad \forall\, n \geq 1.$$

*and*

$$\eta_s \sum_{l=0}^{n} k\|\nabla \mathbf{u}_h^l\|_0^2 \leq \mathrm{Re}\|\mathbf{u}_h^0\|_0^2 + \|\tau_h^0\|_{L^1} \tag{3.3.43}$$

*Proof.* Note that in the limiting case, $\alpha = \beta = 0$ and $\tau_h^n$ is itself a conformation tensor for $n \geq 0$. The result then immediately follows from two estimate (3.3.27) and (3.3.28) since $C_1 = 0$ and $C_2^n = 0$ for all $n \geq 0$. This completes the proof. □

We would like to remark that in general, the algorithm is expected to have the better stability than what is seen in the above analysis for example (3.3.27). Let us look closely at the Theorem 3.3.2, where the lower bound for the conformation tensor (3.3.31) has been crucially used to drive the inequality (3.3.36). However, since the positivity

is added from the positivity of the previous step conformation tensor, the lower bound (3.3.31) is not sharp. Hence in general, our new discretizations lead to more stable schemes than what has been presented. In summary, apparently, the stability analysis indicates that keeping the positivity of the conformation tensor in the discrete level is crucial in the numerical stability and that the so-called "the high Weissenberg number problem" will not be presented in our scheme as is the case for currently available schemes in the literatures. This issue shall be further presented in the section 3.4 that discusses the global existence and uniqueness of discrete solution. Based on the new frameworks in this section and §3.4, we then confirmed the common belief affirmatively that the positivity preserving scheme shall allow to overcome the difficulty arising in simulating the time dependent non-Newtonian fluid flows. Furthermore, we also have extended the answer to other models.

### 3.3.3 Volume preserving scheme for computations of characteristic feet

The stability analysis indicates that algorithmic developments for the volume preserving schemes in the discrete sense are crucial. In this section, we shall elaborate that the volume preserving scheme can be devised for both two and three space dimensions ($d = 2$ or 3) and their algorithmic details.

Under the assumption that the volume preserving scheme $y^n$ is used, we obtain that for $\Pi_h^{\mathbf{V}} = Q_h$, the $L^2$ projection, we have

$$\|\Pi_h^{\mathbf{V}} \left(\mathbf{u}_h^n \circ y^n\right) \|_0 \leq \|\mathbf{u}_h^n \circ y^n\|_0 = \|\mathbf{u}_h^n\|_0. \tag{3.3.44}$$

For the stress field, a simple nodal value interpolant $\Pi_h^{\mathbf{S}} = I_h$ will result in

$$\|\Pi_h^{\mathbf{S}} \left(\tau_{A,h}^n \circ y^n\right) \|_{L^1} = \|\tau_{A,h}^n\|_{L^1}. \tag{3.3.45}$$

We would like to remark that from the stability analysis performed in the previous section, it is not so apparent that a computational realization of preserving $\det(E_h) = 1$ is crucial. Hence, we shall restrict our concern only on how to integrate the following ordinary differential equation for the computation of the characteristic feet.

$$\begin{aligned} \frac{d}{ds} y(x,t,s) &= \mathbf{u}(y(x,t,s),s), \\ y(x,t,t) &= x. \end{aligned} \tag{3.3.46}$$

where

$$\mathbf{divu} = 0. \tag{3.3.47}$$

The equation (3.3.46) is often called the source-free dynamical systems due to (3.3.47). For such a system, the solution $y(x,t,s) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is often called the flow map or phase flow and it has the following property.

$$\det\left(\frac{\partial y(x,t,s)}{\partial x}\right) = 1, \quad \forall x \in \mathbb{R}^d, \ \forall s \in \mathbb{R}. \tag{3.3.48}$$

We shall begin with introducing some popular scheme (see (3.3.52) below) to compute (3.3.46) and showing that (3.3.52) is indeed volume-preserving scheme for $d = 2$

but not for $d = 3$. We will then introduce some volume preserving scheme to solve (3.3.46) for $d = 3$, which is due to Feng and Shang in 1995, [28]. We remark that the scheme developed by Feng and Shang [28] does not seem to be applied in simulating incompressible flows.

In literatures, the following second order numerical scheme for solving (3.3.46) seems to be most popular and it seems to first appear in [83].

First, we integrate the equation (3.3.46) using the mid-point rule to obtain :

$$\frac{1}{k}\left(x - y(x,t,s)\right) = \frac{1}{2}\,\mathbf{u}\left(y\left(x,t,s+\frac{k}{2}\right), s+\frac{k}{2}\right) + O(k^2), \tag{3.3.49}$$

where $k = t - s$.

Second, the right hand side is approximated by a second order accurate extrapolation. Namely,

$$\mathbf{u}\left(x, s+\frac{k}{2}\right) = \frac{3}{2}\mathbf{u}(x,s) - \frac{1}{2}\mathbf{u}(x,s-k) + O(k^2). \tag{3.3.50}$$

The following approximation shall also be used :

$$y\left(x,t,s+\frac{k}{2}\right) = \frac{x + y(x,t,s)}{2} + O(k^2). \tag{3.3.51}$$

For notational conveniences, let us denote $y^n = y(x,t,s)$ and $y^{n+\frac{1}{2}} = (x+y^n)/2$. Hence we have the following implicit approximations :

$$\frac{1}{k}(x-y^n) := \frac{1}{2}\left(\frac{3}{2}\mathbf{u}(y^{n+\frac{1}{2}},s) - \frac{1}{2}\mathbf{u}(y^{n+\frac{1}{2}},s-k)\right). \tag{3.3.52}$$

To see that the mid-point rule (3.3.52) results in the volume-preserving scheme, let us take the derivative with respect to $x$ for both sides of (3.3.52). We then obtain the following :

$$\frac{1}{k}\left(I - F(s)\right) = A(y^{n+\frac{1}{2}})\left(I + F(s)\right), \tag{3.3.53}$$

where

$$A(y^{n+\frac{1}{2}}) = \frac{1}{2}\left(\frac{3}{2}\nabla\mathbf{u}(y^{n+\frac{1}{2}},s) - \frac{1}{2}\nabla\mathbf{u}(y^{n+\frac{1}{2}},s-k)\right) \tag{3.3.54}$$

and

$$F(s) = \frac{\partial y^n}{\partial x}. \tag{3.3.55}$$

We solve (3.3.53) for $F(s)$ and obtain that

$$F(s) = \left(I + \frac{k}{2}A(y^{n+\frac{1}{2}})\right)^{-1}\left(I - \frac{k}{2}A(y^{n+\frac{1}{2}})\right) \tag{3.3.56}$$

Under the assumption that some appropriate finite element space for the velocity field is used so that the divergence free condition of the velocity field is imposed in the discrete sense, we have

$$\mathrm{tr}A(y^{n+\frac{1}{2}}) = 0. \tag{3.3.57}$$

From (3.3.57), we conclude that $\det F(s) = 1$ identically.

The main reason why such an algorithm is popular seems that it has the volume preserving property. On the other hand, it is easy to see that the algorithm may not result in the volume-preserving scheme for $d = 3$. Note that for $d = 3$, under the assumption that $\mathrm{tr}A = 0$, for $H$ given as follows,

$$H = \left(I + \frac{k}{2}A\right)^{-1}\left(I - \frac{k}{2}A\right) \tag{3.3.58}$$

we have $\det H = 1 \Leftrightarrow \det A = 0$.

Our purpose here is that by reviewing the volume-preserving scheme in three dimension developed by Feng and Shang in [28], we wish to make sure such a volume preserving scheme can be devised in three dimension and confirm our numerical scheme can be implemented. As far as the author is concerned, such a special algorithmic detail has not been implemented in the context of the semi-Lagrangian scheme.

The basic idea of constructing the volume preserving scheme for $d = 3$ is based upon the following observation. Following the idea of H. Weyl, we have :

$$\mathbf{u}(y(x,t,s),s) = \begin{pmatrix} 0 \\ \dfrac{\partial v_1}{\partial y_3} \\ -\dfrac{\partial v_1}{\partial y_2} \end{pmatrix} + \begin{pmatrix} \dfrac{\partial v_2}{\partial y_2} \\ -\dfrac{\partial v_2}{\partial y_1} \\ 0 \end{pmatrix}, \tag{3.3.59}$$

where

$$v_1 = -\int_{y_3}^{x_3} \left( u_2(y_1, y_2, w, s) + \frac{\partial v_2}{\partial y_1}(y_1, y_2, w, s) \right) dw \tag{3.3.60}$$

$$+ \int_{y_2}^{x_2} u_3(y_1, w, x_3, s)\, dw \quad \text{and} \tag{3.3.61}$$

$$\tag{3.3.62}$$

$$v_2 = -\int_{y_2}^{x_2} u_1(y_1, w, y_3, s)\, dw$$

The actual expressions for $\frac{\partial v_1}{\partial y_2}$, $\frac{\partial v_1}{\partial y_3}$, $\frac{\partial v_2}{\partial y_2}$ and $\frac{\partial v_2}{\partial y_1}$ as follows :

$$\frac{\partial v_1}{\partial y_2} = -u_3(y_1, y_2, y_3, s) \tag{3.3.63}$$

$$\frac{\partial v_1}{\partial y_3} = u_2(y_1, y_2, y_3, s) - \int_{y_2}^{x_2} \frac{\partial u_1}{\partial y_1}(y_1, w, y_3, s)\, dw \tag{3.3.64}$$

$$\frac{\partial v_2}{\partial y_2} = u_1(y_1, y_2, y_3, s) \tag{3.3.65}$$

$$\frac{\partial v_2}{\partial y_1} = -\int_{y_2}^{x_2} \frac{\partial u_1}{\partial y_1}(y_1, w, y_3, s)\, dw. \tag{3.3.66}$$

From this, it is easy to see that (3.3.59) holds and by construction $\mathbf{u}^1$ and $\mathbf{u}^2$ given as follows are divergence free :

$$\mathbf{u}^1 = \begin{pmatrix} 0 \\ \frac{\partial v_1}{\partial y_3} \\ -\frac{\partial v_1}{\partial y_2} \end{pmatrix} \quad \text{and} \tag{3.3.67}$$

$$\mathbf{u}^2 = \begin{pmatrix} \frac{\partial v_2}{\partial y_2} \\ -\frac{\partial v_2}{\partial y_1} \\ 0 \end{pmatrix} \tag{3.3.68}$$

Let us now denote $S_i^k$ by the volume preserving scheme for

$$\frac{d}{ds} y(x,t,s) = \mathbf{u}^i(y(x,t,s),s), \tag{3.3.69}$$

$$y(x,t,s) = x \tag{3.3.70}$$

with $i = 1, 2$, then the following composition is trivially volume preserving :

$$y^n = S_2^k \circ S_1^k \tag{3.3.71}$$

Moreover, assuming $S_i^k$ is of second order accurate, it is easy to see that the composition (3.3.71) is of second order. The above idea on the composition is from Feng and Shang, [28].

## 3.4 On the global existence and uniqueness of discrete solutions

In this section, based on the stability results obtained in the previous section, we shall prove the main result in this chapter, the global existence and uniqueness of discrete solutions by taking a specific numerical discretization scheme (3.2.21), (3.2.22) and (3.2.23). We also introduce some implementation details from which the actual computations can be performed. Of course, such an implementation should be practical or at least doable in practice.

In this section and also in the following chapters, we shall often use the following notation $\lesssim$ and $\gtrsim$ to avoid keeping writing generic constants, namely, When we write

$$x_1 \lesssim y_1 \quad \text{and} \quad x_2 \gtrsim y_2, \tag{3.4.1}$$

then there exist constants $c_1$ and $c_2$ such that

$$x_1 \leq c_1 y_1 \quad \text{and} \quad x_2 \geq c_2 y_2. \tag{3.4.2}$$

We shall begin this section by writing some specific algorithms applicable in actual computations.

**Algorithm 3.4.1.** *Assume that* $\mathbf{u}_h^n$ *and* $\tau_{A,h}^n$ *have been defined.*

*To proceed to* $\mathbf{u}_h^{n+1}$ *and* $\tau_{A,h}^{n+1}$*, we shall perform the following iterations denoted by* $\left\{\mathbf{u}_h^{n+1,\ell}, \tau_{A,h}^{n+1,\ell}\right\}_{\ell=1}^{\infty}$*, with* $\mathbf{u}_h^{n+1,0} = \mathbf{u}_h^n$ *and* $\tau_{A,h}^{n+1,0} = \tau_{A,h}^n$.

**Step One :**

*Solve the momentum equation and continuity equations :*

$$\begin{aligned} \mathrm{Re}\left(\frac{\mathbf{u}_h^{n+1,\ell} - \Pi_h^{\mathbf{V}}(\mathbf{u}_h^n \circ y^n)}{k}\right) + \nabla_h p_h^{n+1,\ell} - \eta_s \Delta_h \mathbf{u}_h^{n+1,\ell} &= \mathrm{div}_h \tau_{A,h}^{n+1,\ell-1} \\ \mathbf{divu}_h^{n+1,\ell} &= 0 \end{aligned}$$

**Step Two :**

*Compute the deformation tensor* $E_h(t^n, t^{n+1,\ell})$.

$$E_h(t^n, t^{n+1,\ell}) = (I - k(R_h^{n+1,\ell})_h)^{-1}, \tag{3.4.3}$$

*where*

$$(R_h^{n+1,\ell})_h = \sum_{i=1}^{N_h} \frac{1}{|K_i|} \int_{K_i} \nabla_h \mathbf{u}_h^{n+1,\ell}\, dx. \tag{3.4.4}$$

**Step Three :**

*Update the stress* $\tau_{A,h}^{n+1,\ell}$ *based on* $E_h(t^n, t^{n+1,\ell})$ *:*

$$\begin{aligned} \left(\frac{1}{k} + \alpha^{n+1,\ell-1}\right)\tau_{A,h}^{n+1,\ell} &= \frac{1}{k} E_h(t^n, t^{n+1,\ell})(\tau_{A,h}^n \circ y^n) E_h(t^n, t^{n+1,\ell})^T \\ &+ \beta^{n+1,\ell-1} I. \end{aligned} \tag{3.4.5}$$

We observe that the main obstacle in developing algorithmic development described in (3.4.1) is in the computation of $E_h(t^n, t^{n+1,\ell})$. Namely, $I - k(R_h^{n+1,\ell})_h$ should be invertible. Hence the following assumption does not seem to be avoidable :

The time step size $k$ is sufficiently small so that $I - k(R_h^{n+1,\ell})_h$ is invertible. More precisely, we shall require the condition that

$$k \, \|(R_h^{n+1,\ell})_h\|_\infty < 1, \quad \forall \ell = 1, \cdots \tag{3.4.6}$$

Under the assumption (3.4.6), we shall have that

$$\left\| \left( I - k \left( R_h^{n+1,\ell} \right)_h \right)^{-1} \right\|_\infty \leq \frac{1}{1 - k \left\| \left( R_h^{n+1,\ell} \right)_h \right\|_\infty}. \tag{3.4.7}$$

Further note that

$$\begin{aligned} \left\| \left( R_h^{n+1,\ell} \right)_h \right\|_\infty &\leq \max_K \frac{1}{|K|} \int_K \left| R_h^{n+1,\ell} \right| \, dx \\ &\leq \max_K \left\| R_h^{n+1,\ell} \right\|_{0,K} \\ &\leq \max_K \left| \mathbf{u}_h^{n+1,\ell} \right|_{1,K}. \end{aligned} \tag{3.4.8}$$

For the following analysis, the step size $k$ shall not be too restrictive, namely, it should be enough that $k$ behaves like $O(h)$, where $h$ is the mesh size under the assumption that the triangulation is uniform. To be more specific, we invoke the following well-known local inverse inequality :

**Lemma 3.4.1.** *For $K \in \mathcal{T}_h$ and $\mathbf{u} \in \mathcal{P}^m$*

$$|\mathbf{u}|_{1,K} \lesssim h_K^{-1} \|\mathbf{u}\|_{0,K}, \tag{3.4.9}$$

*where $h_K$ is the mesh size of the element $K$.*

We then conclude from the Lemma 3.4.1 that

$$\left\| \left( R_h^{n+1,\ell} \right)_h \right\|_\infty \leq \max_K h_K^{-1} \left\| \mathbf{u}_h^{n+1,\ell} \right\|_{0,K}.$$

In general, our stability analysis does not guarantee the uniform boundedness of the gradient of velocity, so the stability for the scheme (3.2.21), (3.2.22) and (3.2.23) is dependent on the mesh size, namely conditional. On the other hand, from the uniform boundedness of $\|\mathbf{u}_h^n\|_0$ for all $n \geq 0$, we see that the stability condition is not too restrictive.

Since we use only the uniform boundedness of $L^2$ norm of the velocity field, $\eta_s$ is allowed to be zero. For simplicity of the argument in the following theorem, $\beta$ shall be assumed to be constant. Indeed, such contribution of non-linearity does not hinder rigorous analysis but adds some trivial complications. To ease our presentation, let us denote

$$\overline{\mathbf{u}}_h^{n+1,\ell+1} = \mathbf{u}_h^{n+1,\ell+1} - \mathbf{u}_h^{n+1,\ell}, \quad \overline{\tau}_h^{n+1,\ell+1} = \tau_h^{n+1,\ell+1} - \tau_h^{n+1,\ell}, \tag{3.4.10}$$

$$\overline{E}_h^{n+1,\ell+1} = E_h(t^n, t^{n+1,\ell+1}) - E_h(t^n, t^{n+1,\ell}), \tag{3.4.11}$$

and

$$C_\ell := \left\| \left( I - k \left( R_h^{n+1,\ell} \right)_h \right)^{-1} \right\|_\infty, \tag{3.4.12}$$

where $\ell = 0, \cdots,$.

**Theorem 3.4.1.** *Assume that $k$ is sufficiently small, then the discrete problem (3.2.21), (3.2.22) and (3.2.23) produces a unique global discrete solution.*

*Proof.* We shall define the procedure (3.2.21), (3.2.22) and (3.2.23) as an operator $\Phi$ by the following :

$$\Phi(X_h^{n+1,\ell}) = X_h^{n+1,\ell+1}, \tag{3.4.13}$$

where

$$X_h^{n+1,\ell} = \left(\mathbf{u}_h^{n+1,\ell}, \tau_{A,h}^{n+1,\ell}\right) \quad \text{with} \quad X_h^{n+1,0} = \left(\mathbf{u}_h^{n}, \tau_{A,h}^{n}\right). \tag{3.4.14}$$

It is then enough to show that $\Phi$ is contractive for some appropriate norm $\|\cdot\|$ for a sufficiently small $k$. Namely,

$$\|\Phi(X_h^{n+1,\ell}) - \Phi(X_h^{n+1,\ell-1})\| \leq c\|X_h^{n+1,\ell} - X_h^{n+1,\ell-1}\|, \tag{3.4.15}$$

where $c$ is some constant less than 1. This shall result that as $\ell \to \infty$,

$$\mathbf{u}_h^{n+1,\ell} \longrightarrow \mathbf{u}_h^{n+1} \quad \text{in } L^2 \tag{3.4.16}$$

and

$$\tau_{A,h}^{n+1,\ell} \longrightarrow \tau_{A,h}^{n+1} \quad \text{in } L^1 \tag{3.4.17}$$

We shall first consider the momentum equation. By subtracting the equation for $\mathbf{u}_h^{n+1,\ell}$ from the equation for $\mathbf{u}_h^{n+1,\ell+1}$, we obtain that

$$\begin{aligned}\frac{\mathrm{Re}}{k}\|\overline{\mathbf{u}}_h^{n+1,\ell+1}\|_0^2 &+ \eta_s|\overline{\mathbf{u}}_h^{n+1,\ell+1}|_{1,\Omega}^2 \\ &\leq \left\|\left(\mathcal{D}(\overline{\mathbf{u}}_h^{n+1,\ell+1})\right)_h\right\|_\infty \sum_{i,j=1}^d \left\|\left(\overline{\tau}_{A,h}^{n+1,\ell}\right)_{ij}\right\|_{L^1(\Omega)} \\ &\leq \left\|\left(\nabla(\overline{\mathbf{u}}_h^{n+1,\ell+1})\right)_h\right\|_\infty \sum_{i,j=1}^d \left\|\left(\overline{\tau}_{A,h}^{n+1,\ell}\right)_{ij}\right\|_{L^1(\Omega)}\end{aligned} \tag{3.4.18}$$

From this and the Lemma 3.4.1, we have,

$$\|\overline{\mathbf{u}}_h^{n+1,\ell+1}\|_0^2 \lesssim k \max_{K\in\mathcal{T}_h} h_K^{-1}\|\overline{\mathbf{u}}_h^{n+1,\ell+1}\|_{0,K} \sum_{i,j=1}^d \left\|\left(\overline{\tau}_{A,h}\right)_{ij}\right\|_{L^1(\Omega)}. \tag{3.4.19}$$

For a sufficiently small $k$, we have that

$$\|\overline{\mathbf{u}}_h^{n+1,\ell+1}\|_0 \lesssim \sum_{i,j=1}^d \left\|\left(\overline{\tau}_{A,h}^{n+1,\ell}\right)_{ij}\right\|_{L^1(\Omega)}, \tag{3.4.20}$$

Now we shall consider the constitutive equation. By subtracting the equation for $\tau_{A,h}^{n+1,\ell-1}$ from the equation for $\tau_{A,h}^{n+1,\ell}$, we obtain that

$$\begin{aligned}\left(\frac{1}{k}+\alpha^{n+1,\ell}\right)\sum_{i,j=1}^d \left\|\left(\overline{\tau}_{A,h}^{n+1,\ell}\right)_{ij}\right\|_{L^1(\Omega)} &\lesssim \frac{1}{k}\|\overline{E}_h^\ell\|_\infty\|E_h(t^n,t^{n+1,\ell})\|_\infty \\ &+ \frac{1}{k}\|E_h(t^n,t^{n+1,\ell-1})\|_\infty\|\overline{E}_h^{\ell-1}\|_\infty,\end{aligned}$$

where the following result is used :

$$\|\tau_{A,h}^{n}\|_{L^1(\Omega)} \lesssim 1. \tag{3.4.21}$$

Further, from the assumption that $\alpha^n \geq c > 0$, we conclude that

$$\begin{aligned} \sum_{i,j=1}^{d} \left\| \left(\overline{\tau}_{A,h}^{n+1,\ell}\right)_{ij} \right\|_{L^1(\Omega)} &\lesssim \|\overline{E}_h^{\ell}\|_{\infty} \|E_h(t^n, t^{n+1,\ell})\|_{\infty} \\ &+ \|E_h(t^n, t^{n+1,\ell-1})\|_{\infty} \|\overline{E}_h^{\ell-1}\|_{\infty}. \end{aligned} \tag{3.4.22}$$

Note that for any matrices $A$ and $B$,

$$A^{-1} - B^{-1} = A^{-1}\,(B - A)\,B^{-1}, \tag{3.4.23}$$

from which, we have

$$\begin{aligned} \|\overline{E}_h(t^n, t^{n+1,\ell})\|_{\infty} &\lesssim k C_{\ell} C_{\ell-1} \|\nabla \overline{\mathbf{u}}_h^{n+1,\ell}\|_{0,K} \\ &\lesssim k h_K^{-1} C_{\ell} C_{\ell-1} \|\overline{\mathbf{u}}_h^{n+1,\ell}\|_0. \end{aligned} \tag{3.4.24}$$

Since $C_\ell$ and $C_{\ell-1}$ are bounded, one can choose $k$ appropriately so that

$$\sum_{i,j=1}^{d} \left\| \left(\overline{\tau}_{A,h}^{n+1,\ell}\right)_{ij} \right\|_{L^1(\Omega)} \leq c \|\overline{\mathbf{u}}_h^{n+1,\ell}\|_0, \tag{3.4.25}$$

where $c < 1$. Further such $c$ can be chosen so that

$$\|\overline{\mathbf{u}}_h^{n+1,\ell+1}\|_0 \lesssim \sum_{i,j=1}^{d} \left\| \left(\overline{\tau}_{A,h}^{n+1,\ell}\right)_{ij} \right\|_{L^1(\Omega)} \leq c\|\overline{\mathbf{u}}_h^{n+1,\ell}\|_0. \tag{3.4.26}$$

This shows that $\Phi$ is contractive with respect to $\|\cdot\|$ and completes the global existence of discrete solution. The uniqueness follows from the standard argument. This completes the proof. $\square$

# Chapter 4

# Efficient iterative techniques: multigrid method and preconditioned MINRES method

## 4.1 Introduction

The main purpose of this chapter is to discuss the solution techniques for the resulting discrete systems of the viscoelastic models obtained by applying the discretization schemes introduced in the chapter 3.

This chapter is motivated and devoted to construct robust and efficient iterative techniques and much will be discussed on the new framework and mathematical foundations for the convergence rate of the method of successive subspace corrections for the singular problems. This chapter has been discussed mostly following the new paper in review by Lee, Wu, Xu and Zikatanov, [55]. The main reason why we are concerned with the study on the singular problem is because an efficient multigrid method for the Laplace equation with pure Neumann boundary condition will be crucial in developing the efficient solver for viscoelastic flows (see §4.3 below for a detailed discussion on this issue).

Recall that the discretization techniques proposed in the previous chapter is based upon the semi-Lagrangian scheme for the reformulated constitutive equations. Such a reformulation proves to be important for the stability of the discrete system. Although, such a discussion leads us to be more interested in solving the viscoelastic systems of

the reformulated form. It is also noteworthy that the application of the semi-Lagrangian scheme without reformulating the constitutive equation leads to the following system of equation as an operator form :

$$\begin{pmatrix} -\frac{1}{k}\frac{\mathrm{We}}{2\mu_p}I & \mathcal{D} & 0 \\ -\mathrm{div} & \frac{\mathrm{Re}}{k}I - \eta_s\Delta & \nabla \\ 0 & -\mathbf{div} & 0 \end{pmatrix} \begin{pmatrix} \tau \\ \mathbf{u} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{h} \\ \mathbf{f} \\ \mathbf{g} \end{pmatrix}, \tag{4.1.1}$$

where $k$ is the time step size, $\mathbf{h}, \mathbf{f}$ and $\mathbf{g}$ are source terms. This is nothing else than the three field Stokes equation. This may be considered as an additional contribution of the current thesis work for the numerical simulation of viscoelastic models. That is, we have shown that the complicated viscoelastic model can be reduced into the three field Stokes equation by the semi-Lagrangian scheme based on the Lie derivative rather than the classical material derivative, [68].

The discretization of the reformulated equations (see the Algorithm (3.4.1)) require us to solve rather the following system of equations :

$$\begin{pmatrix} \frac{\mathrm{Re}}{k}I - \eta_s\Delta & \nabla \\ -\mathbf{div} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}. \tag{4.1.2}$$

This chapter is in general devoted to construct efficient iterative solvers for systems like (4.1.1) and (4.1.2). Although, we shall be rather focused on the latter system in that we are more concerned with the stable discretizations and its resulting discrete

systems, it is straightforward to see that the efficient method for the system (4.1.2) can be also immediately used for the system given in (4.1.1).

There are many iterative solvers available for the system of equations (4.1.1), such as the Uzawa-type methods [93, 90, 82] and the Krylov subspace methods, [93, 90, 52, 75] like the GMRES or MINRES. However, plain applications of such methods can not be efficient and preconditioners play an important role to make fast and robust solvers and how to construct "good" preconditioner is still a difficult problem for both numerically and mathematically and it is still an active research area. We wish to identify the problem of efficient preconditioning and try to handle such a problem in a quite general framework. Our approach to solve the resulting discrete saddle problem (4.1.1) is based on the Krylov space method. Especially, we shall apply the preconditioned MINRES method with the preconditioner being a block diagonal.

## 4.2 Analysis of an abstract variational problem

In this paragraph, we shall review an abstract framework well adapted to solution of a variety of linear boundary value problems with a constraint, such as the Stokes problem like the aforementioned system (4.1.2). The following discussion are mostly based on the monograph by Girault and Raviart, [30] and will be revisited for the convergence analysis of successive subspace corrections for singular systems below.

Let $V$ and $W$ be two (real) Hilbert spaces with norms, $\|\cdot\|_V$ and $\|\cdot\|_W$ respectively. Let $V^*$ and $W^*$ be their corresponding dual spaces and let $\|\cdot\|_{V^*}$ and $\|\cdot\|_{W^*}$ denote their dual norms. As usual, we denote by $\langle\cdot,\cdot\rangle$ the duality pairing between the spaces $V$ and $V^*$ or $W$ and $W^*$.

We introduce two bilinear continuous forms:

$$m(\cdot,\cdot) \quad : \quad V \times V \mapsto \mathbb{R}; \quad m(u,v) \leq \|m\| \|u\|_V \|v\|_V, \quad \forall u, v \in V, \tag{4.2.1}$$

$$b(\cdot,\cdot) \quad : \quad V \times W \mapsto \mathbb{R}; \quad b(u,p) \leq \|b\| \|u\|_V \|p\|_W, \quad \forall u \in V, p \in W, \tag{4.2.2}$$

where

$$\|m\| = \sup_{u,v \in V, u,v \neq 0} \frac{m(u,v)}{\|u\|_V \|v\|_V}, \quad \|b\| = \sup_{u \in V, p \in W, u,p \neq 0} \frac{b(u,p)}{\|u\|_V \|p\|_W}. \tag{4.2.3}$$

Then we consider the following variational problem :

Given $f \in V^*$ and $\chi \in W^*$, find a pair $(u,p) \in V \times W$ such that

$$\begin{aligned} m(u,v) + b(v,p) &= \langle f, v \rangle, \quad \forall v \in V, \\ b(u,\mu) &= \langle \chi, \mu \rangle, \quad \forall \mu \in W. \end{aligned} \tag{4.2.4}$$

A special theory was developed by Brezzi [15] for this type of problems. We now discuss about this theory.

**Theorem 4.2.1 (Brezzi, 1974).** *The variational problem (4.2.4) is well-posed if and only if the following LBB-conditions hold*

$$\inf_{u \in \mathcal{N}} \sup_{v \in \mathcal{N}} \frac{m(u,v)}{\|u\|_V \|v\|_V} = \inf_{v \in \mathcal{N}} \sup_{u \in \mathcal{N}} \frac{m(u,v)}{\|u\|_V \|v\|_V} = \eta > 0, \tag{4.2.5}$$

*where* $\mathcal{N} = \{v \in V : b(v, \mu) = 0, \forall \mu \in W.\}$ *and*

$$\inf_{q \in W} \sup_{v \in V} \frac{b(v, q)}{\|v\|_V \|q\|_W} = \mu > 0. \tag{4.2.6}$$

This inf-sup condition is crucial especially for the error analysis. Moreover, for successful numerical simulations, a stable finite element pairs, say $V_h \subset V$ and $W_h \subset W$ should be chosen. Namely, by stable pairs, we mean that the inf-sup condition holds for pair of spaces $V_h \times W_h$ with corresponding constants $\eta_h$ and $\mu_h$ are generic, namely they are independent of $h$, [30, 82].

We would like to remark that although for classical formulations of Navier-Stokes equations, such a theory, the Theorem 4.2.1 is well-developed, the corresponding theory for the axisymmetric formulations is difficult to find. This issue has been worked out in the Appendix A.

In the remainder of this section, we shall study further the inf-sup condition (4.2.6) as it will be needed for the analysis of successive subspace corrections later in this chapter. With the bilinear forms $m(\cdot, \cdot)$ and $b(\cdot, \cdot)$, we associate two linear operators $M \in \mathcal{L}(V, V^*)$ and $B \in \mathcal{L}(V, W^*)$ defined by :

$$\langle Mu, v\rangle = m(u, v), \quad \forall u, v \in V, \tag{4.2.7}$$

$$\langle Bv, \mu\rangle = b(v, \mu), \quad \forall v \in V, \quad \forall \mu \in W. \tag{4.2.8}$$

Let $B^* \in \mathcal{L}(W, V^*)$ be the dual operator of $B$, namely,

$$\langle B^*\mu, v\rangle = \langle \mu, Bv\rangle = b(v, \mu) \quad \forall v \in V, \quad \forall \mu \in W. \tag{4.2.9}$$

The problem (4.2.4) may then be equivalently written in the form :

Find $(u, p) \in V \times W$ satisfying

$$\begin{aligned} Mu + B^* p &= f \quad \text{in } V^* \\ Bu &= \chi \quad \text{in } W^* \end{aligned} \tag{4.2.10}$$

We set

$$\mathcal{N} = \mathcal{N}(B), \tag{4.2.11}$$

where $\mathcal{N}(B)$ is the null space of the operator $B$ and more generally, for each $\chi \in W^*$, we define the affine manifold

$$\mathcal{N}(\chi) = \{v \in V | Bv = \chi\} \tag{4.2.12}$$

Equivalently, we have :

$$\mathcal{N}(\chi) = \{v \in V \,|\, b(v, \mu) = \langle \chi, \mu \rangle \quad \forall \mu \in W\}, \tag{4.2.13}$$

$$\mathcal{N} = \mathcal{N}(0). \tag{4.2.14}$$

We also define the polar set $\mathcal{N}^\circ$ of $\mathcal{N}$ by

$$\mathcal{N} = \{g \in V^* \,|\, \langle g, v \rangle = 0 \quad \forall v \in \mathcal{N}.\} \tag{4.2.15}$$

**Lemma 4.2.1.** *The following three properties are equivalent:*

*(i) there exists a constant $\beta > 0$ such that*

$$\inf_{u \in W} \sup_{v \in V} \frac{b(v,u)}{\|v\|_V \|u\|_W} \geq \beta; \tag{4.2.16}$$

*(ii) the operator $B^* : W \mapsto \mathcal{N}^\circ$ is an isomorphism and*

$$\|B^*\mu\|_{V^*} \geq \beta \|\mu\|_W \quad \forall \mu \in W; \tag{4.2.17}$$

*(iii) the operator $B : \mathcal{N}^\perp \mapsto W^*$ is an isomorphism and*

$$\|Bv\|_{W^*} \geq \beta \|v\|_V \quad \forall v \in \mathcal{N}^\perp. \tag{4.2.18}$$

*Proof.* Let us show the properties (i) and (ii) are equivalent. It is easy to note that

$$\|B^*\mu\|_{V^*} = \sup_{v \in V, v \neq 0} \frac{\langle B^*\mu, v \rangle}{\|v\|_V} \geq \beta \|\mu\|_W \quad \forall \mu \in W, \tag{4.2.19}$$

So, (4.2.16) and (4.2.17) are equivalent. Hence (ii) implies (i). To prove that (i) implies (ii), it remains only to show that, under the condition (4.2.17), $B^*$ is an isomorphism from $W$ onto $\mathcal{N}^\circ$. Clearly, from (4.2.17), $B^*$ is a one to one operator from $W$ onto its range $\mathcal{R}(B^*)$ with a continuous inverse. Hence $B^* : W \mapsto \mathcal{R}(B^*)$ is an isomorphism so that $\mathcal{R}(B^*)$ is a closed subspace of $V^*$. Thus it remains to prove

$$\mathcal{R}(B^*) = \mathcal{N}^\circ. \tag{4.2.20}$$

For this, we apply the Closed Range Theorem of Banach which asserts that

$$\mathcal{R}(B^*) = (\mathcal{N}(B))^\circ = \mathcal{N}^\circ. \tag{4.2.21}$$

This proves that (i) implies (ii).

We now show that (ii) is equivalent to (iii). First, we observe that $\mathcal{N}^\circ$ may be isometrically identified with $(\mathcal{N}^\perp)^*$. Namely there is an isometric bijection

$$i : (\mathcal{N}^\perp)^* \mapsto \mathcal{N}^\circ. \tag{4.2.22}$$

As a consequence, we have that $B : \mathcal{N}^\perp \mapsto W^*$ is an isomorphism if and only if $B^* : W \mapsto (\mathcal{N}^\perp)^* = \mathcal{N}^\circ$ is an isomorphism. Therefore properties (ii) and (iii) are equivalent. This completes the proof. □

## 4.3 Preconditioned Minimum Residual Method

In this section, we shall now make a brief review on the preconditioned minimum residual method together with its algorithmic detail and some mathematical fact on its convergence property.

For conveniences, we shall denote the system of equation (4.2.10) by the following :

$$\Upsilon \mathbf{p} = \mathbf{q}, \tag{4.3.1}$$

where

$$\Upsilon = \begin{pmatrix} M & B \\ B^* & 0 \end{pmatrix} \tag{4.3.2}$$

and

$$\mathbf{p} = \begin{pmatrix} u \\ p \end{pmatrix} \quad \text{and} \quad \mathbf{q} = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{4.3.3}$$

There are several version of preconditioned MINRES method. Especially, we shall consider a natural generalization of the preconditioned Conjugate Gradient (PCG) method, called the PCR method, [36, 52]. We shall consider the block diagonal preconditioner $\hat{\Upsilon}$ given by the following:

$$\hat{\Upsilon} = \begin{pmatrix} \hat{M} & 0 \\ 0 & \hat{S} \end{pmatrix}. \tag{4.3.4}$$

A stable version of preconditioned MINRES based on a three-term recurrence can be given as follows :

**Algorithm 4.3.1.** *Assume that an initial guess* $\mathbf{p}_0$ *is given.*

*Step 1. Initialization :*

$$\mathbf{r}_0 := \mathbf{q} - \Upsilon \mathbf{q}_0,$$

$$\mathbf{h}_{-1} := 0,$$

$$\mathbf{h}_0 := \hat{\Upsilon}^{-1} \mathbf{r}_0.$$

*Step 2. Iteration :*

$$\gamma := \frac{\left(\mathbf{r}_m, \hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_m\right)}{\left(\Upsilon\mathbf{h}_m, \hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_m\right)}$$

$$\mathbf{p}_{m+1} := \mathbf{p}_m + \gamma\mathbf{h}_m,$$

$$\mathbf{r}_{m+1} := \mathbf{r}_m - \gamma\Upsilon\mathbf{h}_m,$$

$$\varsigma_0 := \frac{\left(\hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_m, \Upsilon\hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_m\right)}{\left(\hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_m, \Upsilon\mathbf{h}_m\right)},$$

$$\varsigma_1 := \frac{\left(\hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_m, \Upsilon\hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_{m-1}\right)}{\left(\hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_{m-1}, \Upsilon\mathbf{h}_{m-1}\right)},$$

$$\mathbf{h}_{m+1} := \hat{\Upsilon}^{-1}\Upsilon\mathbf{h}_m - \varsigma_0\mathbf{h}_m - \varsigma_1\mathbf{h}_{m-1}.$$

We are in a position to discuss the convergence property of the Algorithm 4.3.1. Let us first define a condition number, $\kappa(\hat{\Upsilon}^{-1}\Upsilon)$ by

$$\kappa(\hat{\Upsilon}^{-1}\Upsilon) := \frac{\max\left\{|\lambda| : \lambda \in \sigma(\hat{\Upsilon}^{-1}\Upsilon)\right\}}{\min\left\{|\lambda| : \lambda \in \sigma(\hat{\Upsilon}^{-1}\Upsilon)\right\}}. \tag{4.3.5}$$

The convergence property can then be given in terms of $\kappa$ as follows :

**Theorem 4.3.1.** *Let $\Upsilon$ be symmetric, $\hat{\Upsilon}$ be non-negative definite and $\mathbf{p}$ be an exact solution to $\Upsilon\mathbf{p} = \mathbf{q}$. Then the m-th iterate $\mathbf{p}_m$ of the Algorithm 4.3.1 satisfies*

$$\|\hat{\Upsilon}^{-1/2}\Upsilon(\mathbf{p}_m - \mathbf{p})\|_0 \leq \frac{2c(\kappa)^{\varrho}}{1 + c(\kappa)^{2\varrho}}\|\hat{\Upsilon}^{-1/2}\Upsilon(\mathbf{p}_0 - \mathbf{q})\|_0, \tag{4.3.6}$$

*where $c(\kappa) = \dfrac{\kappa - 1}{\kappa + 1}$, $\kappa = \kappa(\hat{\Upsilon}^{-1}\Upsilon)$, and $\dfrac{m}{2} - 1 < \varrho \leq \dfrac{m}{2}, \quad \forall \varrho \in \mathbf{Z}$.*

From the convergence analysis contained in the Theorem 4.3.1, the robust and efficient preconditioner should make the condition number $\kappa \lesssim 1$. In the next section,

we shall discuss how to construct such a preconditioner for the system of equation we are concerned with.

### 4.3.1 On a schur complement operator arising from (non) Newtonian fluid flows simulations

In this section, as briefly mentioned in the previous section, we shall discuss how to construct an efficient solver for saddle point problems of the following form :

$$\begin{pmatrix} kI - \Delta_h & \nabla_h \\ -\mathbf{div} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_h \\ p_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ 0 \end{pmatrix}. \tag{4.3.7}$$

Here $k$ is related to the time step-size and Reynolds number, $\Delta_h$ and $\nabla_h$ are certain discrete Laplacian and gradient operators on a pair of stable finite element spaces $\mathbf{V}_h \times W_h$. An actual form of discrete system (4.1.2) can be recovered from (4.3.7) by some normalization. Author indeed finds that the formulation (4.3.7) is more stable to apply the preconditioned MINRES from the computational point of view.

Let us apply the block Gaussian elimination to (4.3.7) to obtain

$$\begin{pmatrix} kI - \Delta_h & \nabla_h \\ 0 & -\mathbf{div}(kI - \Delta_h)^{-1}\nabla_h \end{pmatrix} \begin{pmatrix} \mathbf{u}_h \\ p_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ -\mathbf{div}(kI - \Delta_h)^{-1}\mathbf{f}_h \end{pmatrix}. \tag{4.3.8}$$

From the (4.3.8) and in view of Theorem 4.3.1, we observe that in order to construct an efficient preconditioner for the MINRES method, we shall be needing to construct robust preconditioners for both

$$kI - \Delta_h \tag{4.3.9}$$

and

$$-\mathbf{div}(kI - \Delta_h)^{-1}\nabla_h \tag{4.3.10}$$

It is rather an easy task to construct an efficient preconditioner for an operator like $kI - \Delta_h$ since it is symmetric and positive definite. Especially, for unstructured grids, which is often the case for flow problems, many public codes are available for such elliptic operators although some theoretical analysis is still being developed. Note that in our numerical works, the AMG developed by Chang, Xu and Zikatanov, [18] is used. Assuming that a spectrally equivalent solver is available for an elliptic operator like (4.3.9), we shall restrict our concern only on how to handle the schur complement operator given in (4.3.10).

To find the behavior of an operator (4.3.10), the following result is intriguing, which is contained in [78] and [13].

**Theorem 4.3.2.** *The schur complement operator $kI - \Delta_h$ is spectrally equivalent to*

$$kI + Q_h(-\Delta_H)_N^{-1}Q_h^T, \tag{4.3.11}$$

*where $h$ is the spatial mesh size, $H = \sqrt{k}$, $Q_h : L^2(\Omega) \mapsto \mathbf{V}_h$ is the $L^2$ projection and $(-\Delta_H)_N$ is the Laplace operator with Neumann boundary condition.*

We note that the result is still true even when $(-\Delta_H)_N$ is simply replaced by $(-\Delta_h)_N$. We then conclude that it is crucial to study the singular problem like $(-\Delta_h)_N$, the Laplace problem with pure Neumann boundary condition to construct an efficient solver for the system (4.3.7). This is the topic to be handled in next section.

## 4.4 Multigrid analysis for singular systems

This section shall discuss the problem to construct efficient solvers for the Laplace equation with pure Neumann boundary condition as it is crucial to develop robust iterative solvers for viscoelastic flow problems as pointed out in the previous sections.

However, in this section, we shall provide some general mathematical foundation on the convergence analysis of the method of subspace corrections applied to singular problems in a Hilbert space setting.

Let us briefly review on the method of subspace corrections. The method of subspace corrections is an abstract linear iterative method. Many iterative techniques based on an old and simple strategy "divide and conquer" such as Jacobi method, Gauss-Seidel method, point or block relaxation methods also fall into this the category. The method of subspace corrections are largely divided into two categories, the *parallel* (Jacobi, block Jacobi, additive Schwarz methods) and *successive* (Gauss-Seidel, block Gauss-Seidel and multiplicative Schwarz method) subspace correction. There are certainly a lot of works related to the convergence of these methods for positive definite systems. However, most

of them are restricted to study the case when the problem is symmetric positive definite. An extensive theoretical study of the subspace correction methods can be found in Bramble and Zhang [12], Hackbusch [35], Trottenberg, Oosterlee and Schüller [85] and a survey paper by Xu [89]. Our presentation here will also use some of the results described by Xu and Zikatanov [91] and the main result obtained in [91] can be also found in the Appendix B.

For convergence rate estimates related to the semidefinite case, we refer to some recent works by Chang and Sun [19] and Marek and Szyld [62]. Both of them provide an algebraic convergence analysis for the multiplicative Schwarz method. Similar results for classical iterative methods have been obtained in [7], [50], [61], [5], and [17].

From an abstract point of view, Xu and Zikatanov [91] have established an optimal result given by an identity for the convergence rate of the method of subspace corrections applied to the symmetric positive definite problems (see also Appendix B below). The main result of this section is on an extension of the identity to the case when the problem is singular. We would like stress on the point that an extension of the result from [91] to the semidefinite case is not at all straightforward. A careful look at the theory for the in [91] shows that the convergence rate estimate crucially relies on the fact that the ranges of a subspace solver and its Hilbert-adjoint are identical under certain assumptions (see the Lemma 4.1 in [91] and also Appendix B).

To prove a convergence result in the semi-definite case, among other things, one needs first to make the right assumptions, and second to introduce an appropriate "replacements" for the $a$-adjoint operators, which in turn makes the analysis quite different from the definite case and oftentimes much more elaborate. Such technical changes surely

are not obvious, and at the last two sections, we show on simple examples, that the assumptions we have made are in fact necessary for convergence of the iterative method. We also point out the similarities and relations with the P-regularity for matrix splittings (see H. Keller [50]), since the latter is a well-known criterion for the convergence of stationary iterative method for singular problems.

Finally, we remark that the result obtained here for the singular problem can also be used to make a new framework on the convergence analysis of the method of subspace corrections applied to the symmetric positive definite problems. Some initial illustrations are made in the Appendix B.

Let us begin this chapter with some preliminaries on the singular problems posed in a Hilbert space setting and some notation useful for the following discussions.

### 4.4.1 Preliminaries

In this section, we shall formulate singular problems in a Hilbert space setting and introduce various useful notation for further discussions that follow.

Let $V$ be a Hilbert space with an inner product $(\cdot,\cdot)_V = (\cdot,\cdot)$ and a corresponding norm $\|\cdot\|_V = \|\cdot\|$ and let $V^*$ denote the space of bounded linear functionals on $V$.

We introduce a symmetric bilinear form $a : V \times V \mapsto \mathbb{R}$ satisfying the following conditions :

$$a(v,v) \geq 0 \quad \text{and} \quad a(v,w) \lesssim \|v\|\|w\|, \quad \forall v, \forall w \in V. \tag{4.4.1}$$

If $a(\cdot,\cdot)$ is semi-definite, then the seminorm induced from $a(\cdot,\cdot)$ we denote with $|\cdot|_a$, while if $a(\cdot,\cdot)$ is a positive definite, then we write $(\cdot,\cdot)_a$ for the corresponding inner

product and $\|\cdot\|_a$ for the norm. On few occasions, whenever the restriction of $a(\cdot,\cdot)$ on a subspace $W \subset V$, is positive definite, then $|\cdot|_a$ on $W$ will be replaced by $\|\cdot\|_a$.

Some general standard notation we will use is as follows: For any closed space $W \subset V$, $W^{\perp}$ denotes the orthogonal complement of $W$ with respect to the inner product, $(\cdot,\cdot)$; for two subspaces $N$ and $W$ of $V$ with $N$ being closed subspace of $W$, $W/N$ denotes the quotient space of $W$; for a continuous linear operator $T : V \mapsto V$, by $\mathcal{N}(T)$ and $\mathcal{R}(T)$, we denote the null space of $T$ and the range of $T$ respectively.

Related to the bilinear form are the spaces $\mathcal{N}$ and $\mathcal{N}^{\circ}$ defined by

$$\mathcal{N} = \{v \in V : a(v,w) = 0 \quad \forall w \in V\} \tag{4.4.2}$$

and

$$\mathcal{N}^{\circ} = \{f \in V^* : \langle f, v\rangle = 0 \quad \forall v \in \mathcal{N}\} \tag{4.4.3}$$

respectively, where $\langle\cdot,\cdot\rangle$ is the duality pairing between $V^*$ and $V$.

The variational formulation of the semidefinite (singular) problem then is: Given $f \in \mathcal{N}^{\circ}$, find $u \in V$ such that

$$a(u,v) = \langle f, v\rangle, \quad \forall v \in V. \tag{4.4.4}$$

An important assumption is that $a(\cdot,\cdot)$ satisfies the following coercivity condition:

$$a(v,v) \gtrsim \|v\|^2_{V/\mathcal{N}}, \quad \forall v \in V. \tag{4.5}$$

It is well-known that (4.5) together with $f \in \mathcal{N}^{\circ}$ implies that the problem (4.4.4) is solvable, although there might be infinitely many solutions.

### 4.4.2 Identity for the Gauss-Seidel method for the system

In this section, we shall consider the singular problem posed on the finite dimensional space and discuss the convergence analysis for the Gauss-Seidel method for the problem.

This section shall then be believed to illustrate the main idea of the mathematical analysis of the method of successive subspace corrections for the singular problem.

Let us introduce a basis $\{\phi_i\}_{i=1}^n$ for $V$ to translate the abstract version of the singular problem (4.4.4) into the explicit algebraic problem. Given a basis $\{\phi_i\}_{i=1}^n$, there exists a unique $\nu = (\nu_i) \in \mathbb{R}^n$ such that

$$v = \sum_{i=1}^{n} \nu_i \phi_i, \quad \forall v \in V, \tag{4.6}$$

The vector $\nu$ is the representation of $v$, via this fixed basis. In a standard way, we obtain the matrix representation of the bilinear form and the right hand side of (4.4.4):

$$\mathcal{A} = \Big(\mathcal{A}_{ij}\Big)_{i,j=1,\cdots,n} \quad \text{and} \quad \eta = (\eta_i)_{i=1,\cdots,n}, \tag{4.7}$$

where

$$\mathcal{A}_{ij} = a(\phi_j, \phi_i) \quad \text{and} \quad \eta_i = \langle f, \phi_i \rangle. \tag{4.8}$$

As it is well known, solving the semidefinite system (4.4.4) is equivalent to the solution of the algebraic system:

$$\mathcal{A}\mu = \eta, \tag{4.9}$$

where $\mathcal{A}$ is symmetric and nonnegative definite with positive diagonals in $\mathbb{R}^{n\times n}$ and $\eta \in \mathcal{N}(\mathcal{A})^{\perp} \subset \mathbb{R}^n$. Note that since $\mathcal{A}$ is symmetric, $\mathcal{R}(\mathcal{A}) = \mathcal{N}(\mathcal{A})^{\perp}$.

We consider the following matrix splitting :

$$\mathcal{A} = \mathcal{D} - \mathcal{L} - \mathcal{L}^T, \tag{4.10}$$

where $\mathcal{D}$ is the diagonal, $-\mathcal{L}$ is the lower triangular of $\mathcal{A}$ respectively and $\mathcal{L}^T$ is the transpose of $\mathcal{L}$.

The Gauss-Seidel method is then given by

$$\mu^l = \mu^{l-1} + (\mathcal{D} - \mathcal{L})^{-1}(\eta - \mathcal{A}\mu^{l-1}), \quad l = 1, 2, ... \tag{4.11}$$

where $\mu^0$ is an initial guess.

We shall denote

$$|\nu|^2_{\mathcal{A}} = \nu^T \mathcal{A}\nu = (\mathcal{A}\nu, \nu)_{\ell^2} = (\mathcal{A}\nu, \nu).$$

We are in a position to discuss the convergence rate of the Gauss-Seidel method.

**Theorem 4.4.1.** *Let*

$$\mathcal{E}_{\mathcal{A}} = \mathcal{I} - (\mathcal{D} - \mathcal{L})^{-1}\mathcal{A} \tag{4.12}$$

*and*

$$|\mathcal{E}_{\mathcal{A}}|^2_{\mathcal{A}} = \sup_{\nu \in \mathcal{R}(\mathcal{A})} \frac{|\mathcal{E}_{\mathcal{A}}\nu|^2_{\mathcal{A}}}{|\nu|^2_{\mathcal{A}}}. \tag{4.13}$$

*Then the convergence rate for Gauss-Seidel (4.11) method applied to (4.9) is given as follows :*

$$|\mathcal{E}_{\mathcal{A}}|^2_{\mathcal{A}} = \frac{c_0}{1+c_0}, \tag{4.14}$$

*where*

$$c_0 = \sup_{\nu \in \mathcal{R}(\mathcal{A})} \inf_{\mathrm{c} \in \mathcal{N}(\mathcal{A})} \frac{(\mathcal{L}\mathcal{D}^{-1}\mathcal{L}^T(\nu + \mathrm{c}), (\nu + \mathrm{c}))}{|\nu|^2_{\mathcal{A}}} \tag{4.15}$$

*Proof.* It is easy to see that the convergence rate of Gauss-Seidel method is given by the following energy norm :

$$|\mathcal{E}_{\mathcal{A}}|^2_{\mathcal{A}} = \sup_{\nu \in \mathcal{R}(\mathcal{A})} \frac{(\mathcal{A}\mathcal{E}_{\mathcal{A}}\nu, \mathcal{E}_{\mathcal{A}}\nu)}{(\mathcal{A}\nu, \nu)} = \sup_{\nu \in \mathcal{R}(\mathcal{A})} \frac{(\mathcal{A}(\mathcal{I} - \mathcal{B}\mathcal{A})^*(\mathcal{I} - \mathcal{B}\mathcal{A})\nu, \nu)}{(\mathcal{A}\nu, \nu)} \tag{4.16}$$

where "$*$" denotes the adjoint operator with respect to $(\mathcal{A}\cdot, \cdot)$ inner product. We observe that

$$\mathcal{B} = (\mathcal{D} - \mathcal{L})^{-1} \text{ and } (\mathcal{I} - \mathcal{B}\mathcal{A})^* = \mathcal{I} - \mathcal{B}^T\mathcal{A}$$

In fact, "$*$" can not be well-defined in this case, here we understand it in the above sense. A simple computation yields that

$$(\mathcal{I} - \mathcal{B}\mathcal{A})^*(\mathcal{I} - \mathcal{B}\mathcal{A}) = \mathcal{I} - (\mathcal{A} + \mathcal{S})^{-1}\mathcal{A},$$

where

$$\mathcal{S} = \mathcal{L}\mathcal{D}^{-1}\mathcal{L}^T.$$

From this observation, we obtain that

$$\|\mathcal{E}\|_{\mathcal{A}}^2 = 1 - \inf_{\nu \in \mathcal{A}} \frac{((\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu, \nu)_{\mathcal{A}}}{(\mathcal{A}\nu, \nu)} \tag{4.17}$$

The remaining work is devoted to the following quantity :

$$\frac{1}{c_0} = \left( \inf_{\nu \in \mathcal{A}} \frac{((\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu, \nu)_{\mathcal{A}}}{(\mathcal{A}\nu, \nu)_A} \right)^{-1} = \sup_{\nu \in \mathcal{A}} \frac{(\nu, \nu)_{\mathcal{A}}}{((\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu, \nu)_{\mathcal{A}}}. \tag{4.18}$$

Let us denote $\mathcal{M} = \mathcal{A}^{1/2}(\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\mathcal{A}^{1/2}$ and consider

$$\begin{aligned}
\frac{1}{c_0} &= \sup_{\nu \in \mathcal{A}} \frac{(\nu, \nu)_{\mathcal{A}}}{(\mathcal{A}^{-1/2}\mathcal{M}\mathcal{A}^{-1/2}\nu, \nu)_{\mathcal{A}}} \\
&= \sup_{\nu \in \mathcal{A}} \frac{((\mathcal{A}^{-1/2}\mathcal{M}\mathcal{A}^{-1/2}\nu, \nu)_{\mathcal{A}}}{(\mathcal{A}^{-1/2}\mathcal{M}\mathcal{A}^{-1/2}\nu, \mathcal{A}^{-1/2}\mathcal{M}\mathcal{A}^{-1/2}\nu)_{\mathcal{A}}} \\
&= \sup_{\nu \in \mathcal{A}} \frac{((\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu, \tilde{v})_{\mathcal{A}}}{((\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu, (\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu)_{\mathcal{A}}}
\end{aligned}$$

Now set $\omega + \mathrm{c}(\omega) = (\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu$, where $\omega \in \mathcal{A}, \mathrm{c}(\omega) \in \mathcal{N}(A)$ and we note that $\omega = P(\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu$ and $\mathrm{c}(\omega)$ is uniquely determined by $\omega$, where $P : V \mapsto \mathcal{R}(\mathcal{A})$ is a projection defined by

$$\mathcal{A}P\nu = \mathcal{A}\nu \quad \forall \nu \in \mathbb{R}^n.$$

We then obtain that

$$\begin{aligned}
&\left( \inf_{\nu \in \mathcal{A}} \frac{((\mathcal{A}+\mathcal{S})^{-1}\mathcal{A}\nu, \nu)_{\mathcal{A}}}{(\nu, \nu)_{\mathcal{A}}} \right)^{-1} \\
&= \sup_{\{(\omega + \mathrm{c}(\omega)) : \omega \in \mathcal{R}(\mathcal{A})\}} \frac{((\mathcal{A}+\mathcal{S})(\omega + \mathrm{c}(\omega)), (\omega + \mathrm{c}(\omega)))}{(\omega, \omega)_{\mathcal{A}}}.
\end{aligned}$$

Now we claim that

$$\sup_{\{(\omega+\mathrm{c}(\omega)):\omega\in\mathcal{R}(\mathcal{A})\}} \frac{((\mathcal{A}+\mathcal{S})(\omega+\mathrm{c}(\omega)),(\omega+\mathrm{c}(\omega)))}{(\omega,\omega)_{\mathcal{A}}}$$

$$= \sup_{\omega\in\mathcal{R}(\mathcal{A})} \inf_{\mathrm{c}\in\mathcal{N}(\mathcal{A})} \frac{((\mathcal{A}+\mathcal{S})(\omega+\mathrm{c}),(\omega+\mathrm{c}))}{(\omega,\omega)_{\mathcal{A}}}$$

Our proof is based on the following observation. Namely, for a fixed $\omega \in \mathbb{R}^n$, the infimum is taken by $\mathrm{c}(\omega) \in \mathcal{N}(\mathcal{A})$. Let us assume that $\xi \in \mathcal{N}(\mathcal{A})$ is given as follows :

$$\tilde{\xi} = \arg\left( \inf_{\mathrm{c}\in\mathcal{N}(\mathcal{A})} \frac{((\mathcal{A}+\mathcal{S})(\omega+\mathrm{c}),(\omega+\mathrm{c}))}{(\omega,\omega)_{\mathcal{A}}} \right). \tag{4.19}$$

It is then easy to see that

$$((\mathcal{A}+\mathcal{S})(\omega+\tilde{\xi}),\tilde{d}) = 0, \quad \forall \tilde{d} \in \mathcal{N}(\mathcal{A}) \tag{4.20}$$

This would imply that $(\mathcal{A}+\mathcal{S})(\omega+\tilde{\xi}) \in \mathcal{R}(\mathcal{A})$. Now let us set $(\mathcal{A}+\mathcal{S})(\omega+\tilde{\xi}) = \mathcal{A}\nu'$ for some $\nu' \in \mathbb{R}^n$ or equivalently $(\mathcal{A}+\mathcal{S})(\omega+\tilde{\xi}) = \mathcal{A}P\nu' = \mathcal{A}\nu$. Hence we conclude that $\tilde{\xi} = \mathrm{c}(\omega)$. This completes the proof. □

In the following section, we shall generalize the result in this section to the general successive subspace corrections to the singular problem in a Hilbert space setting.

### 4.4.3 MSSC: The Method of Successive Subspace Corrections

To introduce the method of subspace corrections, we assume that $V$ can be decomposed into closed subspaces $V_i \subset V\,(i = 1, \cdots, J)$ such that

$$V = \sum_{i=1}^{J} V_i. \tag{4.21}$$

Associated with each subspace $V_i$, $i = 1, \cdots, J$, we introduce subspace problems on each $V_i$, defined through approximate bilinear forms $a_i(\cdot, \cdot)$. In the analysis, the following subspaces $\mathcal{N}_i$ and $\widetilde{\mathcal{N}}_i$ of $V_i$ are needed

$$\mathcal{N}_i = \{v_i \in V_i : a(v_i, w_i) = 0 \quad \forall w_i \in V_i\} \tag{4.22}$$

and

$$\widetilde{\mathcal{N}}_i = \{v_i \in V_i : a_i(v_i, w_i) = 0 \quad \forall w_i \in V_i\} \tag{4.23}$$

It is easy to see that

$$\mathcal{N}_i = \mathcal{N} \cap V_i. \tag{4.24}$$

We remark that $\mathcal{N}_i \neq \widetilde{\mathcal{N}}_i$ in general. For example $\widetilde{\mathcal{N}}_i = \{0\}$ is oftentimes a case of interest. Now, we are in a position to introduce the algorithm of MSSC.

**Algorithm 4.4.1 (MSSC).** *Let $u^0 \in V$ be given,*

**for** $l = 1, \ldots$ *until convergence,*

$u_0^{l-1} = u^{l-1}$

**for** $i = 1, \cdots, J$

*Let* $\hat{e}_i \in V_i$ *solve*

$$a_i(\hat{e}_i, v_i) = \langle f, v_i \rangle - a(u_{i-1}^{l-1}, v_i), \quad \forall v_i \in V_i \tag{4.25}$$

$u_i^{l-1} = u_{i-1}^{l-1} + \hat{e}_i$

**endfor**

$u^l = u_J^{l-1}$

**endfor**

In the equation (4.25), we may simply set $a_i(\cdot, \cdot) = a(\cdot, \cdot)$ on $V_i$, namely we may solve subspace or local problems exactly. Observe that the algorithm (4.4.1) above is not always well-defined since the equation (4.25) may not be solvable. In the next paragraph, we provide assumptions under which the MSSC is well-defined and the convergence is guaranteed.

### 4.4.4 Assumptions on subspaces and subspace solvers

We introduce the assumptions needed to prove our abstract result.

**Assumption A1.** *A decomposition of* $V$ *consists of closed subspaces* $V_i \subset V$, *with* $i = 1, \cdots, J$ *satisfying*

$$V = \sum_{i=1}^{J} V_i. \tag{A1.1}$$

*Moreover, for each* $i = 1, \cdots, J$, *the followings hold true :*

$$a(v_i, v_i) \gtrsim \|v_i\|^2_{V_i/\mathcal{N}_i}, \quad \forall v_i \in V_i, \tag{A1.2}$$

*and*

$$\sup_{v_i \in V_i} \frac{a_i(v_i, w_i)}{\|v_i\|} \gtrsim \|w_i\|_{V_i/\widetilde{\mathcal{N}}_i}, \quad \forall w_i \in V_i. \tag{A1.3}$$

**Assumption A2.**

$$\textit{For each } i = 1, \cdots, J, \quad \widetilde{\mathcal{N}}_i \subseteq \mathcal{N}_i.$$

**Assumption A3.** *For each $i = 1, \cdots, J$, there exist $\omega \in (0, 2)$ such that*

$$a(T_i v, T_i v) \leq \omega a(T_i v, v), \quad \forall v \in V, \tag{A3.1}$$

*and*

$$a(T_i v, T_i v) \gtrsim \|T_i v\|^2, \quad \forall v \in V. \tag{A3.2}$$

Since in general, it is not true that the sum of closed subspaces is closed (see e.g. [76]) and (4.21) is a necessary condition for the convergence even in positive definite case, of subspace correction method (see [91]), $(A1.1)$ is assumed here as well.

It is perhaps worth pointing out that in finite dimensional case, $(A1.1)$ is automatically satisfied. In the assumptions above $T_i : V \mapsto V_i$ are linear operators, called subspace solvers, which are related to $a_i(\cdot, \cdot)$. Their definition of is postponed to (4.4.32) and the fact that their actions are unambiguously defined is contained in Lemma 4.4.1 and (4.4.32) below.

**Lemma 4.4.1.** *Under the assumptions* $(A1.2), (A1.3)$ *and* $(A2)$*, the equation* (4.25) *is solvable.*

*Proof.* Recall that the problem (4.25) is to find $v_i \in V_i$ such that

$$a_i(v_i, w_i) = a(u - u^{l-1}, w_i), \quad \forall w_i \in V_i, \tag{4.4.25}$$

where $u$ is a solution to

$$a(u, v) = \langle f, v \rangle, \quad \forall v \in V. \tag{4.4.26}$$

To show that the equation (4.4.25) is solvable, we invoke the well-known fact (see e.g. [30], p. 58) that $(A1.3)$ is equivalent to saying that

$$B_i : V_i/\widetilde{\mathcal{N}}_i \mapsto (V_i/\widetilde{\mathcal{N}}_i)^* \text{ is an isormophism,} \tag{4.4.27}$$

where $B_i : V_i \mapsto V_i^*$ is an operator defined by

$$\langle B_i v_i, w_i \rangle = a_i(v_i, w_i), \quad \forall v_i, \forall w_i \in V_i. \tag{4.4.28}$$

Based on this fact, it is enough to show that for any given $v \in V$, $a(v, \cdot) \in (V_i/\widetilde{\mathcal{N}}_i)^*$. But this is evident from the assumption $(A2)$, and the fact that

$$a(v, w_i) = 0, \quad \forall w_i \in \widetilde{\mathcal{N}}_i. \tag{4.4.29}$$

This proves that the problem (4.4.25) is well-posed on $V_i/\widetilde{\mathcal{N}}_i$ and so it is solvable on $V_i$.

Further, if $a_i(\cdot,\cdot) = a(\cdot,\cdot)$, similarly, we use the fact that the assumption $(A1.2)$ is equivalent that

$$A_i : V_i/\mathcal{N}_i \mapsto (V_i/\mathcal{N}_i)^* \text{ is an isormophism,} \tag{4.4.30}$$

where $A_i : V_i \mapsto V_i^*$ is an operator defined by

$$\langle A_i v_i, w_i\rangle = a(v_i, w_i), \quad \forall v_i, \forall w_i \in V_i \tag{4.4.31}$$

and observe that $a(v,\cdot) \in (V_i/\mathcal{N}_i)^*$.

This proves that subspace problems have solution on $V_i$ for both cases: when the subspace problems are exact or approximate. □

By the Lemma 4.4.1, for any given $v \in V$, we can define $T_i v \in V_i/\widetilde{\mathcal{N}}_i$ to be the unique solution to

$$a_i(T_i v, w_i) = a(v, w_i), \quad \forall w_i \in V_i. \tag{4.4.32}$$

Namely, $T_i : V \mapsto V_i$ is well-defined and it will be called the subspace solver. If $a_i(\cdot,\cdot) = a(\cdot,\cdot)$, the corresponding subspace solver is still well-defined on $V_i/\mathcal{N}_i$, which will be denoted by $P_i$ and it will be called the exact solver.

### 4.4.5 Some Remarks on the Abstract Assumptions

In this section, we shall discuss and elaborate some meaning of the assumptions, $(A1)$, $(A2)$ and $(A3)$ made in the previous section. The assumptions $(A1.2)$, $(A1.3)$ and (A2) cover two very important cases, when $\mathcal{N}_i = \{0\}$, namely each subspace problem is well-posed. This would correspond to a domain decomposition method. When $\mathcal{N}_i \neq$

$\{0\}$, these assumptions give that each subspace problem is consistent. Such settings correspond to a multigrid method, in which every coarse space contains the null space of $a(\cdot,\cdot)$.

Finally, let us consider the assumptions (A1.2) and (A3.2), which are given for the subspace solvers, $P_i$, exact solver and $T_i$. Basically, by assumptions (A1.2) and (A3.2), we control the range of $P_i$ and $T_i$ respectively. More precisely, $(A1.2)$ assumes that the subspace problem is well-posed on $V_i/\mathcal{N}_i$ and $(A3.2)$ also assumes the coercivity of $a(\cdot,\cdot)$ on $\mathcal{R}(T_i)$.

Note that we may not be able to assume the coercivity of $a(\cdot,\cdot)$ on $\mathcal{R}(P_i)$ since $P_i$ is invariant operator on $\mathcal{N}_i$. The first question about the assumption (A1.2) is that it might be deduced from the coercivity of $a(\cdot,\cdot)$ on $V/\mathcal{N}$ assumed previously, see the equation (4.5) in the section (4.4.1). We shall see that this is not the case immediately below. Now, let us consider the assumption (A3.2). The question is why we need it. This assumption requires $T_i$ reduces an error in the space $\mathcal{N}^{\perp}$ and can be said to control the range of $T_i$ equivalently. Indeed, in §4.4.8, we shall see that without this assumption, we can not guarantee the convergence of the method. Moreover, both assumptions shall be used crucially to prove $Q(\mathcal{R}(P_i))$ and $Q(\mathcal{R}(T_i))$ are closed in the following subsection 4.4.6.

We are in a position to elaborate the necessity of two assumptions $(A1.2)$ and $(A3.2)$. In doing so, we shall use some examples to clarify our ideas. Let us first recall the space $\ell^2$ which is often used to provide some peculiar examples in the infinite

dimensional Hilbert space. The space $\ell^2$ is defined as follows :

$$\ell^2 = \left\{ x = (x_1, x_2, \cdots) : \|x\|_{\ell^2} = \left( \sum_{i=1}^{\infty} |x_i|^2 \right)^{1/2} < \infty, \quad x_i \in \mathbb{R} \right\}. \tag{4.4.33}$$

It is well-known that $\ell^2$ is a Hilbert space and the orthonormal system $\{e_1, e_2 \cdots\}$ in $\ell^2$ is maximal, where

$$e_i = (0, 0, \cdots, 0, 1, 0, \cdots) \tag{4.4.34}$$

with the solitary 1 being in the i-th position in the sequence.

By the maximal orthonormal system, we mean that for any $f \in \ell^2$, the following holds true :

$$f = \sum_{i=1}^{\infty} (f, e_i) e_i. \tag{4.4.35}$$

In other words,

$$\ell^2 = \overline{\text{span}\{e_1, e_2, \cdots, \}}, \tag{4.4.36}$$

where $\text{span}\{e_1, e_2, \cdots, \}$ is the set of finite linear combinations from the collection of $\{e_1, e_2, \cdots, \}$.

We also recall that the coercivity of the bilinear form $a(\cdot, \cdot)$ is assumed. Namely we have

$$a(v, v) \gtrsim \|v\|^2_{V/\mathcal{N}}, \quad \forall v \in V. \tag{4.4.37}$$

One may easily notice that such a condition (4.4.37) is not automatically satisfied. See the following example :

**Example 4.4.1.** *Let us set* $V = \ell^2$ *and consider the bilinear form* $a(\cdot,\cdot)$ *defined by*

$$a(e_i, e_j) = \frac{1}{i}\delta_{ij} \quad i = 1, 2, ..., \tag{4.4.38}$$

*where* $\delta_{ij}$ *is the Kronecker delta function.*

Throughout this subsection, we shall assume that $V = \ell^2$.

#### 4.4.5.1 On the Assumption $(A1.2)$

Let us consider a closed subspace $V_i$ of $V$ and recall that a closed subspace $\mathcal{N}_i$ is defined by

$$\mathcal{N}_i = \{v_i \in V_i : a(v_i, u_i) = 0, \quad \forall u_i \in V_i\}. \tag{4.4.39}$$

The Assumption (A1.2) states that

$$a(v_i, v_i) \gtrsim \|v_i\|^2_{V_i/\mathcal{N}_i} \quad \forall v_i \in V_i \tag{4.4.40}$$

and it is given for the well-posedness of the subspace problem.

The main question to be asked here can be formulated as follows :

**Question 4.4.1.** *Can we deduce (4.4.40) from (4.4.37)?*

We shall answer this Question 4.4.1 negatively by providing an example in which (4.4.40) does not hold although (4.4.37) holds true.

Let us define a bilinear form $a(\cdot,\cdot)$ by the followings:

$$a(e_{2i-1}, e_{2j-1}) = 0 \quad \forall i, j = 1, \cdots \tag{4.4.41}$$

and

$$a(e_{2i}, e_{2j}) = \delta_{ij} \quad \forall i, j = 1, \cdots . \tag{4.4.42}$$

We then note that by the definition,

$$\mathcal{N} = \overline{\text{span}\{e_{2i-1} : i = 1, \cdots\}} \tag{4.4.43}$$

and

$$\mathcal{N}^{\perp} = \overline{\text{span}\{e_{2i} : i = 1, \cdots\}}. \tag{4.4.44}$$

We shall now consider a closed subspace $V_i$ of $V$ given as follows :

$$V_i = \overline{\text{span}\{w_1, \cdots\}}, \quad w_i = \alpha_i \left( e_{2i-1} + \frac{1}{2^i} e_{2i} \right), \tag{4.4.45}$$

where $\alpha_i$ is chosen so that

$$\|w_i\|_V = 1. \tag{4.4.46}$$

Namely, $\alpha_i$ has the following explicit form :

$$\alpha_i = \frac{1}{\sqrt{1 + \frac{1}{4^i}}} \approxeq 1, \quad \forall i \geq 1. \tag{4.4.47}$$

We note that

$$a(w_i, w_j) = \delta_{ij} \frac{\alpha_i \alpha_j}{2^{i+j}}, \quad \forall w_i \in V_i. \tag{4.4.48}$$

On the other hand, we have

$$\|w_i\|_{V_i/\mathcal{N}_i} = \|w_i\|_V = 1, \tag{4.4.49}$$

since $\mathcal{N}_i = V_i \cap \mathcal{N} = \{0\}$.

We conclude that for the bilinear form $a(\cdot,\cdot)$ defined by the relations (4.4.41) and (4.4.42), the Assumption (4.4.37) holds true, but the Assumption $(A1.2)$ does not hold.

#### 4.4.5.2 On the Assumption (A3.2)

We recall the well-known fact that since $\mathcal{N}$ is a closed subspace for $V$, we have the following unique decomposition of $V$ :

$$V = \mathcal{N} \oplus \mathcal{N}^{\perp} \tag{4.4.50}$$

and the projection $Q : V \mapsto \mathcal{N}^{\perp}$ is surjective and hence $Q(V)$ is closed. The question to be raised and discussed in this section is the following :

**Question 4.4.2.** *Assume that $W \subset V$ is closed, then is $Q(W)$ necessarily closed in $V$?*

This is subtle and indeed true if we can control the kernel part of $\forall w$ in $W$. For example, under the following condition

$$\|Qw\| \gtrsim \|w\|, \quad \forall w \in W. \tag{4.4.51}$$

we can answer the Question (4.4.2) affirmatively. The Assumption (A3.2) is about such a control for $\mathcal{R}(T_i)$ or equivalently, we may say that it controls the "twistedness" of the space $\mathcal{R}(T_i)$.

As shall be seen immediately in the following, we may not guarantee the closedness of $Q(W)$ although $W$ is closed. Namely, for the closedness of $Q(W)$, such a control (4.4.51) is necessary. The key idea in constructing a counter example for the Question (4.4.2) is to look at a subspace $W$ for which the kernel is not well-controlled. Indeed, the subspace $V_i$ defined in the previous section works also for this purpose. Let us define $W \subset V$ as follows :

$$W = \overline{\text{span}\{w_1, \cdots\}}, \quad w_i = \alpha_i \left( e_{2i-1} + \frac{1}{2^i} e_{2i} \right), \tag{4.4.52}$$

where $\alpha_i$ is given as follows :

$$\alpha_i = \frac{1}{\sqrt{1 + \frac{1}{4^i}}}. \tag{4.4.53}$$

We then notice that

$$Qw_i = \frac{1}{\sqrt{1 + 4^i}} e_{2i}. \tag{4.4.54}$$

Hence, $\overline{Q(W)} = \mathcal{N}^\perp$. Namely, $Q(W)$ is dense in $\mathcal{N}^\perp$. We shall now show that there exists $\breve{w} \in \mathcal{N}^\perp$ for which no pre-image in $W$ of $Q$ exists. Namely, $Q : W \mapsto \mathcal{N}^\perp$ is not surjective and so $Q(W)$ is not closed. Let us choose $\breve{w} \in \mathcal{N}^\perp$ as follows :

$$\breve{w} = \{0, 1, 0, 1/2, 0, 1/4, 0, \cdots, \} = \sum_{i=1} \frac{1}{2^i} e_{2i} \tag{4.4.55}$$

and assume there exists $w \in W$ such that $Qw = \breve{w}$. It is clear then that $w$ should be of the following form :

$$w = \sum_{i=1} \mu_i w_i = \sum_{i=1} \mu_i \alpha_i \left( e_{2i-1} + \frac{1}{2^i} e_{2i} \right) = \sum_{i=1} \left( e_{2i-1} + \frac{1}{2^i} e_{2i} \right). \tag{4.4.56}$$

Namely, $\mu_i = \alpha_i^{-1}$. Now from the fact that the norm of $w$ is given as follows,

$$\|w\|^2 = \sum_{i=1} \left( 1 + \frac{1}{4^i} \right) = \infty. \tag{4.4.57}$$

we are led to a contradiction.

### 4.4.6 On the convergence rate of the MSSC

In this section, we prove several auxiliary results, centered around Theorem 4.4.3, which is later used to extend the estimates from [91] to the case of semidefinite problems.

To write a more convenient expression for the error, let $u \in V$ be a solution to (4.4.4) and $\{u^l : l = 0, \cdots\}$ be the iterates generated by (4.4.1) respectively. The following relation is standard:

$$u - u^l = E_J(u - u^{l-1}) = ... = E_J^l(u - u^0), \tag{4.4.58}$$

where

$$E_J = (I - T_J) \cdots (I - T_1). \tag{4.4.59}$$

The operator $E_J$ is referred to as the error transfer operator and if $a(\cdot,\cdot)$ is symmetric positive definite, the uniform convergence result is obtained usually by proving that

$$\|E_J\|_a = \sup_{\|v\|_a=1} \|E_J v\|_a < 1. \tag{4.4.60}$$

When $a(\cdot,\cdot)$ is semi-definite, the convergence rate can be estimated by the following norm of $E_J$, namely

$$|E_J|_a = \sup_{v\in\mathcal{N}^\perp} \frac{|E_J v|_a}{\|v\|_a}. \tag{4.4.61}$$

Introducing

$$E_{J,a} = (I - QT_J)\cdots(I - QT_1). \tag{4.4.62}$$

where $Q : V \mapsto \mathcal{N}^\perp$ is the orthogonal projection onto $\mathcal{N}^\perp$ with respect to $(\cdot,\cdot)$ inner product. The following relations are obvious:

$$|E_J|_a = \|E_{J,a}\|_a = \sup_{v\in\mathcal{N}^\perp} \frac{\|E_{J,a} v\|_a}{\|v\|_a}, \tag{4.4.63}$$

To see, why (4.4.63) holds, we observe that

$$|E_J v|_a^2 = a(E_J v, E_J v) = a(QE_J v, QE_J v) = \|E_{J,a} v\|_a^2, \quad \forall v \in \mathcal{N}^\perp \tag{4.4.64}$$

and from the definition of $T_i$, we have that for $c \in \mathcal{N}$,

$$a_i(T_i c, v_i) = a(c, v_i) = 0, \quad \forall v_i \in V_i. \tag{4.4.65}$$

Namely, $T_i c \in \tilde{\mathcal{N}}_i, \quad \forall c \in \mathcal{N}$.

To make our idea more clear, we shall introduce the following notation. For any continuous operator $T : V \mapsto V$, $T_a : V \mapsto V$ shall denote

$$T_a := QT. \tag{4.4.66}$$

Also, for any closed subspace $W \subset V$, $W_a$ is the space defined by

$$W_a := \{u \in V \,|\, u = Qw\,;\, w \in W\}. \tag{4.4.67}$$

With this notation in hand, the operator $E_{J,a}$ defined in (4.4.62) can be written as follows:

$$E_{J,a} = (I - T_{J,a}) \cdots (I - T_{1,a}). \tag{4.4.68}$$

One can easily see that (4.4.63) and (4.4.68) provide us with a hint that the study of the convergence rate for the MSSC for (4.4.4) can be performed by restricting subspace solvers $T_i$ and error transfer operator $E_J$ to $\mathcal{N}^{\perp}$ and applying the result from [91] for the symmetric positive definite case. This shall also give us a way to avoid the difficulty arising from the non-trivial null space of $a(\cdot, \cdot)$. The difficulty is, after writing out the estimate in terms of $T_{i,a}$ to go back to the original subspace solvers.

#### 4.4.6.1 Some technical Lemmas

In this section, we prove several technical lemmas which establish the assumptions needed to prove the auxiliary result Theorem 4.4.3.

**Lemma 4.4.2.** *Under the assumptions* $(A1.2), (A1.3)$ *and* $(A3.2)$*, for each* $i = 1, \cdots, J$*, the followings hold true*

*(a)* $T_i c = 0, \quad \forall c \in \mathcal{N}$.

*(b)* $\mathcal{N}(T_{i,a}) \cap V_{i,a} = \{0\}$.

*(c)* $\mathcal{R}(T_i)$ *and* $\mathcal{R}(T_{i,a})$ *are closed.*

*Proof.* First, we will prove (a). Recall that for any $c \in \mathcal{N}$, $T_i c \in \tilde{\mathcal{N}}_i$ (see (4.4.65). Now by the assumption $(A3.2)$,

$$\|T_i c\|^2 \lesssim a(T_i c, T_i c) = 0. \tag{4.4.69}$$

Namely $T_i c = 0$. This completes the proof of (a). Second, we shall show that $\mathcal{N}(T_{i,a}) \cap V_{i,a} = \{0\}$. It is enough to show that for $v_i \in V_i$,

$$T_{i,a} v_i = Q T_i v_i = 0 \quad \text{implies that } Q v_i = 0 \tag{4.4.70}$$

Again from the assumption (A3.2), we have that if $Q T_i v_i = 0$, then $T_i v_i = 0$. Now from the definition of $T_i$, we obtain that

$$0 = a_i(T_i v_i, w_i) = a(v_i, w_i), \quad \forall w_i \in V_i. \tag{4.4.71}$$

Hence, $v_i \in \mathcal{N}_i$, and since $\tilde{\mathcal{N}}_i \subset \mathcal{N}_i$, we conclude that $Q v_i = 0$, which gives (b). It remains to show that $\mathcal{R}(T_i)$ is closed. By (A1.2), for any $v \in V$, there exists $v_i \in V_i / \mathcal{N}_i$ such that

$$a(v_i, w_i) = a(v, w_i), \quad \forall w_i \in V_i. \tag{4.4.72}$$

From this, we obtain that

$$a_i(T_i v_i, w_i) = a(v_i, w_i) = a(v, w_i) = a_i(T_i v, w_i), \quad \forall w_i \in V_i. \tag{4.4.73}$$

with $T_i v_i \in V_i/\widetilde{\mathcal{N}}_i$. The relation (4.4.73) together with the fact that $T_i c = 0$ for all $c \in \mathcal{N}$, we proved that $\mathcal{R}(T_i) = T_i(V_i/\mathcal{N}_i)$. Our task is now to show that $T_i(V_i/\mathcal{N}_i)$ is closed. From the inequality

$$\begin{aligned} \|T_i v_i\|_{V_i} \|v_i\|_{V_i/\mathcal{N}_i} &= \|T_i v_i\|_{V_i/\widetilde{\mathcal{N}}_i} \|v_i\|_{V_i} \\ &\gtrsim a_i(T_i v_i, v_i) \\ &= a(v_i, v_i) \\ &\gtrsim \|v_i\|^2_{V_i/\mathcal{N}_i}. \end{aligned} \tag{4.4.74}$$

it follows that $\mathcal{R}(T_i)$ is closed, and this in turn implies that $\mathcal{R}(T_{i,a})$ is closed, by using (A3.2). This completes the proof. □

**Lemma 4.4.3.** *If* $(A1), (A2)$ *and* $(A3)$ *are satisfied, then for* $i = 1, \cdots, J$ *we have the following:*

*(a)* $V_{i,a}$ *is closed and* $\mathcal{N}^{\perp} = \sum_{i=1}^{J} V_{i,a}$.

*(b)* $T_{i,a} : V_{i,a} \mapsto V_{i,a}$ *is an isomorphism.*

*(c)* $a(T_{i,a} v, T_{i,a} v) \leq \omega a(T_{i,a} v, v), \quad \forall v \in V.$

*Proof.* We shall show (c),(a) and (b) in that order. The item (c) follows from ($A3.1$), by

$$a(T_{i,a}v, T_{i,a}v) = a(T_iv, T_iv) \leq \omega a(T_iv, v) = \omega a(T_{i,a}v, v), \quad \forall v \in V. \tag{4.4.75}$$

Since $V_i$ is closed, it is easy to show that $V_{i,a}$ is closed from the assumption (A1.2). We now observe that by (A1),

$$\mathcal{N}^{\perp} = QV = \sum_{i=1}^{J} QV_i = \sum_{i=1}^{J} V_{i,a}. \tag{4.4.76}$$

This proves (a). Finally, to show (b), it is enough to establish that

$$\mathcal{R}(T_{i,a}) = V_{i,a}. \tag{4.4.77}$$

It is obvious that

$$\mathcal{R}(T_{i,a}) \subset V_{i,a}. \tag{4.4.78}$$

The harder part of the proof is to show the reverse inclusion. From part (a) of the Lemma 4.4.2 and (4.4.78), we may view $T_{i,a}$ as an operator from $V_{i,a}$ to itself Note that in such case, the adjoint operator of $T_{i,a}$, denoted by $T_{i,a}^*$ is well-defined. To prove

$$\mathcal{R}(T_{i,a}) \supset V_{i,a}. \tag{4.4.79}$$

we use the fact that from (4.4.75), it follows that $I - T_{i,a}$ is non-expansive,i.e. $\|I - T_{i,a}\|_a \leq 1$. This then implies that

$$\mathcal{N}(T_{i,a}) = \mathcal{N}(T_{i,a}^*).$$

and applying the Closed Range Theorem we obtain

$$V_{i,a} = \mathcal{R}(T_{i,a}) \oplus \mathcal{N}(T_{i,a}^*) = \mathcal{R}(T_{i,a}) \oplus \mathcal{N}(T_{i,a}). \tag{4.4.80}$$

Finally, from the part (b) of Lemma 4.4.2, namely $\mathcal{N}(T_{i,a}) = \{0\}$, we complete the proof of (4.4.77). This gives that $T_{i,a} : V_{i,a} \mapsto V_{i,a}$ is onto, and hence

$$T_{i,a}(V_i) = T_{i,a}(V_{i,a}) = \mathcal{R}(T_{i,a}) = V_{i,a}.$$

On the other hand, part (a) of the Lemma 4.4.2 shows that $T_{i,a}$ is also one-to-one on $V_{i,a}$. Applying the Open Mapping Theorem we obtain that $T_{i,a} : V_{i,a} \mapsto V_{i,a}$ is an isomorphism and this completes the proof of (b). □

#### 4.4.6.2 An identity for the convergence factor of the MSSC

Let us recall a result established in [91].

**Theorem 4.4.2.** *Assume that $a(\cdot,\cdot)$ is positive definite. Then under the assumptions that*

**(B0)** $V = \sum_{i=1}^{J} V_i$, *with $V_i$ closed for $i = 1, \cdots, J$,*

**(B1)** $T_i : V_i \mapsto V_i$ *is isomorphic for each $i = 1, \cdots, J$,*

*and*

**(B2)** $\|T_i v\|_a^2 \le \omega a(T_i v, v) \quad \forall v \in V$ *with some* $\omega \in (0,2)$, *we have*

$$\|E\|_a^2 = \|(I - T_J)\cdots(I - T_1)\|_a^2 = \frac{c_0}{1 + c_0},$$

*where*

$$c_0 = \sup_{\|v\|_a = 1} \inf_{\sum_{i=1}^J v_i = v} \sum_{i=1}^J (\bar{T}_i^{-1} T_i^* w_i, T_i^* w_i)_a \ \ \textit{with} \ \ w_i = \sum_{j=i}^J v_j - T_i^{-1} v_i.$$

The auxiliary result stated below can be obtained by a straightforward manner from (4.4.63), Lemma 4.4.3 and Theorem 4.4.2.

**Theorem 4.4.3.** *Under the assumptions* $(A1), (A2)$ *and* $(A3)$, *we obtain*

$$|E_J|_a^2 = |(I - T_J)\cdots(I - T_1)|_a^2 = \frac{c_0}{1 + c_0}, \tag{4.4.81}$$

*where*

$$c_0 = \sup_{v \in \mathcal{N}^\perp} \inf_{\sum_{i=1}^J v_i = v} \frac{\sum_{i=1}^J (\bar{T}_{i,a}^{-1} T_{i,a}^* u_i, T_{i,a}^* u_i)_a}{\|v\|_a^2}, \tag{4.4.82}$$

$u_i = \sum_{j=i}^J v_j - T_{i,a}^{-1} v_i$ *and* $v_i \in V_{i,a}$.

*Proof.* Notice that $c_0$ in (4.4.82) is expressed by the operators $T_{i,a}$, the restriction of $T_i$ on $\mathcal{N}^\perp$. Based on the simple observation (4.4.63) made in the previous section, to obtain (4.4.81), it is then enough to verify three assumptions in the Theorem 4.4.2 in

terms of $V_{i,a}$ and $T_{i,a}$ for $i = 1, \cdots, J$. But these are the same as (a),(b) and (c) in the Lemma 4.4.3. This completes the proof. □

The constant $c_0$ in (4.4.82) is expressed in terms of $T_{i,a}$, the restriction of $T_i$ onto $\mathcal{N}^{\perp}$. In the following section, we shall discuss how to rewrite $c_0$ in terms of the subspace solvers $T_i$.

### 4.4.7 An abstract convergence result

We begin this section by defining an analogue of adjoint operator of $T_i$ with respect to $a(\cdot,\cdot)$. Note that the Hilbert adjoint of $T_i$ with respect to $a(\cdot,\cdot)$ is not well defined since the latter is not an inner product. Thanks to ($A$1.3) and (A2), we notice that for any given $v \in V$, there is a unique $w_i \in V_i/\widetilde{\mathcal{N}}_i$ such that

$$a_i(v_i, w_i) = a(v, v_i), \quad \forall v_i \in V_i. \tag{4.4.83}$$

We then set $T_i^* v = w_i$. The following important relation is satisfied by $T_i^*$:

$$a(T_i v, w) = a_i(T_i v, T_i^* w) = a(v, T_i^* w), \quad \forall v, \forall w \in V, \tag{4.4.84}$$

Note that if $a(\cdot,\cdot)$ is an inner product, this is just the definition for the Hilbert adjoint. In accordance with this definition, we define also the symmetrization $\bar{T}_i$ of $T_i$ as follows:

$$\bar{T}_i = T_i + T_i^* - T_i^* T_i. \tag{4.4.85}$$

We observe that from the assumption $(A1.3)$, we have for $c \in \mathcal{N}$,

$$0 = \sup_{v_i \in V_i} \frac{a(c, v_i)}{\|v_i\|} = \sup_{v_i \in V_i} \frac{a_i(v_i, T_i^* c)}{\|v_i\|} \gtrsim \|T_i^* c\|_{V_i/\widetilde{\mathcal{N}}_i}. \tag{4.4.86}$$

We conclude that $T_i^* c \in \widetilde{\mathcal{N}}_i \subset \mathcal{N}_i$. This then implies that

$$\bar{T}_{i,a} = Q\bar{T}_i. \tag{4.4.87}$$

because

$$\begin{aligned} \bar{T}_{i,a} &= T_{i,a} + T_{i,a}^* - T_{i,a}^* T_{i,a} \\ &= QT_i + QT_i^* - QT_i^* QT_i \\ &= QT_i + QT_i^* - QT_i^* T_i = Q\bar{T}_i. \end{aligned}$$

Let us consider now the $(\cdot, \cdot)$ orthogonal projections $Q_i : V \mapsto V_i$ defined by

$$\langle Q_i v, v_i \rangle = \langle v, v_i \rangle, \quad \forall\, v \in V, \forall\, v_i \in V_i. \tag{4.4.88}$$

We have the following

**Lemma 4.4.4.** *Under the assumptions* $(A1.2), (A3.1)$ *and* $\mathcal{N}_i = \{0\}$*, the following hold true:*

*(a)* $\mathcal{R}(T_i) = \mathcal{R}(T_i^*) = \mathcal{R}(\bar{T}_i) = V_i$.

*(b)* $T_i, T_i^*$ *and* $\bar{T}_i$ *are all isomorphic from* $V_i$ *to* $V_i$.

*(c)* $Q_{i,a} : V_i \mapsto V_{i,a}$ *is an isomorphism, where* $Q_{i,a} = Q\, Q_i$.

*Proof.* We shall consider $T_i, T_i^*$ and $\bar{T}_i$ as operators from $V_i$ to itself. By the assumption $(A2)$ and $\mathcal{N}_i = \{0\}$, it is straightforward to see that $\mathcal{N}(T_i) = \mathcal{N}(T_i^*) = \{0\}$. Moreover, from $(A3.1)$, we have that

$$\frac{2-\omega}{\omega} a(T_i v, T_i v) \lesssim a(\bar{T}_i v, v), \quad \forall v \in V.$$

This implies that $\mathcal{N}(\bar{T}_i) = \{0\}$. Let us invoke the Banach's Closed Range Theorem to obtain that $\mathcal{R}(T_i) = \mathcal{R}(T_i^*) = \mathcal{R}(\bar{T}_i) = V_i$. This proves (a). The part (b) is a easy consequence of (a) due to the Open Mapping Theorem. Now from the assumption $(A1.2)$, it is easy to check that $Q_{i,a} : V_i \mapsto V_{i,a}$ is one to one and onto. This completes the proof. □

In the rest of this section, we first consider the most general case when we could remove the restriction operator $Q$ and obtain the closed form for $c_0$ in terms of the real subspace solvers $T_i$. We then state several interesting corollaries.

**Theorem 4.4.4.** *Assume* $(A1), (A2)$ *and* $(A3.1)$ *and that* $T_i = P_i$, *whenever* $\mathcal{N}_i \neq \{0\}$ *and define* $K_1$ *and* $K_2$ *by*

$$K_1 \;:=\; \{i \in \{1, \cdots, J\} : \mathcal{N}_i := \{0\}\} \quad \textit{and} \quad K_2 = \{1, \cdots, J\} \backslash K_1$$

*respectively. Then* $c_0$ *for the energy norm of* $E_J$ *in* (4.4.81) *can be written as follows:*

$$c_0 \;=\; \sup_{\|v\|_a = 1, v \in \mathcal{N}^\perp} \inf_{c \in \mathcal{N}} \inf_{\sum_i v_i = v + c} C(v_1, \cdots, v_J, v), \tag{4.4.89}$$

*where*

$$C(v_1,\cdots,v_J,v) \;=\; \sum_{i\in K_1}(\bar{T}_i^{-1}T_i^*u_i,T_i^*u_i)_a+\sum_{j\in K_2}|P_j(\sum_{k=j+1}^{J}v_k)|_a^2$$

*with*

$$u_i=\sum_{j=i}^{J}v_j-T_i^{-1}v_i \quad \textit{and} \quad v_i\in V_i.$$

*Proof.* Due to the assumption that $T_i = P_i$, whenever $\mathcal{N}_i \neq \{0\}$, we may apply the Theorem 4.4.3 without the assumption $(A3.2)$ to obtain that

$$c_0 \;=\; \sup_{\|v\|_a=1,v\in\mathcal{N}^\perp}\inf_{\sum_i\check{v}_i=v}\left(\sum_{i=1}^{J}(\bar{T}_{i,a}^{-1}T_{i,a}^*\check{u}_i,T_{i,a}^*\check{u}_i)_a\right), \tag{4.4.90}$$

where $\check{u}_i=\sum_{j=i}^{J}\check{v}_j-T_{i,a}^{-1}\check{v}_i$ with $\check{v}_i\in V_{i,a}$. Our main task is to replace $T_{i,a},T_{i,a}^*$ and $\bar{T}_{i,a}$ by $T_i,T_i^*$ and $\bar{T}_i$ and also replace $\check{v}_i\in V_{i,a}$ by some $v_i\in V_i$ for the expression in the right hand side of (4.4.90). Let us set

$$c_0(v)=\inf_{\sum_i\check{v}_i=v}\left(\sum_{i=1}^{J}(\bar{T}_{i,a}^{-1}T_{i,a}^*\check{u}_i,T_{i,a}^*\check{u}_i)_a\right), \tag{4.4.91}$$

The easy part of the proof is to replace $T_i$ by $P_i$ for $i\in K_2$ and obtain that

$$c_0(v)=\inf_{\sum_{i=1}^{J}\check{v}_i=v}\left(\sum_{i\in K_1}(\bar{T}_{i,a}^{-1}T_{i,a}^*\check{u}_i,T_{i,a}^*\check{u}_i)_a+\sum_{j\in K_2}|P_j(\sum_{k=j+1}^{J}\check{v}_k)|_a^2\right), \tag{4.4.92}$$

where $\check{u}_i = \sum_{j=i}^{J} \check{v}_j - T_{i,a}^{-1}\check{v}_i$ and $\check{v}_i \in V_{i,a}$.

We may also replace $\check{v}_i \in V_{i,a}$ by some $v_i \in V_i$ by the isomorphic operator $Q_{i,a} : V_i \mapsto V_{i,a}$ (see Lemma 4.4.4). Namely, if we define

$$v_i := Q_{i,a}^{-1}\check{v}_i \in V_i. \tag{4.4.93}$$

We notice that from the assumption that $T_i = P_i$ whenever $\mathcal{N}_i \neq \{0\}$, for $c \in \mathcal{N}$,

$$T_i^* c \in \widetilde{\mathcal{N}}_i \subset \mathcal{N}_i = \{0\}, \quad \text{for } i \in K_1 \tag{4.4.94}$$

and

$$P_i c \in \mathcal{N}_i, \quad \text{for } i \in K_2. \tag{4.4.95}$$

The first term of (4.4.92) can then be handled by observing that

$$\bar{T}_{i,a}^{-1} T_{i,a}^* \check{v}_i = \bar{T}_i^{-1} Q_{i,a}^{-1} Q_{i,a} T_i^* v_i = \bar{T}_i^{-1} T_i^* v_i, \quad \forall v_i \in V_i. \tag{4.4.96}$$

Note that the expression $\bar{T}_i^{-1} T_i^*$ makes sense due to the Lemma 4.4.4. Exploiting (4.4.95), $c_0(v)$ in (4.4.92) can be written in terms of $v_i$, $T_i$, $T_i^*$ and $\bar{T}_i$ as follows.

$$c_0(v) = \inf_{\sum_{i=1}^{J} v_i = v + c} \left( \sum_{i \in K_1} (\bar{T}_i^{-1} T_i^* u_i, T_i^* u_i)_a + \sum_{j \in K_2} |P_j( \sum_{k=j+1}^{J} v_k)|_a^2 \right), \tag{4.4.97}$$

where $u_i = \sum_{j=i}^{J} v_j - T_i^{-1} v_i$ and $v_i \in V_i$. In the equation (4.4.97), $c \in \mathcal{N}$ arises due to the fact that each $\check{v}_i \in V_{i,a}$ has been replaced by $v_i \in V_i$. To complete the proof, it remains to show that the following two quantities are equal:

$$c_0(v) = \inf_{\sum_i v_i = v+c} \left( \sum_{i \in K_1} (\bar{T}_i^{-1} T_i u_i, T_i u_i)_a + \sum_{j \in K_2} |P_j( \sum_{k=j+1}^{J} v_k)|_a^2 \right)$$

and

$$\bar{c}_0(v) = \inf_{c \in \mathcal{N}} \inf_{\sum_i v_i = v+c} \left( \sum_{i \in K_1} (\bar{T}_i^{-1} T_i u_i, T_i u_i)_a + \sum_{j \in K_2} |P_j( \sum_{k=j+1}^{J} v_k)|_a^2 \right).$$

It is straightforward to see that $\bar{c}_0(v) \leq c_0(v)$. The reverse inequality also follows easily from the following observation that for any choice $c \in \mathcal{N}$ and for a given $v \in \mathcal{N}^{\perp}$, we may find a decomposition of $v$ such that $\sum_{i=1}^{J} v_i = v + c$. This completes the proof. □

In case when we use exact solvers, $P_i$, we shall have the following expression.

**Corollary 4.4.1.** *Assume* $(A1), (A2)$, $(A3.1)$ *and* $T_i = P_i$ *for* $i = 1, \cdots, J$. *Then* $c_0$ *for the MSSC is given by*

$$c_0 = \sup_{v \in \mathcal{N}^{\perp}} \inf_{c \in \mathcal{N}} \inf_{\sum_i v_i = v+c} \frac{\sum_{i=1}^{J} |P_i(\sum_{j=i+1}^{J} v_j)|_a^2}{\|v\|_a^2}, \tag{4.4.98}$$

*where* $v_i \in V_i$.

This expression is quite useful especially for the analysis of multigrid method with the Gauss-Seidel smoothing. To illustrate this, we denote $T_i$ by one Gauss-Seidel sweep on $V_i$, then error transfer operator based on this type of smoothing can be written as

followings :

$$E_J = (I - T_1) \cdots (I - T_J) = \Pi_{k=1}^{J} \Pi_{l=1}^{n_k} (I - P_k^l), \tag{4.4.99}$$

where $n_k$ is the number of nodes in the k-th subspace with $k = 1, \cdots, J$. For more details, refer to [89].

### 4.4.8 Relations between abstract assumptions (A1), (A2), (A3) and the P-regularity

In this section, we assume that $V$ is a finite dimensional Hilbert space, namely $\dim V = n < \infty$. The goal of this section is to illustrate and elaborate our theory in more concrete situations by taking the abstract theory in the previous section down to the finite dimensional setting. Note that most of existing works on iterative methods for singular problems are performed on the finite dimensional setting, this section shall then provide some clues on how much our theory is different from the existing works and extended as well.

We shall first translate the abstract version of the singular problem (4.4.4) into the explicit algebraic problem by introducing a basis $\{\phi_i\}_{i=1}^{n}$ for $V$. We then introduce the P-regularity for matrix splittings, following H. Keller [50] (1965) and discuss how it is related to our assumptions $(A1), (A2)$ and $(A3)$. For simplicity, we shall assume that $J = 1$ and $V$ is not decomposed, i.e. $V = V_1$. In this case of interest are the relationships between (A2), (A3) and the P-regularity.

Recall that given a basis $\{\phi_i\}_{i=1}^n$, we obtain the matrix representation of the bilinear form and the right hand side of (4.4.4):

$$\mathcal{A} = \Big(\mathcal{A}_{ij}\Big)_{i,j=1,\cdots,n} \quad \text{and} \quad \eta = (\eta_i)_{i=1,\cdots,n}, \tag{4.4.100}$$

where

$$\mathcal{A}_{ij} = a(\phi_j, \phi_i) \quad \text{and} \quad \eta_i = \langle f, \phi_i \rangle. \tag{4.4.101}$$

As it is well known, solving the semidefinite system (4.4.4) is equivalent to the solution of the algebraic system:

$$\mathcal{A}\mu = \eta. \tag{4.4.102}$$

We consider a linear iterative method, based on the following matrix splitting of $\mathcal{A}$:

$$\mathcal{A} = \mathcal{B} - \mathcal{C}. \tag{4.4.103}$$

and for a given initial guess $\mu^0$, for $l = 1, \cdots$, until convergence we obtain the next iterate from

$$\mu^l = \mu^{l-1} + \mathcal{B}^{-1}(\eta - \mathcal{A}\mu^{l-1}). \tag{4.4.104}$$

Note that on the variational setting, $\mathcal{B}$ is related to the approximate problem to the original problem (4.4.25). Let the approximate problem to (4.4.102) be given by the bilinear form $b(\cdot, \cdot)$, then $\mathcal{B}$ is a matrix given by:

$$\mathcal{B} = \Big(\mathcal{B}_{ij}\Big)_{i,j=1,\cdots,n} = \Big(b(\phi_j, \phi_i)\Big)_{i,j=1,\cdots,n}. \tag{4.4.105}$$

The subspace solver, denoted by $T$ can be defined as follows:

$$b(Tv, w) = a(v, w), \quad \forall v, w \in V. \tag{4.4.106}$$

Let us denote $\widetilde{T}$ by the matrix representation of the operator $T$. It is then given by $\widetilde{T} = \mathcal{B}^{-1}\mathcal{A}$. Correspondingly, the error transfer matrix $\widetilde{\mathcal{E}}$ satisfying the relation

$$\mu - \mu^l = \widetilde{\mathcal{E}}(\mu - \mu^{l-1}) \tag{4.4.107}$$

is given by

$$\widetilde{\mathcal{E}} = I - \widetilde{T} = I - \mathcal{B}^{-1}\mathcal{A}. \tag{4.4.108}$$

We note that $\mathcal{N} = \mathcal{N}(\mathcal{A})$ and $\widetilde{\mathcal{N}} = \mathcal{N}(\mathcal{B})$, where $\widetilde{\mathcal{N}}$ is the null space of the bilinear form $b$ (see (4.23)). With this notation, we can rewrite the assumptions (A2) and (A3) as follows:

$$\mathcal{N}(\mathcal{B}) \subseteq \mathcal{N}(\mathcal{A}). \tag{4.4.109}$$

There exist $\omega \in (0, 2)$ such that

$$(\mathcal{A}\widetilde{T}\widetilde{v}, \widetilde{T}\widetilde{v})_{\ell^2} \leq \omega(\mathcal{A}\widetilde{T}\widetilde{v}, \widetilde{v})_{\ell^2}, \quad \forall \widetilde{v} \in \mathbb{R}^n, \tag{4.4.110}$$

and

$$(\mathcal{A}\widetilde{T}\nu, \widetilde{T}\nu)_{\ell^2} \geq \alpha(\widetilde{T}\nu, \widetilde{T}\nu)_{\ell^2}, \quad \forall \nu \in \mathbb{R}^n, \tag{4.4.111}$$

where $(\cdot,\cdot)_{\ell 2}$ is the discrete $\ell^2$ inner product and $\alpha$ is some positive constant. We shall simply denote $(\cdot,\cdot)_{\ell 2}$ by $(\cdot,\cdot)$ throughout this section.

**Remark 4.4.1.** *In general, a classical iterative procedure assumes $\mathcal{B}$ is invertible, namely $\widetilde{\mathcal{N}} = \mathcal{N}(\mathcal{B}) = \{0\}$.*

**Definition 4.4.1.** *The splitting* (4.4.103) *is called P-regular if $\mathcal{B}$ is invertible and $\mathcal{B} + \mathcal{B}^T - \mathcal{A}$ is positive definite.*

It is well-known [50] that the convergence of the iterative procedure (4.4.104) is equivalent to

$$\lim_{l\to\infty} \widetilde{\mathcal{E}}^l = P_{\mathcal{N}}, \tag{4.4.112}$$

where $P_{\mathcal{N}}$ is a projection onto $\mathcal{N}$, which is not necessarily an orthogonal projection (see the example (4.4.115) and (4.4.121) below).

The property (4.4.112) is called semi-convergent for $\mathcal{A}$. Under the assumption that $\mathcal{A}$ is semi-positive definite, Keller (1965) showed that the P-regularity is a sufficient condition under which (4.4.112) holds true (see [50]). Since then the P-regularity is used as an important criterion for the convergence of the iterative method based on the matrix splitting (see e.g. [5], [7], [17] and references cited therein).

Based on the our new abstract theory, it is simple to show that $\widetilde{\mathcal{E}}$ satisfies a property (4.4.112) under the assumptions (A2) and (A3). In the following, we shall see that the P-regularity is much stronger than the abstract assumptions (A2) and (A3). Namely, we have the same result as Keller, but under the weaker assumptions.

We shall begin our discussion on how the abstract assumptions (A2) and (A3) are related to the P-regularity by the following simple but lemma.

**Lemma 4.4.5.** *The following two inequalities are equivalent, namely for* $\omega \in (0,2)$*,*

$$\begin{aligned} (\mathcal{A}\widetilde{T}\nu, \widetilde{T}\nu) &\leq \omega(\mathcal{A}\widetilde{T}\nu, \nu), \quad \forall \nu \in \mathcal{R}(\mathcal{A}) \quad \Longleftrightarrow \\ \left(\frac{2}{\omega} - 1\right)(\mathcal{A}\nu, \nu) &\leq \left(\left(\mathcal{B} + \mathcal{B}^T - \mathcal{A}\right)\nu, \nu\right), \quad \forall \nu \in \mathcal{R}(\mathcal{A}). \end{aligned}$$

**Theorem 4.4.5.** *Assume that the splitting* (4.4.103) *is P-regular, then* $(A2)$ *and* $(A3)$ *hold true.*

*Proof.* Let us assume that the splitting (4.4.103) is P-regular. (A2) is then trivially satisfied. By the Lemma 4.4.5, the positive definiteness of $\mathcal{B} + \mathcal{B}^T - \mathcal{A}$ implies $(A3.1)$. It is enough to show $(A3.2)$. We shall assume that $(A3.2)$

$$(\mathcal{A}\widetilde{T}\nu, \widetilde{T}\nu) \geq \alpha(\widetilde{T}\nu, \widetilde{T}\nu), \quad \forall \nu \in \mathbb{R}^n, \tag{4.4.113}$$

does not hold. Then by a compactness argument, we can find some $\nu \neq 0 \in \mathcal{R}(\mathcal{A})$ such that $\widetilde{T}\nu \in \mathcal{N}(\mathcal{A})$. This means that $\widetilde{T}\nu = \mathrm{c}$ for some $\mathrm{c} \in \mathcal{N}(\mathcal{A})$. However, this implies that $\mathcal{A}\nu = \mathcal{B}\mathrm{c}$ and we have

$$(\mathcal{A}\nu, \mathrm{c}) = 0 \Rightarrow (\mathcal{B}\mathrm{c}, \mathrm{c}) = 0 \Rightarrow \left((\mathcal{B} + \mathcal{B}^T)\mathrm{c}, \mathrm{c}\right) = 0. \tag{4.4.114}$$

A contradiction to the fact that $\mathcal{B} + \mathcal{B}^T$ is positive on $\mathcal{N}(\mathcal{A})$. This completes the proof.

□

We note that in general, the assumptions (A2) (even with $\widetilde{\mathcal{N}} = \{0\}$) and (A3) do not necessarily imply the P-regularity. Let us consider the following example due to

Cao [17].

$$\mathcal{A} = \begin{pmatrix} 1/2 & -1 \\ -1 & 2 \end{pmatrix}. \tag{4.4.115}$$

$\mathcal{A}$ in (4.4.115) is a singular symmetric positive semi-definite matrix. Now we consider its splitting given as follows:

$$\mathcal{A} = \mathcal{B} - \mathcal{C}, \tag{4.4.116}$$

where

$$\mathcal{B}^{-1} = \begin{pmatrix} -1 & -1 \\ 0 & 1/4 \end{pmatrix} \quad \text{and} \quad \mathcal{H} = \mathcal{B} + \mathcal{B}^T - \mathcal{A} = \begin{pmatrix} -5/2 & -3 \\ -3 & 6 \end{pmatrix}. \tag{4.4.117}$$

We shall see that this splitting satisfies the assumptions (A2) and (A3), but it does not satisfy the P-regularity. To verify this, note that

$$\mathcal{R}(\mathcal{A}) = \text{span}\left\{\begin{pmatrix} -1 \\ 2 \end{pmatrix}\right\} \quad \text{and} \quad \mathcal{N}(\mathcal{A}) = \text{span}\left\{\begin{pmatrix} 2 \\ 1 \end{pmatrix}\right\}, \tag{4.4.118}$$

and

$$\left(\mathcal{H}\begin{pmatrix} -1 \\ 2 \end{pmatrix}, \begin{pmatrix} -1 \\ 2 \end{pmatrix}\right) = 163/2. \tag{4.4.119}$$

This implies that the splitting (4.4.116) satisfies ($A$3.1). It is also easy to see that $\mathcal{R}(\widetilde{T}) \cap \mathcal{N}(\mathcal{A}) = \{0\}$, so ($A$3.2) holds true. We however, further note that

$$\left( \mathcal{H} \left( \begin{array}{c} 2 \\ \\ 1 \end{array} \right), \left( \begin{array}{c} 2 \\ \\ 1 \end{array} \right) \right) = -16. \tag{4.4.120}$$

Namely, the splitting (4.4.116) does not satisfy the P-regularity.

One can notice that $\tilde{\mathcal{E}} = I - \mathcal{B}^{-1}\mathcal{A}$ is semi-convergent for $\mathcal{A}$. More precisely,

$$\widetilde{\mathcal{E}}^{\ell} = \widetilde{\mathcal{E}}, \quad \forall\, l \geq 1 \quad \text{and} \quad \widetilde{\mathcal{E}} = P_{\mathcal{N}}. \tag{4.4.121}$$

Namely, (A2) and (A3) are sufficient for the convergence of iterative method based on the splitting (4.4.116). The following example shall show that ($A$3.2) can not be eliminated for the convergence. Let us now consider the following splitting (4.4.116) with

$$\mathcal{B}^{-1} = \left( \begin{array}{cc} 2 & 2 \\ \\ -1 & 0 \end{array} \right). \tag{4.4.122}$$

Some simple manipulation shows that

$$\mathcal{R}(\widetilde{T}) = \mathcal{N}(\mathcal{A}).$$

This implies that the splitting (4.4.116) based on (4.4.122) satisfies clearly (A3.1) but it does not satisfy (A3.2). It is then straightforward to see that the energy norm of $\widetilde{\mathcal{E}}$ is 1.

Namely,

$$|\widetilde{\mathcal{E}}|^2_{\mathcal{A}} = \sup_{v \in \mathcal{N}(\mathcal{A})^{\perp}} \frac{(\mathcal{A}\widetilde{\mathcal{E}}v, \widetilde{\mathcal{E}}v)}{(\mathcal{A}v, v)} = 1. \tag{4.4.123}$$

This means that the method does not converge.

The aim of the following theorem is to provide the infinite dimensional analogue of the P-regularity.

**Theorem 4.4.6.** *If we assume that*

$$\inf_{c \in \mathcal{N}} b(c, c) \geq 0, \tag{4.4.124}$$

*and* $\widetilde{\mathcal{N}} = \{0\}$*, then* $(A3)$ *implies that the splitting* (4.4.103) *is P-regular.*

*Proof.* By the Lemma 4.4.5, the assumption $(A3.1)$ implies that $\mathcal{B} + \mathcal{B}^T - \mathcal{A}$ is positive definite on $\mathcal{R}(\mathcal{A})$. Now we shall show that under the assumption $(A3.2)$ and (4.4.124), $\mathcal{B} + \mathcal{B}^T$ is positive definite on $\mathcal{N}(\mathcal{A})$. We assume that $\mathcal{B} + \mathcal{B}^T$ is non-negative definite on $\mathcal{N}(\mathcal{A})$. Then since $\inf_{c \in \mathcal{N}} b(c, c) \geq 0$, there exists $\mathrm{c} \in \mathcal{N}(\mathcal{A})$ such that

$$\left(\left(\mathcal{B} + \mathcal{B}^T\right)\mathrm{c}, \mathrm{c}\right) = 0. \tag{4.4.125}$$

This implies that $(\mathcal{B}\mathrm{c}, \mathrm{c}) = 0$. From this together with (4.4.124), we conclude that $(\mathcal{B}\mathrm{c}, \tilde{d}) = 0, \quad \forall \tilde{d} \in \mathcal{N}(\mathcal{A})$, namely, $\mathcal{B}\mathrm{c} \in \mathcal{R}(\mathcal{A})$. This leads to the contradiction to $(A3.2)$ since it implies $\mathrm{c} \in \mathcal{R}(\widetilde{T})$. This completes the proof. □

By the Theorems 4.4.5 and 4.4.6, we conclude that under the additional assumption (4.4.124), the assumptions (A2) with $\widetilde{\mathcal{N}} = \{0\}$ and (A3) are equivalent to the

P-regularity and obtain an infinite dimensional P-regularity. In short, we have showed that the well-known sufficient condition for the convergence of iterative method applied to the singular problem is much stronger.

In the rest of this section, we shall consider the convergence rate of the Gauss-Seidel method. The Gauss-Seidel method can be also viewed as an iterative method based on the matrix splitting (4.4.103) with

$$\mathcal{B} = \mathcal{D} - \mathcal{L}^T \quad \text{and} \quad \mathcal{C} = \mathcal{L}, \tag{4.4.126}$$

where $\mathcal{D}$ is the diagonal, $\mathcal{L}$ is the lower triangular part of $\mathcal{A}$ and $\mathcal{L}^T$ denotes the transpose of $\mathcal{L}$.

In this case, the error transfer matrix $\widetilde{\mathcal{E}}$ is

$$\widetilde{\mathcal{E}} = I - (\mathcal{D} - \mathcal{L})^{-1}\mathcal{A}. \tag{4.4.127}$$

The convergence will then follow if $\mathcal{B} + \mathcal{B}^T - \mathcal{A}$ is positive definite, namely $\mathcal{D}$ is positive. For the convergence rate, the following observation is crucial to apply the Theorem 4.4.4, namely, we have

$$\widetilde{\mathcal{E}} = (I - (\mathcal{D} - \mathcal{L})^{-1}\mathcal{A}) = (I - P_n)\cdots(I - P_1), \tag{4.4.128}$$

where $P_i = \dfrac{a(e_i, \cdot)}{a(e_i, e_i)} e_i$ with $\{e_1, \cdots, e_n\}$ being the canonical basis for $\mathbb{R}^n$. Based on this observation, we obtain the following result.

**Corollary 4.4.2.** *If $\mathcal{A}$ is symmetric and positive semi definite with positive diagonal, then the convergence rate of the Gauss-Seidel iterative method is given as follows:*

$$|\widetilde{\mathcal{E}}|_{\mathcal{A}}^2 = \frac{c_0}{1+c_0},$$

*where*

$$c_0 = \sup_{v \in \mathcal{N}(\mathcal{A})^{\perp}} \inf_{c \in \mathcal{N}(\mathcal{A})} \frac{(\mathcal{S}(v-c),(v-c))}{(v,v)_a}, \tag{4.4.129}$$

*where* $\mathcal{S} = \mathcal{L}\mathcal{D}^{-1}\mathcal{L}^T$.

### 4.4.9 Multigrid method for Neumann problems

Consider the following differential equation with Neumann boundary conditions:

$$-\Delta u = f \quad \text{in } \Omega, \quad \frac{\partial u}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega, \tag{4.4.130}$$

where $\dfrac{\partial u}{\partial \mathbf{n}}$ is the exterior normal derivative of $u$ and $\Omega$ is a polygonal domain in $\mathbb{R}^d$ with $d = 1, 2$ or 3. The variational problem corresponding to (4.4.130) can be given as follows: find $u \in H^1(\Omega)$ such that

$$a(u,v) = \langle f, v\rangle, \quad \forall v \in H^1(\Omega), \tag{4.4.131}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v dx \quad \text{and} \quad \langle f, v\rangle = \int_\Omega f v dx. \tag{4.4.132}$$

The null space $\mathcal{N}$ of $a(\cdot,\cdot)$ is given by

$$\mathcal{N} = \text{span}\{1\} \tag{4.4.133}$$

and it is well-known that for solvability of (4.4.131), it is necessary that $f$ satisfy the following compatibility condition, namely

$$\int_\Omega f\,dx = 0. \tag{4.4.134}$$

It is then easy to see that the solution to (4.4.130) exists and is unique (in the weak sense) on the quotient space $H^1(\Omega)/\mathcal{N}$.

Throughout this section, we also assume that $\Omega$ is triangulated with a nested sequence of quasi-uniform triangles $\mathcal{T}_k = \{\tau_k^i\}$ of size $h_k$, where the quasi-uniformity constants are independent of $k$ and $h_k \sim \gamma^k$ with $\gamma \in (0,1)$ for $k = 1, \cdots, J$. Associated with each $\mathcal{T}_k$, we have the finite element space of continuous piecewise linear functions $V_k \subset H^1(\Omega)$. In this setting, we have that

$$V_1 \subset \cdots \subset\ V_k \subset \cdots \subset V_J, \tag{4.4.135}$$

and each subspace $V_i$ contains the null space, namely the constant function. We shall then be interested in solving the following equations resulting from the finite element discretization : find $u_h \in V$ with $h = h_J$ and $V = V_J$ such that

$$a(u_h, v) = \langle f, v\rangle, \quad \forall v \in V. \tag{4.4.136}$$

Under the settings outlined above, we consider a multigrid method with the Gauss-Seidel method as a smoother. From the abstract theory given in the previous section, we can obtain the following convergence result.

**Theorem 4.4.7.** *The convergence rate of the multigrid method with one Gauss-Seidel smoothing is given as follows:*

$$|E_J|_a^2 = \frac{c_0}{1+c_0}, \tag{4.4.137}$$

*where*

$$c_0 = \sup_{v\in\mathcal{N}^\perp} \inf_{c\in\mathcal{N}} \inf_{\sum_{k=1}^J \sum_{i=1}^{n_k} v_k^i = v+c} \frac{\sum_{k=1}^J \sum_{i=1}^{n_k} |P_k^i(\sum_{(l,j)>(k,i)} v_l^j)|_a^2}{(v,v)_a}, \tag{4.4.138}$$

*where $n_k$ is the number of nodal points in each subspace $V_k$ with $k = 1, \cdots, J$ and moreover, $c_0$ is bounded independently of the mesh size $h$ and the number of levels $J$.*

*Proof.* By the observation (4.4.99) in the previous section, we may set the space decomposition as follows: $V = \sum_{k=1}^{J} \sum_{i=1}^{n_k} V_k^i$, where $V_k^i$ corresponds to the function space spanned by a continuous piecewise linear function $\phi_k^i$ with $\phi_k^i(x) = 1$ at $x = x_k^i$ and zero at the other nodes. In this setting, since $\mathcal{N}_k^i = \{0\}$ and $T_k^i = P_k^i$, by a direct application of the Corollary 4.4.1, the $c_0$ is given by (4.4.138). Now, given $v \in \mathcal{N}^\perp$ and $c \in \mathcal{N}$, let $w = v + c \in V$. Now consider the following decomposition,

$$w = \sum_{k=1}^{J} v_k = \sum_{k=1}^{J} \sum_{i=1}^{n_k} v_k^i, \quad v_k^i \in V_k^i,$$

where

$$v_k = \sum_{i=1}^{n_k} v_k^i = (Q_k - Q_{k-1})w.$$

and $Q_k$ is the $L^2$ projection onto $V_k$. Note that $Q_J w = w$ and

$$\sum_{(l,j)>(k,i)} v_l^j = \sum_{j=i+1}^{n_k} v_k^j + \sum_{l=k+1}^{J} \sum_{j=1}^{n_l} v_l^j = \sum_{j=i+1}^{n_k} v_k^j + w - Q_k w.$$

Since

$$a(P_k^i P_k w, v_k^i) = a(P_k w, v_k^i) = a(w, v_k^i) = a(P_k^i w, v_k^i), \quad \forall w \in V, \quad \forall v_k^i \in V_k^i$$

and $a(\cdot,\cdot)$ is $V_k^i$-elliptic, we have that $P_k^i P_k w = P_k^i w \quad \forall w \in V$. From this relation, we have

$$\begin{aligned} P_k^i \sum_{(l,j)>(k,i)} v_l^j &= P_k^i \sum_{j=i+1}^{n_k} v_k^j + P_k^i (w - Q_k w) \qquad (4.4.139) \\ &= P_k^i \sum_{j=i+1}^{n_k} v_k^j + P_k^i P_k (w - Q_k w) \\ &= P_k^i \sum_{j=i+1}^{n_k} v_k^j + P_k^i (P_k w - Q_k w). \end{aligned}$$

This then yields (with $\Omega_k^i = \mathrm{supp}\phi_k^i$)

$$\begin{aligned}
\sum_{i=1}^{n_k} |P_k^i \sum_{(l,j)>(k,i)} v_l^j|_a^2 &= \sum_{i=1}^{n_k} |P_k^i \sum_{j=i+1}^{n_k} v_k^j + P_k^i(P_k w - Q_k w)|_a^2 \\
&\lesssim \left( \sum_{i=1}^{n_k} |P_k^i \sum_{j=i+1}^{n_k} v_k^j|_a^2 + \sum_{i=1}^{n_k} |P_k^i(P_k w - Q_k w)|_a^2 \right) \\
&\lesssim \left( \sum_{i=1}^{n_k} |\sum_{j\in N_k(i)} v_k^j|_{a,\Omega_k^i}^2 + \sum_{i=1}^{n_k} |(P_k w - Q_k w)|_{a,\Omega_k^i}^2 \right),
\end{aligned}$$

where $N_k(i) = \{j \in \{1,\cdots,J\} : \Omega_k^j \cap \Omega_k^i \neq \emptyset\}$. Now by $(Q_k - Q_{k-1})^2 = Q_k - Q_{k-1}$ and a standard interpolation argument it follows that

$$\begin{aligned}
\sum_{i=1}^{n_k} |\sum_{j\in N_k(i)} v_k^j|_{a,\Omega_k^i}^2 &\lesssim \sum_{i=1}^{n_k} \sum_{j\in N_k(i)} |v_k(x_k^j)|^2 h_k^{d-2} \lesssim h_k^{-2} \sum_{i=1}^{n_k} h_k^d |v_k(x_k^j)|^2 \\
&= C h_k^{-2} \|v_k\|_0^2 = C h_k^{-2} \|(Q_k - Q_{k-1}) v_k\|_0^2 \\
&\lesssim h_k^{-2} h_{k-1}^2 |v_k|_a^2 = \gamma^{-2} |v_k|_a^2.
\end{aligned}$$

Hence

$$\begin{aligned}
\sum_{i=1}^{n_k} |P_k^i \sum_{(l,j)>(k,i)} v_l^j|_a^2 &\lesssim \left( |v_k|_a^2 + |(P_k - Q_k)w|_a^2 \right) \qquad (4.4.140) \\
&\lesssim \left( |(Q_k - Q_{k-1})w|_a^2 + |(P_k - Q_k)w|_a^2 \right).
\end{aligned}$$

The proof is completed by applying the following known estimates (see Bramble and Zhang [12] or Xu [89]):

$$\sum_{k=1}^{J} \left( |(Q_k - Q_{k-1})w|_a^2 + |(P_k w - Q_k w)|_a^2 \right) \lesssim |w|_a^2, \quad \forall w \in H^1(\Omega).$$

□

# Chapter 5

# Numerical studies on a falling sphere through viscoelastic fluids

## 5.1 Introduction

The motion of a falling sphere through viscoelastic fluids is a well-studied problem with a broad range of practical application, [63]. Over the past decade, significant advances have been made in numerical algorithms for steady and transient computation of viscoelastic flows and also in quantitative experimental techniques for resolving the spatial and temporal characteristics of both the particle motion and the velocity field within the fluid. These advances, when combined with careful rheological characterization of the test fluids studied and simulated, have led to a greater understanding of the motion of particles in complex fluids.

A falling sphere in a polymeric fluid in general reach a terminal velocity, although sometimes it accompanies some transient oscillations ,[73, 74, 2, 63, 10]. Very recently, a striking experimental result has been announced. Namely, a sphere falling in a worm-like micellar fluid does not approach a steady terminal velocity; instead it undergoes continual oscillations as it falls, as shown in Figure 5.1.1 below, [45, 21].

There have been several reports on flow instabilities, [31] observed in a worm-like micellar fluid such as the shear banding, [69, 60] and spurt instabilities, [72]. Those instabilities have been attributed to a flat region displayed by the shear-stress flow curve

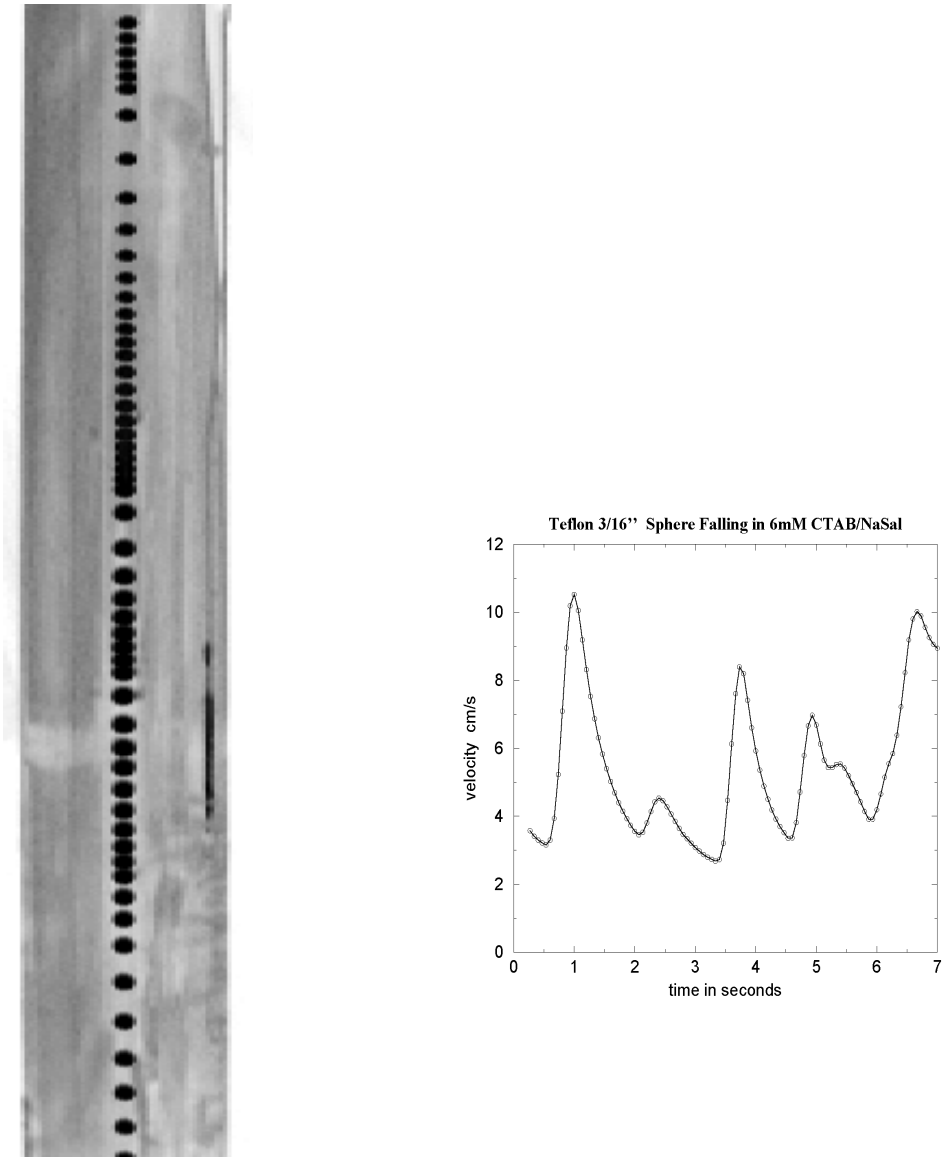


Fig. 5.1.1. (Left) Collage of video images showing the decent of a 3/16-inch-diameter teflon sphere in an aqueous solution of $6.0\,mM$ CTAB/NaSal (image shown is $50cm$ in height, with $\Delta t = 0.13\,s$). (Right) Velocity vs time for a 1/4-inch-diameter teflon sphere falling through $9.0\,mM$ CTAB/NASAL. Originally published in A. Jayaraman and A. Belmonte, [45]. Reprinted with permission from the authors

in worm-like micellar fluids [69, 60]. An oscillation of falling sphere in a worm-like micellar fluid is certainly most recent one of compelling evidences of flow instabilities and it is also believed to be related to a flat region displayed by the shear-stress flow curve in worm-like micellar fluids , [45] and that a flat region displayed by the shear-stress flow curve in worm-like micellar fluids is a manifestation of a nonmonotonic stress-strain rate relation is widely accepted, [69, 60].

The objective of this chapter is motivated by such a belief. Namely, by simulating the Johnson-Segalman model that shows apparent nonmonotonic stress-strain rate relation,(see the Figure 5.2.1 below) and is so suggested to be simulated in [45], we take an initial attempt to search for mathematical models that are responsible for a sustained oscillation of a falling sphere.

Although it is the main goal in this chapter to test the Johnson-Segalman model against a continual oscillation, we have also presented various new numerical observations. It is surprising to note that no numerical studies of the falling sphere with the Johnson-Segalman models is available. Namely, we shall provide missing studies of the motion of a falling sphere problem. In particular, we shall discuss the effects of the slip parameter "$a$" on the oscillations of a falling sphere and explain heuristically how it can be used as a criteria for the formation of negative wake, [38].

This chapter should also be considered to be complementary to the previous chapters 2 and 3 devoted for algorithmic developments. By simulating the falling sphere problem, we shall provide various numerical evidences that our scheme is reliable and robust.

The rest of this chapter is organized as follows. In §5.2, we derive the governing equation for the falling sphere simulation in the sphere frame. The governing equation is then non-dimensionalized in §5.2.1. The main feature of the Johnson-Segalman model exhibiting a non-monotonic shear stress-strain rate for the steady shear flow and the corresponding change of type will then be discussed in §5.2.2. The section 5.3 discusses some algorithmic details and implementations of our schemes for numerical simulations. Finally, in §5.4, we report various new numerical observations from simulating the falling sphere through the Johnson-Segalman model.

In view of the main motivation to test the Johnson-Segalman model in search for a model responsible for a continual oscillation of a falling sphere, our conclusion is that a property of the Johnson-Segalman model exhibiting the non-monotonic shear stress-strain rate relation can not alone be used to explain the continual oscillation of a falling sphere in a worm-like micellar fluid, [45].

## 5.2 Governing equation

The main purpose of this section is to describe a governing equation, especially, in the sphere frame. This type of formulation has been used in the work by Arigo et al, [74] and proved to be effective in the computational viewpoint for the following two reasons. First, it allows us to assume the height of cylinder is infinite. Second, it is not necessary to change the mesh according to the movement of the sphere. The governing equation (Oldroyd-B) written in the lab frame has been used for the work by Bodart and Crochet, [10], in which some moving mesh method had to be used.

Let us assume that a sphere has the radius $r$ and density $\rho_s$ accelerating from rest under the influence of gravity $g$ in a fluid of density $\rho_f$ inside of a cylinder with the radius $R$ and that the extra stress $\sigma$ is decomposed into two parts as follows :

$$\sigma = 2\mu_s \mathcal{D}(\mathbf{u}) + \tau, \tag{5.2.1}$$

where $2\mu_s \mathcal{D}(\mathbf{u})$ is from the Newtonian contribution and $\tau$ is from the polymeric contribution.

In a fixed cylindrical coordinate system with the $z-$axis pointing vertically downward in the direction of gravity, the dimensional force balance equation on the sphere shall be

$$\frac{4\pi}{3} r^3 \rho_s \frac{dU_s(t)}{dt} = \frac{4\pi}{3} r^3 (\rho_s - \rho_f) g + F_d, \tag{5.2.2}$$

where $U_s(t)$ is the axial component of the velocity vector $\mathbf{U}_s(t)$ of the sphere given by $U_s = (e_z, \mathbf{U}_s)$, the magnitude of the gravitational acceleration is denoted by $g$ and $F_d$ is the axial component of the drag force exerted by the fluid on the sphere and it is given by the following :

$$F_d = \left( e_z, \int_{\partial B(0,r)} (pI - \sigma)\, \mathbf{n} dS \right), \tag{5.2.3}$$

where $p$ is the isotropic pressure, $\partial B(0, r)$ is the sphere with radius $r$ and centered at the origin and $\mathbf{n}$ is the inward pointing unit normal vector at the surface of the sphere. Note that the direction of $\mathbf{n}$ is outward in view of the sphere.

The dimensional form of the conservation equations for mass and momentum become :

$$\rho_f \left( \frac{\partial \mathbf{U}}{\partial t} + (\mathbf{U} \cdot \nabla)\mathbf{U} \right) = -\nabla p + \mathrm{div}\sigma + \rho_f \frac{dU_s}{dt} e_z \tag{5.2.4}$$

$$\mathbf{divU} = 0 \tag{5.2.5}$$

respectively. Note that moving to the sphere frame from the fixed lab frame introduces a pseudo force, $\rho_f \frac{dU_s}{dt} e_z$ which is proportional to the acceleration of the sphere in the momentum equations. The Johnson-Segalman constitutive relation for $\tau$, the polymeric contribution of extra stress (5.2.1), [49] is given by

$$\tau + \lambda \frac{\delta_E \tau}{\delta_E t} = 2\mu_p \mathcal{D}(\mathbf{u}), \tag{5.2.6}$$

where $\lambda$ is the characteristic relaxation time and $\mu_p$ is the polymeric viscosity and the objective derivative $\frac{\delta_E}{\delta_E t}$ is so-called the Gordon-Schowalter derivative [32] and acts upon $\tau$ by the following :

$$\frac{\delta_E \tau}{\delta_E t} = \frac{D\tau}{Dt} - \left( \frac{a+1}{2} \nabla \mathbf{u} + \frac{a-1}{2} \nabla \mathbf{u}^T \right) \tau - \tau \left( \frac{a+1}{2} \nabla \mathbf{u} + \frac{a-1}{2} \nabla \mathbf{u}^T \right)^T. \tag{5.2.7}$$

Here $a$ is often called the slip parameter. The physical explanation for the parameter $a$ is that the strands "slip" with respect to motion of the continuum. Namely, the continuum "slips" past the strands, the continuum does not feel all the tension in the strands, [54]. In the final section of this chapter, we shall argue that this parameter $a$ and so-called the extensibility parameter denoted by $L$ in FENE models are closely related.

### 5.2.1 Non-dimensionalization

The purpose of this section is to non-dimensionalization the equations (5.2.4), (5.2.5) and (5.2.6). We shall use the following scales for the non-dimensionalization.

- Length scale : Radius of Sphere $r$.
- Velocity scale : Stokes terminal velocity

$$V_N = \frac{2}{9}\frac{r^2(\rho_s - \rho_f)g}{\mu K_N(r/R)} \quad \text{with} \quad \mu = \mu_s + \mu_p.$$

- Time scale : $\dfrac{r}{V_N}$.
- Stress and Pressure scale : $\dfrac{\mu V_N}{r}$.

In the velocity scale, $K_N(r/R)$ is called the Faxen wall correction factor. In our computations, we have set it to be 1, [37]. We shall now introduce the following non-dimensional variables $\tilde{\mathbf{U}}, \tilde{t}$ and $\tilde{\sigma}$ :

$$\mathbf{U} = V_N \tilde{\mathbf{u}} \qquad t = \frac{r}{V_N}\tilde{t} \qquad \sigma = \frac{\mu V_N}{r}\tilde{\sigma}. \tag{5.2.8}$$

Upon using dimensionless variables $\tilde{\mathbf{U}}, \tilde{t}$ and $\tilde{\sigma}$, we can rewrite (5.2.2) and (5.2.3) as followings :

$$\left(\frac{2}{3}\mathrm{Re}\,\vartheta\right)\frac{d\tilde{\mathbf{U}}_s}{d\tilde{t}} = 3 + \tilde{F}_d \tag{5.2.9}$$

and

$$\tilde{F}_d = \frac{1}{2\pi}\left(e_z, \int_{\partial B(0,r)} (\tilde{p}I - \tilde{\sigma})\,\mathbf{n} d\tilde{S}\right), \tag{5.2.10}$$

where

$$\mathrm{Re} = \frac{\rho_f V_N r}{\mu} \quad \text{and} \quad \vartheta = \frac{\rho_s}{\rho_f}. \tag{5.2.11}$$

Because

$$\frac{4\pi}{3} r^3 \rho_s \frac{V_N^2}{r} \frac{d\tilde{U}_s}{d\tilde{t}} = \frac{4\pi}{3} r^3 (\rho_s - \rho_f) g + \tilde{F}_d \frac{2\pi}{r} \mu V_N r^2, \tag{5.2.12}$$

where

$$\tilde{F}_d = \frac{1}{2\pi}\left(e_z, \int_{\partial B(0,r)} (\tilde{p}I - \tilde{\sigma})\ \mathbf{n} d\tilde{S}\right). \tag{5.2.13}$$

Furthermore, after renaming all the variables without tilde, the equations of mass and momentum (5.2.4), (5.2.5) and the constitutive relation (5.2.6) can be rewritten as follows

$$\mathrm{Re}\left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) = -\nabla p + \eta_s \Delta \mathbf{u} + \mathrm{div}\tau + \mathrm{Re}\frac{dU_s}{dt}\, e_z \tag{5.2.14}$$

$$\mathbf{divu} = 0 \tag{5.2.15}$$

$$\tau + \mathrm{We}\frac{\delta_E \tau}{\delta_E t} = 2\eta_p \mathcal{D}(\mathbf{u}), \tag{5.2.16}$$

where

$$\eta_s = \frac{\mu_s}{\mu}, \quad \eta_p = 1 - \eta_s, \quad \text{and} \quad \mathrm{We} = \frac{\lambda V_N}{r}. \tag{5.2.17}$$

The equations (5.2.14), (5.2.18) and (5.2.16) are then coupled with the dimensionless force balance equation given by the following :

$$\left(\frac{2}{3}\mathrm{Re}\,\vartheta\right)\frac{dU_s}{dt} = 3 + F_d, \tag{5.2.18}$$

where

$$\vartheta = \frac{\rho_s}{\rho_f} \quad \text{and} \quad F_d = \frac{1}{2\pi}\left(e_z, \int_{\partial B(0,r)} (pI - \sigma)\,\mathbf{n} dS\right). \tag{5.2.19}$$

Now let us complete the formulation for a falling sphere simulation by giving initial and boundary conditions. We shall set $t = 0$ is the time when the sphere is released. All velocities and stress are then set to be zero as the initial condition since the fluid and sphere are at rest at time $t = 0$. In the moving reference frame used, the fluid velocity is set to zero on the sphere at all times. At the inflow, outflow and tube walls, the dimensionless axial velocity is $-U_s$ and radial velocity component is zero. Symmetry conditions are given at the centerline. At inlet, zero value of stress is imposed as an inflow boundary condition.

### 5.2.2 Non-monotone shear stress-shear rate curves

The dynamics of spurt in models with non-monotone shear stress-shear rate curves in the Johnson-Segalman model have been studied extensively [54, 27, 31, 29, 60, 33, 92]. These results concern the analysis of parallel shear flows with discontinuous shear rates, the approach to such flows from general initial data, and the possibility of oscillations. In all these works, it is assumed that a priori that the flow is parallel shear flow. Such a non-monotone relation is then also believed to produce an oscillation of a sphere in a

worm-like micellar fluid [4, 45]. Although the situation for the falling sphere experiment and the shear flow is quite different, negating or approving what is believed to be related to a continual oscillation of a falling sphere shall be valuable. In doing so, we shall extract parameter ranges for which a typical non-monotonic relation is achieved for the shear flow of the Johnson-Segalman model and shall then identify relevant parameters to simulate the falling sphere problems.

Following the recent monograph by M. Renardy, [71], we consider the parallel shear flow, in which the velocity field is given by $\mathbf{u} = (v(y), 0, 0)$. Note that the incompressibility condition $\mathbf{divu} = 0$ is automatically satisfied and the velocity gradient is given by

$$\nabla \mathbf{u} = \begin{pmatrix} 0 & v'(y) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.2.20}$$

Note that the shear rate $\dot{\gamma}$ is $v'(y)$ and the stress tensor $\tau$ in simple shear flow has the form :

$$\tau = \begin{pmatrix} \tau_{11}(\dot{\gamma}) & \tau_{12}(\dot{\gamma}) & 0 \\ \tau_{12}(\dot{\gamma}) & \tau_{22}(\dot{\gamma}) & 0 \\ 0 & 0 & \tau_{33}(\dot{\gamma}) \end{pmatrix}. \tag{5.2.21}$$

For the steady parallel shear flows with a pressure $p = Px + q(y)$ with $P$ being a constant, the equations become

$$\begin{aligned}
\tau_{12}' + \eta_s v'' - P &= 0, && (5.2.22)\\
\tau_{22}' - q' &= 0, && (5.2.23)\\
\tau_{11} - \mathrm{We}(1+a)v'\tau_{12} &= 0, && (5.2.24)\\
\tau_{12} - \mathrm{We}v'\left(\frac{(1+a)}{2}\tau_{22} + \frac{a-1}{2}\tau_{11}\right) &= \eta_p v' && (5.2.25)\\
\tau_{22} + \mathrm{We}(1-a)v'\tau_{12} &= 0 && (5.2.26)
\end{aligned}$$

The symbol associated with this system is shown to have the determinant that changes its sign when $T'(\dot{\gamma})$ changes its sign, see the equation 5.2.28 below for an analytic expression of $T'(\dot{\gamma})$. Namely, the change of type occurs when the shear stress becomes a decreasing function of shear rate, [71]. The change of type is then apparently related to the non-monotonic relation between the shear stress and the strain rate, see the Figure (5.2.1) below. Furthermore, the shear stress $\tau_{12}$ at shear rate $\dot{\gamma}$ can be explicitly from the equations (5.2.22 - 5.2.26) as follows :

$$\tau_{12}(\dot{\gamma}) = \frac{\eta_p \dot{\gamma}}{1 + \mathrm{We}^2(1-a^2)\dot{\gamma}^2}. \qquad (5.2.27)$$

From this, we obtain that the total shear rate $T_{12}$ at shear rate $\dot{\gamma}$ is

$$T(\dot{\gamma}) = \eta_s \dot{\gamma} + \frac{\eta_p \dot{\gamma}}{1 + \mathrm{We}^2(1-a^2)\dot{\gamma}^2}. \qquad (5.2.28)$$

It is easy to obtain some specific parameters that show the typical non-monotonic relation. See the Figure 5.2.1. The graph clearly shows the non-monotonic shear stress-strain rate relation for We $= 3$ and $\eta_s = 0.05$ with both choices of $a = 0.9$ and $a = 0.6$.

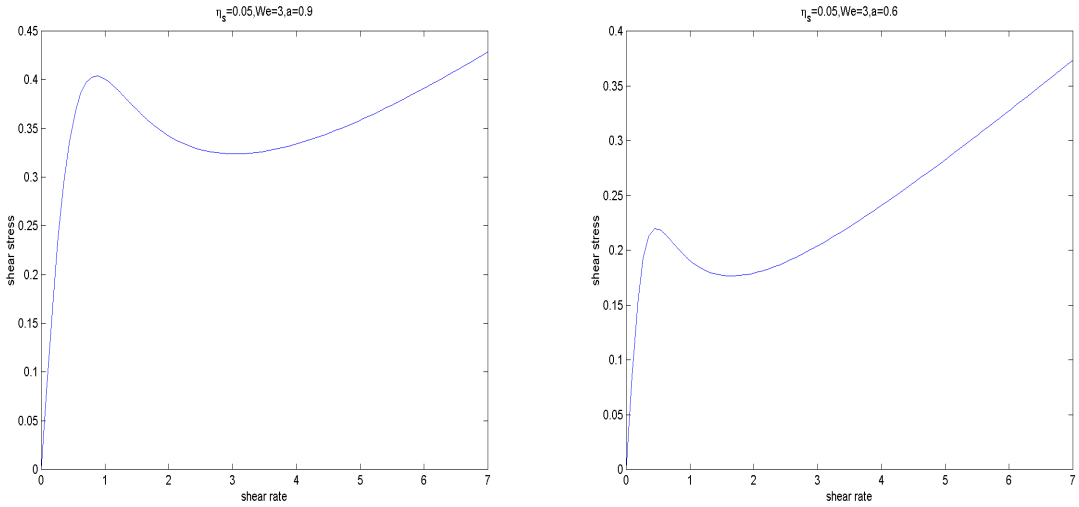


Fig. 5.2.1. non-monotonic shear stress-strain rate relations $a = 0.9$ (left) and $a = 0.6$ (right)

One may notice that as $a$ decreases, the shear thinning occurs at progressively smaller strain rates, $\dot{\gamma}$, [54]. In our computation, we take two sets of fixed data with varying $a$. First set of parameters are given in Table (5.2.1), which show the non-monotonic relation and second set of parameters are given in Table (5.2.2), which does not show the non-monotonic relation, see the Figure (5.2.2).

As shown in the two Tables 5.2.1 and 5.2.2, other parameters such as Re, the Reynolds number, We, the Weissenberg number, $\frac{R}{r}$, an aspect ratio between sphere

Table 5.2.1. Parameters showing a non-monotonic relation

| | | | | |
|---|---|---|---|---|
| $\mathrm{Re} = 0.01$ | $\mathrm{We} = 3$ | $\frac{R}{r} = 0.2$ | $\vartheta = 5$ | $\eta_s = 0.05$ |

Table 5.2.2. Parameters **Not** showing a non-monotonic relation

| | | | | |
|---|---|---|---|---|
| $\mathrm{Re} = 0.01$ | $\mathrm{We} = 4$ | $\frac{R}{r} = 0.2$ | $\vartheta = 5$ | $\eta_s = 0.15$ |

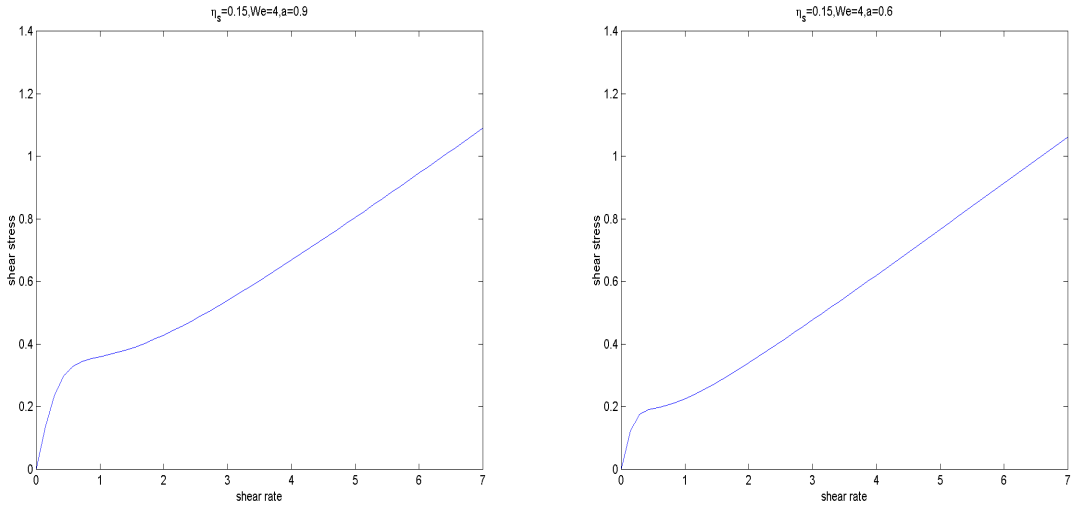


Fig. 5.2.2. monotonic shear stress-strain rate relations with $a = 0.9$ (left) and $a = 0.6$ (right) respectively

radius $r$ and cylinder radius $R$ and the density ratio $\vartheta = \dfrac{\rho_s}{\rho_f}$ are chosen mostly based upon the experimental work of Jayaraman and Belmonte, [45].

In our numerical experiments, both sets of parameters indeed did not produce a continual oscillation, but transient oscillations, see the section of numerical experiments below for more details. Such a result leads us to be more interested in the effects of "$a$" in the pattern of oscillations of a falling sphere.

## 5.3 Some detailed description of the numerical algorithms

In this section, we shall describe numerical algorithm employed to simulate the falling sphere through a cylinder. Since our main motivation is to capture some physically relevant features with a modest Weissenberg number, we simply use the first order explicit schemes developed in the chapter 3.

### 5.3.1 Review on the numerical algorithm

The first order explicit scheme developed in the chapter 3 is not shown to be stable, however, it is cheap and proven to preserve an important physical quantity, namely the positivity of the conformation tensor. The scheme that will be used in this section can be given as follows :

Let us assume that $(\mathbf{u}_h^n, p_h^n, \tau_{A,h}^n) \in \mathbf{V}_h \times W_h \times \mathbf{S}_h$ are defined. We then proceed from $(\mathbf{u}_h^n, p_h^n, \tau_{A,h}^n)$ to $(\mathbf{u}_h^{n+1}, p_h^{n+1}, \tau_{A,h}^{n+1}) \in \mathbf{V}_h \times W_h \times \mathbf{S}_h$ by the following relations :

$$\begin{aligned} \mathrm{Re}\frac{\mathbf{u}_h^{n+1} - \Pi_h^{\mathbf{V}}\left(\mathbf{u}_h^n \circ y^n\right)}{k} &+ \nabla_h p_h^{n+1} + \eta_s A_h \mathbf{u}_h^{n+1} \\ &= \mathrm{div}_h \tau_{A,h}^{n+1} + \mathrm{Re}\left(\frac{dU_s}{dt}\right)^n e_z \end{aligned} \tag{5.3.1}$$

$$\mathbf{divu}_h^{n+1} = 0 \tag{5.3.2}$$

$$\begin{aligned} &\frac{\tau_{A,h}^{n+1} - \Pi_h^{\mathbf{S}}\left(E_h(t^n, t^{n+1})(\tau_{A,h}(t^n) \circ y^n) E_h(t^n, t^{n+1})^T\right)}{k} \\ &\qquad = -\alpha \tau_{A,h}^{n+1} + \beta I, \end{aligned} \tag{5.3.3}$$

where $\alpha = \dfrac{1}{\mathrm{We}}$ and $\beta = \dfrac{\mu_p}{a\mathrm{We}^2}$

$$E_h(t^n, t^{n+1}) = I + k\left(\frac{a+1}{2}\nabla_h \mathbf{u}_h^n \circ y^n + \frac{a-1}{2}(\nabla_h \mathbf{u}_h^n \circ y^n)^T\right) \tag{5.3.4}$$

Note that the pseudo force $\mathrm{Re}\dfrac{dU_s}{dt}\, e_z$ is computed in an explicit manner. We also note that from (5.3.3) and (5.3.4), the term $\mathrm{div}\tau_{A,h}^{n+1}$ of (5.3.1) are computed explicitly from the previous velocity field $\mathbf{u}_h^n$. Namely, except for the equations of mass and momentum, all equations are computed explicitly. For the computation of the equations (5.3.1) and (5.3.2), we apply our fast and robust solver, the preconditioned MINRES

developed and analyzed in the chapter 4. More precisely, we apply following procedures to handle the aforementioned discretizations (5.3.1), (5.3.2) and (5.3.3)

**Algorithm 5.3.1.** *Let* $\mathbf{u}_h^n, p_h^n$ *and* $\tau_{A,h}^n$ *be given.*

*Step 1. Using the previous velocity,* $\mathbf{u}_h^n$ *and* $\mathbf{u}_h^{n-1}$*, we perform the followings :*

*(i) Compute the characteristic feet* $y^n$*.*

*(ii) Compute* $\Pi_h^{\mathbf{V}}(\mathbf{u}^n \circ y^n)$ *by an appropriate interpolation.*

*(iii) Take a discrete gradient to compute* $\nabla_h \left(\Pi_h^{\mathbf{V}}(\mathbf{u}^n \circ y^n)\right)$*.*

*(iv) Compute the drag by :*

$$F_d^n = \frac{1}{2\pi}\left(e_z, \int_{\partial B(0,r)} (p_h^n I - \sigma_h^n)\mathbf{n}\, dS\right). \tag{5.3.5}$$

*Step 2. (i) Compute the deformation gradient* $E_h$ *as follows :*

$$E_h(t^n, t^{n+1}) = I + k\left(\frac{a+1}{2}\nabla_h \mathbf{u}_h^n \circ y^n + \frac{a-1}{2}(\nabla_h \mathbf{u}_h^n \circ y^n)^T\right) \tag{5.3.6}$$

*(ii) Update the stress* $\tau_{A,h}^{n+1}$ *from (5.3.3).*

*(iii) Compute the acceleration of sphere by :*

$$\left(\frac{2}{3}\mathrm{Re}\,\vartheta\right)\left(\frac{dU_s}{dt}\right)^n = 3 + F_d^n \tag{5.3.7}$$

*(iv) Applying the explicit euler method for (5.3.7) to obtain* $U_s^{n+1}$*, namely,*

$$\frac{U_s^{n+1} - U_s^n}{k} \approx \left(\frac{dU_s}{dt}\right)^n = \left(\frac{2}{3}\mathrm{Re}\,\vartheta\right)^{-1}\left(3 + F_d^n\right). \tag{5.3.8}$$

*and use $U_s^{n+1}$ for updating the boundary condition for the equations (5.3.1) and (5.3.2).*

*Step 3. Update $\mathbf{u}_h^{n+1}$ and $p_h^{n+1}$ by solving the equations of mass and momentum (5.3.1) in a coupled manner using the preconditioned MINRES.*

We remark that the Algorithm 5.3.1 is well-defined. In the next subsection, we shall describe more details on how to implement each steps defined in the Algorithm 5.3.1.

### 5.3.2 Detailed description of the numerical implementations

In this subsection, we shall discuss how the Algorithm 5.3.1 is implemented in our numerical experiments, especially, step 1 and step 3.

As is pointed out, our numerical scheme is based on the semi-Lagrangian framework and what is crucial in such a scheme is to compute the characteristic feet $y^n$ and interpolate $u^n \circ y^n$ and $\tau_{A,h}^n \circ y^n$. These tasks can be summarized as the following two steps :

- Backward Integration
- Search-Interpolation Procedure

Namely, we first calculate $y^n$ by some appropriate backward integration and then find the element $K \in \mathcal{T}_h$ hosting it and compute $\Pi_h^{\mathbf{V}}(\mathbf{u}^n \circ y^n)$ and $\Pi_h^{\mathbf{S}}(\tau_{A,h}^n \circ y^n)$.

The computation of $y^n$ is stressed to be important for the stability of the scheme and discussed in the chapter 3. For our numerical experiments, we have used the second

order implicit mid-point rule. More precisely, to compute $y^n \approx y(x,t,s)$, we solve

$$\frac{dy(x,t,s)}{ds} = \mathbf{u}(y(x,t,s),s), \quad y(x,t,t) = x \tag{5.3.9}$$

applying the second order mid-point rule together with the second order extrapolations. This scheme seems to first appear in the work, [83] and its algorithmic details are given below. The second order mid-point rule for (5.3.9) shall result in

$$\frac{x - y(x,t,s)}{k} = \mathbf{u}\left(y\left(x,t,s+\frac{k}{2}\right), s+\frac{k}{2}\right) + O(k^3). \tag{5.3.10}$$

Now the second order extrapolation for $\mathbf{u}(x,s)$ is given as follows :

$$\mathbf{u}\left(x, s+\frac{k}{2}\right) = \frac{3}{2}\mathbf{u}(x,s) - \frac{1}{2}\mathbf{u}(x,s-k) + O(k^2). \tag{5.3.11}$$

Hence the second order accurate approximation $y^n$ to $y(x,t,s)$ is given by the following formula :

$$y^n = x - k\left(\frac{3}{2}\mathbf{u}\left(\frac{x+y^n}{2}, s\right) - \frac{1}{2}\mathbf{u}\left(\frac{x+y^n}{2}, s-k\right)\right). \tag{5.3.12}$$

As is easily observed, the formula is given implicitly. So, we shall use the following iterations to compute $y^n$.

**Algorithm 5.3.2.** *Set*

$$\Delta^0 = k\left(\frac{3}{2}\mathbf{u}(x,s) - \frac{1}{2}\mathbf{u}(x,s-k)\right), \tag{5.3.13}$$

*and perform the following iterations until convergence :*

$$\Delta^m = k\left(\frac{3}{2}\mathbf{u}(x-\frac{1}{2}\Delta^{m-1},s)-\frac{1}{2}\mathbf{u}(x-\frac{1}{2}\Delta^{m-1},s-k)\right). \tag{5.3.14}$$

*Let $\Delta$ be the fixed point, then $y^n = x - \Delta$.*

The Algorithm 5.3.2 needs practically 4 to 5 iterations and shown to be volume preserving for $d = 2$ in the chapter 3. In the current setting, the algorithm (5.3.2) does not make a volume preserving scheme since in our case, $d = 3$.

Note that the algorithm requires to detect the element $K \in \mathcal{T}_h$ in which the points $x - \frac{1}{2}\Delta^l$ is located for all $l$ and interpolate $\mathbf{u}^n$ at the given point $x - \frac{1}{2}\Delta^l$. Such issues are well-studied in the literatures. Especially, we refer to the recent work by Xiu and Karniadakis [88], in which especially the new and fast searching algorithm has been designed for the semi-Lagrangian method that works for unstructured grids such as triangles, quadrilaterals, tetrahedra and hexahedra elements. Once the parent element is detected, we use a simple Lagrange interpolation to obtain $\Pi_h(\mathbf{u}_h^n \circ y^n)$ and $\Pi_h(\tau_{A,h}^n \circ y^n)$.

Finally, we shall consider the solution method to solve the coupled equations of mass and momentum in step 3. In step 3 of the algorithm, we need to solve the following system of equations :

$$\begin{aligned} \frac{\mathrm{Re}}{k}\mathbf{u}_h^{n+1} + \nabla_h p_h^{n+1} + \eta_s A_h \mathbf{u}_h^{n+1} &= \mathbf{F}_h \\ \mathbf{div}\mathbf{u}_h^{n+1} &= \mathbf{G}_h \end{aligned}$$

The solution method for the above system shall be the preconditioned MINRES with AMG preconditioner developed by Chan, Xu and Zikatanov [18] for the elliptic operator $\frac{\mathrm{Re}}{k}I + \eta_s A_h$ and also for the preconditioning the Laplace equation with Neumann boundary condition, we also use same AMG code since it preserves the constant, [18]. The preconditioned MINRES algorithm has been discussed in the chapter 4. See the Algorithm 4.3.1 for more details.

We would like to mention that such an approach makes quite efficient iterative solvers. However, we shall not report detailed numerical experiments here.

## 5.4 Numerical Experiments

In this section, we shall report our main numerical results obtained by simulating the falling sphere through the Johnson-Segalman model. Transient motion of viscoelastic fluids has been studied by many researchers. Among others, we would like to refer two previous works on the sphere falling numerical experiments. First of all, the work of Bodart and Crochet (1994), [10] is similar to our work in that their studies focus on how various parameters affect the pattern of an oscillation of a falling sphere. Their numerical studies are performed using the Oldroyd-B model with two different solvent viscosities and the following observations are made, namely as the solvent viscosity is decreased, the amplitude of the overshoot will increase and the rate of damping will decrease until an under-damped oscillatory response. Their calculations also show graphically that the magnitude of the initial velocity overshoot and the subsequent damped oscillations vary significantly with the constraining effects of the container walls. Secondly, we refer to the study by Harlen (2002),[38] on the formation of negative wake. By simulating various

FENE (Finitely Extensible) models, he proposed a criteria for the formation of negative wake as the extensibility parameter $L$. Namely, for the formation of negative wake, $L$ should be small enough.

Although we do not vary the aspect ratio between the cylinder radius and the sphere radius since we are not concerned with the wall effect, we choose two different viscosities as is the case of the work by Bodart and Crochet and obtained an identical result with them, namely, as is clearly seen from Figure 5.4.1 and Figure 5.4.5 for example, the Oldroyd-B models with two different viscosities $\eta_s = 0.15$ and $\eta_s = 0.05$. Namely, as the viscosity $\eta_s$ decreases from 0.15 to 0.05, the magnitude of overshoot increases.

For the case of negative wake, it seems to be difficult to relate the result by Harlen [38] and our numerical result since we do not consider the FENE (Finitely Extensible) models here. However, we could still relate his work to ours by the simple observation that the extensibility parameter $L$ and the slip parameter "$a$" are closely related physically. The more detailed arguments shall be provided in §5.4.2 below.

### 5.4.1 The slip parameter "a" versus the oscillation of a falling sphere

In this subsection, we shall report our numerical results obtained by simulating the Johnson-Segalman model using both sets of parameters given in the Table 5.2.1 and 5.2.2 with varying "$a$" and $\eta_s$.

We first try the parameter set given in Table 5.2.2. The Figures 5.4.1, 5.4.2, 5.4.3 and 5.4.4 plot the velocity profiles versus time obtained by simulating the Johnson-Segalman model by decreasing the slip parameter "$a$" from 1 down to 0.08 with fixed data sets given in Table 5.2.2. It is clear how the slip parameter "$a$" affects the oscillating

behavior of a falling sphere. Very similar pattern has been observed with different choice of the solvent viscosity except for the magnitude of the overshoot of the sphere velocity.

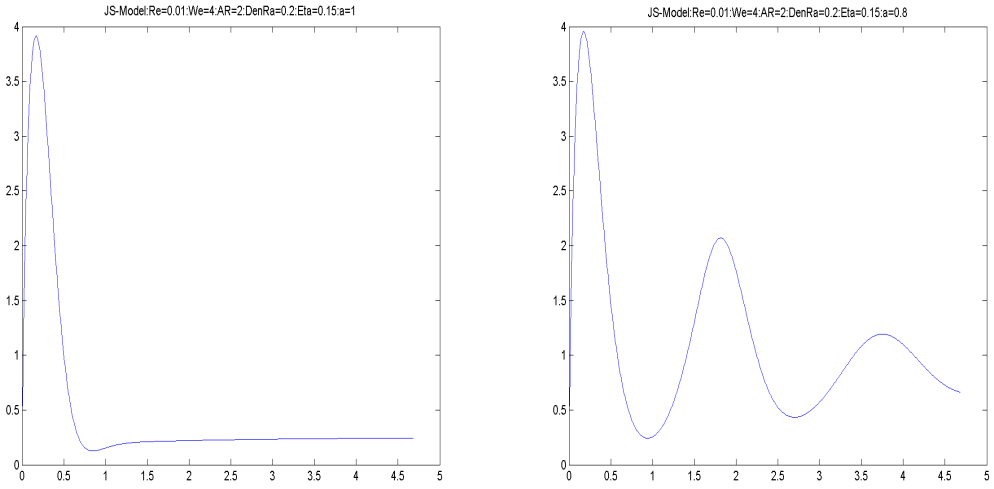


Fig. 5.4.1. J-S model with data sets in Table 5.2.2 : $a = 1$ (left) and $a = 0.8$ (right)

We also investigated how the parameters showing the non-monotone shear stress and rate of strain exhibit different oscillating behavior of a falling sphere. The following Figures 5.4.5, 5.4.6 and 5.4.7 show the velocity profiles versus time obtained by the Johnson-Segalman model with set of parameters given in Table 5.2.1 and varying "$a$" from 1 down to 0.1.

Our numerical experiments for both sets of parameters with varying "$a$" and two choices of the solvent viscosity, $\eta_s$ did not show the sustaining oscillation of the sphere. Considering all the data attained, we may then conclude that the Johnson-Segalman

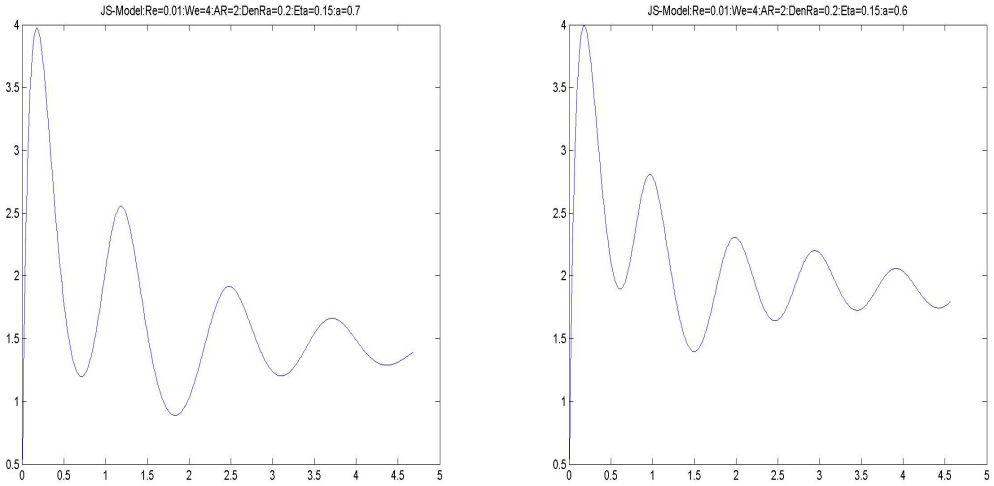


Fig. 5.4.2. J-S model with data sets in Table 5.2.2 : $a = 0.7$ (left) and $a = 0.6$ (right)

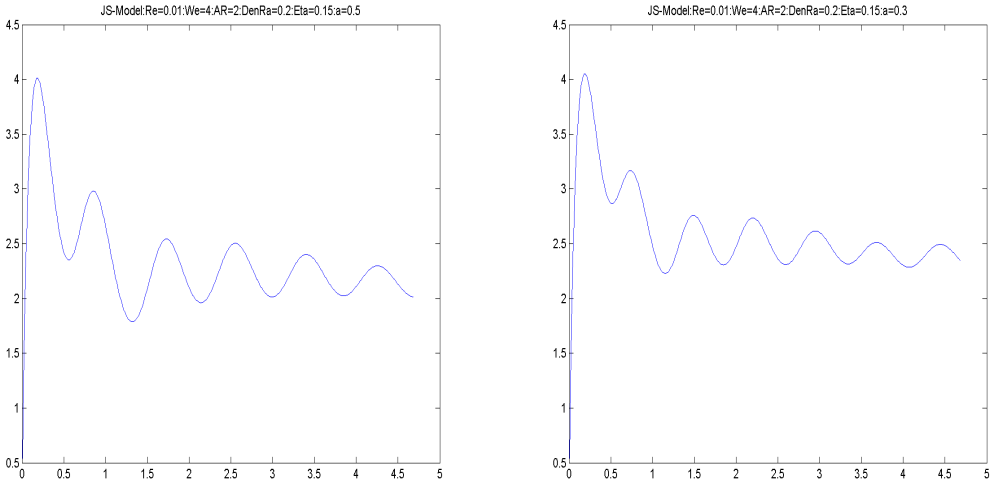


Fig. 5.4.3. J-S model with data sets in Table 5.2.2 : $a = 0.5$ (left) and $a = 0.3$ (right)

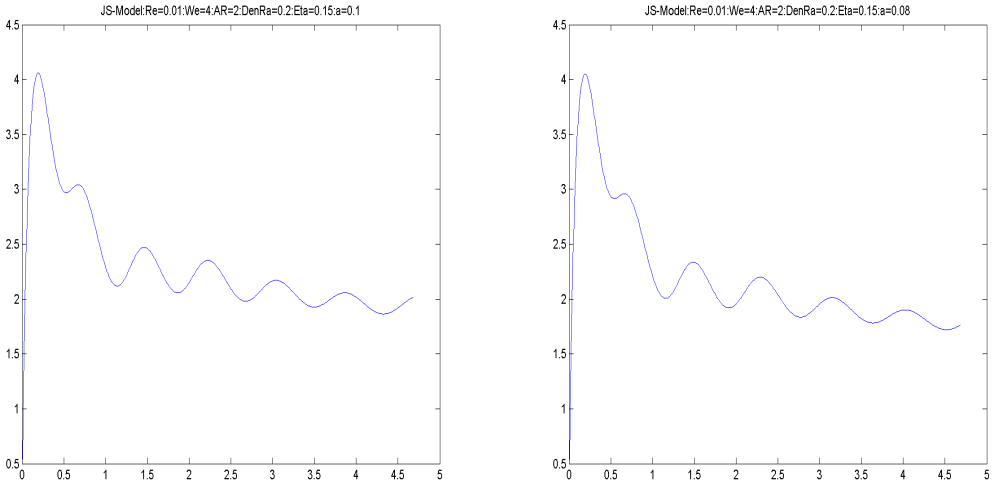


Fig. 5.4.4. J-S model with data sets in Table 5.2.2 : $a = 0.1$ (left) and $a = 0.08$ (right)

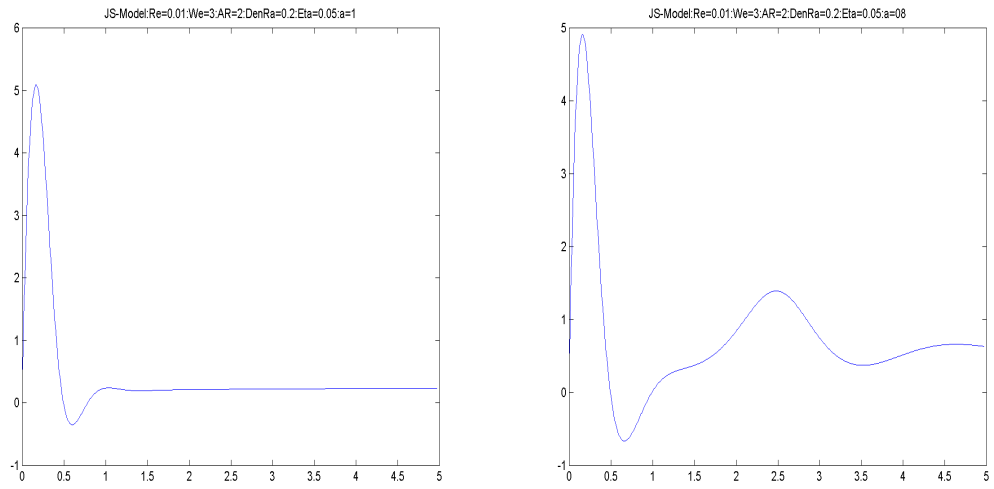


Fig. 5.4.5. J-S model with data sets in Table 5.2.1 : $a = 1$ (left) and $a = 0.8$ (right)

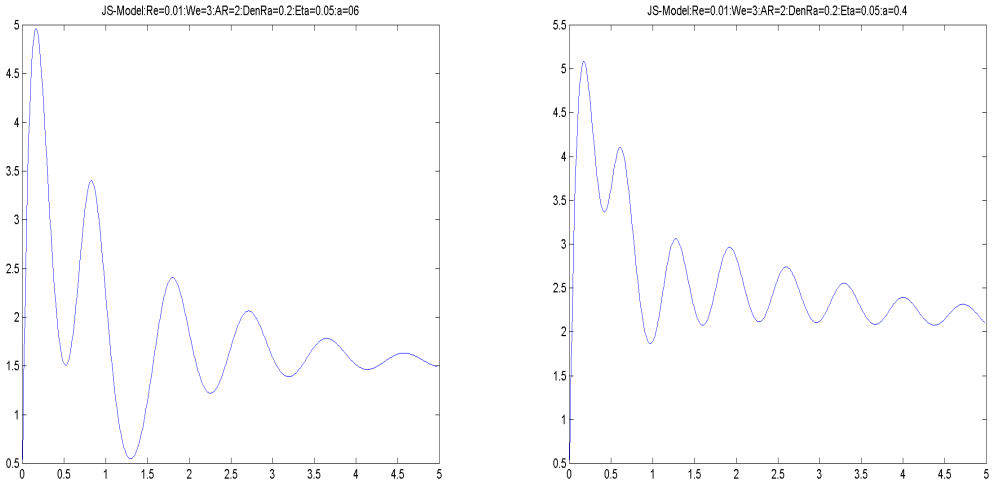


Fig. 5.4.6. J-S model with data sets in Table 5.2.1 : $a = 0.6$ (left) and $a = 0.4$ (right)

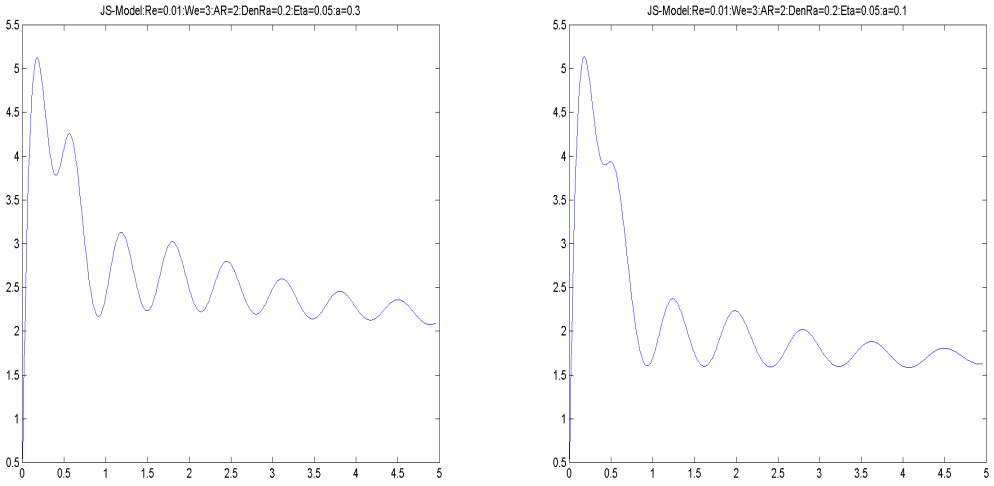


Fig. 5.4.7. J-S model with data sets in Table 5.2.1 : $a = 0.3$ (left) and $a = 0.1$ (right)

models do not show the continuous oscillations. This type of results seems to be further supported by the previous works performed for the simple shear flow by Georgiou and Vlassopoulos in 1998, [29] in which some study on the stability of the time dependent Johnson-Segalman model for the shear flow has been discussed and authors concluded that a non-monotonic relation can not be alone responsible for stick-slip instability.

In the work of Chen and Rothstein in 2004, [21], some similar experimental observations on the continual oscillation have been made and FENE-PM model (see the Appendix C) is suggested for such an experiment. As discussed in the chapter 3, in an energy point of view, it is intriguing that the FENE-PM model has some special feature. Namely, it does not satisfy the usual energy estimates that holds for most models like PTT and Oldroyd-B including the Johnson-Segalman model. We are in a process of attempting to simulate such a model and will leave further discussions for the future works.

### 5.4.2 A formation of the negative wake

In the absence of inertia, the velocity in a Newtonian fluid is fore-aft symmetric. More precisely, along the cylinder axis, the velocity decays monotonically to zero over a distance proportional to the diameter of the cylinder. Hence, in the laboratory frame, the fluid velocity along the axis is always downwards. One particularly intriguing aspect of some viscoelastic fluid, especially shear thinning viscoelastic fluids is that it reveals some remarkable phenomena, namely, the fluid behind the sphere moves upwards, in the opposite direction to the falling sphere. Such an overshoot in the axial velocity is then termed as "Negative Wake" in [39]. It was first reported by Sigli and Coutanceau for

sphere falling through a polyethylene oxide solution in 1977, [79] and also by Hassager in 1979 for a bubble, [39].

Arigo and McKinley, [2] found that the negative wake always appears for a PAA fluid in experiments for the sphere problem. In their experiments, the PAA fluid has a shear-thinning viscosity, shear tension-thickening extensional viscosity. They also confirmed that the value of the elongational stress relative to shear stress in the wake is a key factor for the formation of negative wake. This is then reconfirmed by numerical experiments performed by Harlen in 2002, [38] based on the FENE models including FENE-P and FENE-CR. More quantitative criterion for the formation of negative Wake is presented in (5.4.1).

It is still an open question what physical mechanism produces the opposing force responsible for producing the negative wake [38] while the aforementioned qualitative reasoning is available although there are several proposed criteria for the formation of the negative wake, [38, 25]. The criterion for the formation of negative Wake proposed by Harlen, [38, 63] can be given as follows :

$$\frac{r}{R} \lesssim \frac{\tau_{rz}}{\tau_{zz}}, \tag{5.4.1}$$

where $\tau_{rz}$ is the radial gradient of $\tau$, or a shear stress and $\tau_{zz}$ is the axial gradient of $\tau$ or an extensional stress. Moreover, it is shown that $\tau_{rz} \approx O(L)$ while $\tau_{zz} \approx O(L^2)$. Namely, from the criterion given in (5.4.1), we conclude that the negative wake appears for sufficiently small values of $L$, which has been also shown numerically by Strape and Crochet (1994) from simulating the sphere problem using the FENE-CR model.

Some close connections between the main observations by Harlen and our numerical results are in order. First of all, in simulating the transient motion of sphere in the Oldroyd-B model $a = 1$, it is observed that a negative wake develops at early times in the transient evolution, and as the sphere approaches steady state, the temporary negative wake disappears. Although, it is not explicitly stated in the work by Harlen, it can be deduced that as $L$ becomes smaller, the negative wake is stronger (see the criteria (5.4.1 above). It is our observation that as $a$ decreases, in contrast to the case when $a = 1$, the negative wake still appears in the later stages reaching to the steady state and the negative wake becomes stronger. See the Figure 5.4.8 below for the negative wake obtained by simulations with parameters given in Table 5.2.2 and with $a = 0.6$. It is a snap shot of the profile of $\mathbf{u}_z$ component of the velocity field viewed from the sphere frame.

The figures show that the negative wake is stronger a little bit away from the centerline, which can be interpreted from the criteria, 5.4.1 given by Harlen. Namely, we observe that for the formation of the negative Wake, the shear stress plays an important role and the shear stress due to the sphere formed a little bit away from the axis of symmetry.

Although the Oldroyd-B simulations support our numerical results are consistent with previous works in the literatures, the second connection on the relation between the formation of negative wake and the slip parameter "$a$" cast us a question why.

In the following discussion, we shall show how the slip parameter "$a$" is related to the extensibility parameter $L$ and argue that our numerical observation is natural from the numerical works by Harlen.


Fig. 5.4.8. Negative Wake : $\mathbf{u}_z$ plot with $a = 0.6$

To relate the extensibility parameter $L$ and the slip parameter "$a$", we need to look at $L$ as a measure of the molecular rigidity. Namely, we can make the simple observation that as $L$ becomes smaller, the bead spring becomes more rigid. Hence the surrounding continuum can not stretch the polymer chains as it can for less rigid molecules because there is always some restriction of the movement of the polymer chains with the extensibility parameter, $L$ and this situation worsen as $L$ decreases. Now let us look closely at the slip parameter "$a$". The physical explanation for the slip parameter "$a$" can be found in Larson [54]. That is, the deformation of the continuum "slips" past the strands, the continuum does not feel all the tension in the strands. Thus each strand transmits only a fraction of its tension to the continuum. Imagine that if one attempts to stretch an elastic cord with oily hands, for example, the slippage of one's hands would reduce the felt force. In the view point of the continuum, it can not work towards the

strand as much as it can for less slippery strands just because the strand is slippery. Namely, the more slippery the strands are, the more rigid they are from the continuum point of view. In short, $L^{-1}$ can be thought of as proportional to the slippage. Note that more precise slippage parameter should be written as $\xi = 1 - a$ and as $\xi$ increases, the strand is more slippery. However, we shall be needing to provide more quantitative arguments in view of modelling and it is still an ongoing task but supported by the current numerical experiments.

# Appendix A

## On the well-posedness of axisymmetric stokes equations

Many engineering problems are symmetric. This allows us to reformulate full three dimensional Navier-Stokes equation into the less demanding axisymmetric formulation. However, not much has been known on this formulation. The first result on the axisymmetric formulation of Navier-Stokes equation in terms of the Finite element method is perhaps due to Tabata [81], who proved that Taylor-Hood element pairs are stable elements for the axisymmetric formulations. The well-posedness of axisymmetric problem is due to Bernadi et al, [8] in 1999. Here we shall attack axis-symmetric problems in a quite different context.

The Stokes equation reads:

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{in } \Omega \\ \mathbf{divu} &= 0 \quad \text{in } \Omega \end{aligned} \tag{A.0.1}$$

Let us denote $(r, \theta, z)$ the cylindrical coordinate. We shall assume that the solution $\mathbf{u} = (u_r, u_\theta, u_z)$ is axisymmetric, from which we may reformulate the Stokes equation

(A.0.1) as follows:

$$\begin{aligned} -\frac{\partial^2 u_r}{\partial r^2} - \frac{1}{r}\frac{\partial u_r}{\partial r} - \frac{\partial^2 u_r}{\partial z^2} + \frac{u_r}{r^2} + \frac{\partial p}{\partial r} &= f_r \\ -\frac{\partial^2 u_z}{\partial r^2} - \frac{1}{r}\frac{\partial u_z}{\partial r} - \frac{\partial^2 u_z}{\partial z^2} + \frac{\partial p}{\partial z} &= f_z \\ \frac{u_r}{r} + \frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} &= 0 \end{aligned} \tag{A.0.2}$$

Now we briefly discuss the boundary condition. For simplicity, we shall impose no-slip boundary condition for $\mathbf{u}$ on the boundary denoted by $\partial\Omega$, in axisymmetric formulation, we need to impose symmetry condition on the axis, namely

$$\frac{\partial u_z}{\partial r} = 0 \quad \text{on the z-axis.} \tag{A.0.3}$$

Note that this does not seem to be so-called the natural boundary condition.

As has been seen in (A.0.2), the axisymmetric formulation of Stokes system does not look so beautiful. In this formulation, the direct proof of well-posedness based on inf-sup condition (more precisely Fortin's criterion) has been provided recently by Bernardi et al in [8]. Our proof is different in that we attack the original formulation (A.0.1) with axisymmetric data. The result has been obtained by restricting our concern on the existence of the axisymmetric solution corresponding to axisymmetric data.

Throughout our presentation, we are mainly interested in the space such as $H^{1,0}(\Omega)$ and $\mathbf{H}^{1,0}(\Omega)$. We shall denote corresponding axisymmetric spaces by $H^{1,0}_a(\Omega)$ and $\mathbf{H}^{1,0}_a(\Omega)$.

Let us consider the following weak formulation of the axisymmetric Stokes problem. For any given $\mathbf{f}_a \in \mathbf{L}^2_a(\Omega)$, find $\mathbf{u}_a \in H^{1,0}_a(\Omega)$ and $p_a \in L^2_{a,0}(\Omega)$ such that

$$\begin{aligned}(\nabla \mathbf{u}_a, \nabla \mathbf{v}_a) - (\mathbf{div}\mathbf{u}_a, p_a) &= (\mathbf{f}_a, \mathbf{v}_a), \quad \forall \mathbf{u} \in \mathbf{H}^{1,0}_a(\Omega) && \text{(A.0.4)}\\ (\mathbf{div}\mathbf{u}_a, q_a) &= 0, \quad \forall q_a \in L^2_{a,0}(\Omega) && \text{(A.0.5)}\\ & && \text{(A.0.6)}\end{aligned}$$

Our main objective here is to show that the well-posedness of (A.0.4) and (A.0.5). We first note that it is equivalent to show the following well-known Babuska-Brezzi condition, namely there exists a positive constant $c > 0$ such that

$$\inf_{q_a \in L^2_a(\Omega)} \sup_{\mathbf{u}_a \in \mathbf{H}^{1,0}_a(\Omega)} \frac{(\mathbf{div}\mathbf{u}_a, q_a)}{\|\mathbf{u}_a\|_1} \geq c\|q_a\|_0. \tag{A.0.7}$$

This question shall be answered affirmatively by the following theorem :

**Theorem A.0.1.** *Given $f_a \in L^2_a(\Omega)$, there exists $\mathbf{u}_a$ such that*

$$\mathbf{div}\mathbf{u}_a = f_a. \tag{A.0.8}$$

In order to prove the existence of such vector field, we shall recall the following well-known result.

**Lemma A.0.1.** *For any $q \in L^2(\Omega)$ with $\int_\Omega q dx = 0$, there exists a function $\mathbf{u} \in \mathbf{H}^{1,0}(\Omega)$ such that*

$$\mathbf{div}\mathbf{u} = q \quad \text{with} \quad \|\mathbf{u}\|_1 \leq c\|q\|_0, \tag{A.0.9}$$

*where $c = c(\Omega)$ is a positive constant.*

We are in a position to prove the theorem.

*Proof.* According to the above lemma (A.0.1), for given $q$, we can find a vector field $\mathbf{u}$, which is not necessarily axisymmetric with $\mathbf{u}|_{\partial\Omega} = 0$ and $\mathbf{divu} = q$. Now, for any $\theta \in [0, 2\pi)$, we consider the rotation matrix denoted by $\mathcal{R}_\theta$ and set $\mathbf{u}_\theta = \mathcal{R}_\theta^T \cdot \mathbf{u}(\bar{x})$, where

$$\bar{x} = \mathcal{R}_\theta x. \tag{A.0.10}$$

It is clear to see that since $q$ is axisymmetric,

$$\mathbf{divu}_\theta = q \quad , \forall \theta \in [0, 2\pi). \tag{A.0.11}$$

Now we shall consider the following vector field $\bar{\mathbf{u}}$ defined by

$$\bar{\mathbf{u}} = \frac{1}{2\pi} \int_0^{2\pi} \mathbf{u}_\eta d\eta. \tag{A.0.12}$$

The vector field $\bar{\mathbf{u}}$ shall be axisymmetric and each $\mathbf{u}_\eta$ has divergence $q$, hence $\mathbf{div}\bar{\mathbf{u}} = q$. This completes the proof. □

This then is combined with the LBB condition (A.0.7) and proves that the axisymmetric Stokes system is well-posed.

**Remark A.0.1.** *We shall plan to extend this result to the discrete inf-sup conditions and prove the result in [81].*

# Appendix B

# On the convergence analysis of the MSSC for the symmetric positive definite problems

The purpose of the chapter is two-fold. First of all, we shall provide a proof of the Theorem 4.4.2 given in the chapter 4. The proof written here can be found in the recent work by Xu and Zikatanov [91] and given here for completeness. Second of all, we shall attempt to illustrate that our new framework on the convergence analysis of the MSSC applied to singular problems can also be used for the analysis for symmetric positive case. For this purpose, we shall provide an analysis for the two grid method applied to the symmetric positive definite problem.

## B.1 Proof of the Theorem 4.4.2

The main theorem used in the chapter 4 is as follows :

**Theorem B.1.1.** *If the following assumptions are satisfied*

**(B0)** $V = \sum_{i=1}^{J} V_i$

**(B1)** $T_i : V_i \mapsto V_i$ *is isomorphic for each* $i = 1 : J$

**(B2)** $\|T_i v\|^2 \leq \omega(T_i v, v) \quad \forall v \in V$ *with* $\omega \in (0, 2)$, *then*

$$\|E\|^2_{\mathcal{L}(V,V)} = \|(I - T_J) \cdots (I - T_1)\|^2 = \frac{c_0}{1 + c_0}, \tag{B.1.1}$$

*where*

$$c_0 = \sup_{\|v\|=1} \inf_{\sum_{i=1}^J v_i = v} \sum_{i=1}^J (\bar{T}_i^{-1} T_i^* w_i, T_i^* w_i) \text{ with } w_i = \sum_{j=i}^J v_j - T_i^{-1} v_i.$$

The definitions and the notation are the same as for the subspace correction method described earlier in this paper, but with a nonsingular $a(\cdot,\cdot)$. The proof of this theorem is based on the following lemma:

**Lemma B.1.1.** *Assume that $T_i$ satisfies* **(B1)** *and* **(B2)**. *Then*

1. *$I - T_i$ is nonexpansive.*

2. *$T_i, T_i^*$ and $\bar{T}_i$ have the same kernel: $\mathcal{N}(\bar{T}_i) = \mathcal{N}(T_i) = \mathcal{N}(T_i^*)$.*

3. *$T_i, T_i^*$ and $\bar{T}_i$ have the same range: $\mathcal{R}(\bar{T}_i) = \mathcal{R}(T_i) = \mathcal{R}(T_i^*) = V_i$.*

4. *The following inequality holds:*

$$\frac{2-\omega}{\omega} \|T_i v\|^2 \leq \|v\|^2 - \|(I - T_i)v\|^2 = (\bar{T}_i v, v), \ quadv \ \in \ V$$

5. *As operators restricted on $V_i$, the above (1)–(3) are still valid.*

6. *$T_i, T_i^*$ and $\bar{T}_i$ are all isomorphisms from $V_i$ to itself.*

7. *$\bar{T}_i$ is nonnegative on $V$ and symmetric positive definite on $V_i$.*

The proof of this lemma is rather straightforward, and can be found in [91]. As a consequence of this lemma B.1.1, we obtain that

$$T_i P_i = T_i, \quad T_i^* P_i = T_i^*, \quad \bar{T}_i P_i = \bar{T}_i, \tag{B.1.2}$$

where $P_i$ is the usual elliptic projection (i.e. orthogonal w.r.t. $a(\cdot,\cdot)$ inner product). Note that the first identity can be also be obtained by the definition of $T_i$. But below we use Lemma B.1.1 (3), namely that $\mathcal{R}(T_i) = V_i$ and $\mathcal{R}(T_i^*) = V_i$. The proof of the first relation in B.1.2 is as follows: For any $u \in V$, and $v \in V$ we have

$$a(T_i P_i u, v) = a(P_i u, T_i^* v) = a(u, T_i^* v) = a(T_i u, v). \tag{B.1.3}$$

The flow chart below shows how these identities were obtained:

$$\text{Definition of } T_i^* \rightarrow \text{Definition of } P_i$$
$$\rightarrow \text{Lemma B.1.1 (3)} \rightarrow \text{Definition of } T_i^*$$

Using almost identical argument the second identity in (B.1.2) can is proved by the following

$$a(T_i^* P_i u, v) = a(P_i u, T_i v) = a(u, T_i v) = a(T_i^* u, v). \tag{B.1.4}$$

The third equation in (B.1.2) follows directly from the first two.

*Proof of Theorem 4.4.2.* We first set

$$E_0 = I \text{ and } E_i = (I - T_i)E_{i-1} \text{ for } i = 1 : J. \tag{B.1.5}$$

We then have, with $E_J = E$,

$$\begin{aligned}
\|v\|^2 - \|Ev\|^2 &= \sum_{i=1}^{J} \Big( \|E_{i-1}v\|^2 - \|E_i v\|^2 \Big) \\
&= \sum_{i=1}^{J} \Big( (E_{i-1}v, E_{i-1}v) - ((I - T_i)E_{i-1}v, (I - T_i)E_{i-1}v) \Big) \\
&= \sum_{i=1}^{J} ((I - (I - T_i)^*(I - T_i))E_{i-1}v,\ E_{i-1}v) \\
&= \sum_{i=1}^{J} (\bar{T}_i E_{i-1}v, E_{i-1}v).
\end{aligned}$$

Namely

$$\|v\|^2 - \|Ev\|^2 = \sum_{i=1}^{J} (\bar{T}_i E_{i-1}v, E_{i-1}v). \tag{B.1.6}$$

We now consider the product space $V^J = V \times V \times \ldots \times V$ and $\widetilde{V} = V_1 \times V_2 \times \ldots \times V_J \subset V^J$.

We write the elements in this product space as column vectors:

$$\tilde{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_J \end{pmatrix}, \quad \tilde{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_J \end{pmatrix}, \qquad u_i, v_i \in V \ (i = 1 : J), \quad \tilde{u}, \tilde{v} \in V^J,$$

and use the inner product in the usual way:

$$(\tilde{u}, \tilde{v})_{V^J} = \sum_{i=1}^{J} (u_i, v_i)_V.$$

We introduce the following operators:

$$\underset{\sim}{I} = \begin{pmatrix} I \\ I \\ \vdots \\ I \end{pmatrix}, \quad \underset{\sim}{E} = \begin{pmatrix} I \\ E_1 \\ \vdots \\ E_{J-1} \end{pmatrix}, \quad \underset{\approx}{L} = \begin{pmatrix} I & 0 & 0 & \dots & 0 \\ T_1 & I & 0 & \dots & 0 \\ T_1 & T_2 & I & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_1 & T_2 & T_3 & \dots & I \end{pmatrix}$$

and $\underset{\approx}{\bar{T}} = \mathrm{diag}(\bar{T}_1, \bar{T}_2, \dots, \bar{T}_J)$.

Note that $\underset{\sim}{I}, \underset{\sim}{E} : V \mapsto V^J$, $\underset{\approx}{L} : V^J \mapsto V^J$ and $\underset{\approx}{\bar{T}} : V^J \mapsto \widetilde{V}$. Furthermore $\underset{\approx}{L} : V^J \mapsto V^J$ and $\underset{\approx}{\bar{T}} : \widetilde{V} \mapsto \widetilde{V}$ are apparently isomorphisms.

By the defining relation (B.1.5) for $E_i$, we have the following identities

$$\sum_{j=1}^{i-1} T_j E_{j-1} + E_{i-1} = I, \quad i = 1, 2, \dots, J$$

which can be written in the following compact form:

$$\underset{\approx}{L}\underset{\sim}{E} = \underset{\sim}{I}.$$

In view of (B.1.6), we have

$$\begin{aligned} \|v\|^2 - \|Ev\|^2 &= (\underset{\sim}{\bar{T}}\underset{\sim}{E}v, \underset{\sim}{E}v)_{V^J} \\ &= (\underset{\sim}{\bar{T}}\underset{\sim}{L}^{-1}\underset{\sim}{I}v, \underset{\sim}{L}^{-1}\underset{\sim}{I}v)_{V^J} \\ &= (\underset{\sim}{I}^*(\underset{\sim}{L}^*)^{-1}\underset{\sim}{\bar{T}}\underset{\sim}{L}^{-1}\underset{\sim}{I}v, v). \end{aligned}$$

Namely

$$\|v\|^2 - \|Ev\|^2 = (\underset{\sim}{I}^*(\underset{\sim}{L}^*)^{-1}\underset{\sim}{\bar{T}}\underset{\sim}{L}^{-1}\underset{\sim}{I}v, v). \tag{B.1.7}$$

The derivation so far has been rather straightforward.

We proceed to further modify (B.1.7). Let $\underset{\sim}{T} = \mathrm{diag}(T_1, T_2, \ldots, T_J) : \widetilde{V} \mapsto \widetilde{V}$. We note that

$$[(\underset{\sim}{L}^*\underset{\sim}{T} - \underset{\sim}{\bar{T}})\tilde{v}]_i = T_i^*(\sum_{j=i}^{J} T_j v_j - v_i) \tag{B.1.8}$$

which, thanks to (3) of Lemma B.1.1, implies that $\mathcal{R}(\underset{\sim}{L}^*\underset{\sim}{T} - \underset{\sim}{\bar{T}}) \subset \widetilde{V}$. Since $\underset{\sim}{\bar{T}} : \widetilde{V} \mapsto \widetilde{V}$ is an isomorphism, we can define the following operator (from $\widetilde{V}$ to $\widetilde{V}$):

$$\underset{\sim}{S} = (\underset{\sim}{T}^*\underset{\sim}{L} - \underset{\sim}{\bar{T}})\underset{\sim}{\bar{T}}^{-1}(\underset{\sim}{L}^*\underset{\sim}{T} - \underset{\sim}{\bar{T}}).$$

By (B.1.8), we have

$$(\underset{\sim}{S}\tilde{v}, \tilde{v}) = \sum_{i=1}^{J}(\bar{T}_i^{-1}T_i^*u_i, T_i^*u_i) \text{ with } u_i = \sum_{j=i}^{J} T_j v_j - v_i. \tag{B.1.9}$$

By means of the operator $\underset{\approx}{S}$, we claim that the following relation holds

$$\|v\|^2 - \|Ev\|^2 = \left( \underset{\sim}{T}(\underset{\approx}{S} + \underset{\sim}{T}^* \underset{\sim}{T})^{-1} \underset{\sim}{T}^* v, v \right) \tag{B.1.10}$$

where, with a slight abuse of notation,

$$\underset{\sim}{T} = (T_1, \cdots, T_J) : \widetilde{V} \mapsto V \qquad , \underset{\sim}{T}^* = \begin{pmatrix} T_1^* \\ \vdots \\ T_J^* \end{pmatrix}, \quad V \mapsto \widetilde{V}. \tag{B.1.11}$$

The identity (B.1.10) is one crucial step in our derivation. Its verification can be carried out by direct calculations. First, we have

$$\begin{aligned} \underset{\approx}{S} &= (\underset{\approx}{T}^* \underset{\approx}{L} \bar{\underset{\approx}{T}}^{-1} - \underset{\approx}{I})(\underset{\approx}{L}^* \underset{\approx}{T} - \bar{\underset{\approx}{T}}) \\ &= \underset{\approx}{T}^* \underset{\approx}{L} \bar{\underset{\approx}{T}}^{-1} \underset{\approx}{L}^* \underset{\approx}{T} - \underset{\approx}{T}^* \underset{\approx}{L} - \underset{\approx}{L}^* \underset{\approx}{T} + \bar{\underset{\approx}{T}} \\ &= \underset{\approx}{T}^* \underset{\approx}{L} \bar{\underset{\approx}{T}}^{-1} \underset{\approx}{L}^* \underset{\approx}{T} - \underset{\sim}{T}^* \underset{\sim}{T}, \end{aligned}$$

and we would like to prove that

$$\begin{aligned} \underset{\sim}{I}^* (\underset{\approx}{L}^*)^{-1} \bar{\underset{\approx}{T}} \bar{\underset{\approx}{L}}^{-1} \underset{\sim}{I} &= \underset{\sim}{I}^* \underset{\approx}{T} (S + \underset{\sim}{T}^* \underset{\sim}{T})^{-1} \underset{\approx}{T}^* \underset{\sim}{I} \\ &= \underset{\sim}{T} (S + \underset{\sim}{T}^* \underset{\sim}{T})^{-1} \underset{\sim}{T}^*. \end{aligned}$$

Let $\underset{\sim}{P} : V \mapsto \tilde{V}$, and $\underset{\sim}{P}^* = (P_1, \ldots, P_J) : \tilde{V} \mapsto V$ be defined in a way similar to $\underset{\sim}{T}$ and $\underset{\sim}{T}^*$ (replacing $T_i$ with $P_i$ in the corresponding definitions). Note that

$$\underset{\sim}{T}^{-1}\underset{\sim}{T} = \underset{\sim}{P}, \quad \underset{\sim}{[T^*]}^{-1}\underset{\sim}{T}^* = \underset{\sim}{P}. \tag{B.1.12}$$

Further, a consequence from (B.1.2) is that

$$\underset{\sim}{I}^*(\underset{\sim}{L}^*)^{-1}\underset{\sim}{\bar{T}}\underset{\sim}{L}^{-1}\underset{\sim}{I} = \underset{\sim}{P}^*(\underset{\sim}{L}^*)^{-1}\underset{\sim}{\bar{T}}\underset{\sim}{L}^{-1}\underset{\sim}{P}.$$

Applying the last two relations and (B.1.12) lead to

$$\begin{aligned}
\underset{\sim}{P}^*(\underset{\sim}{L}^*)^{-1}\underset{\sim}{\bar{T}}\underset{\sim}{L}^{-1}\underset{\sim}{P} &= \underset{\sim}{P}^*(\underset{\sim}{L}\underset{\sim}{\bar{T}}^{-1}\underset{\sim}{L}^*)^{-1}\underset{\sim}{P} \\
&= \underset{\sim}{P}^*\underset{\sim}{T}(S + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*\underset{\sim}{P} \\
&= \underset{\sim}{I}^*\underset{\sim}{T}(S + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{I}^*\underset{\sim}{P} \\
&= \underset{\sim}{T}(S + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*,
\end{aligned}$$

which gives (B.1.10). From (B.1.10) then we obtain that

$$\|E\|^2 = 1 - \inf_{\|v\|=1}\left(\underset{\sim}{T}(\underset{\sim}{S} + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*v, v\right). \tag{B.1.13}$$

In regard to the last term in the above identity, we claim that the following relation holds:

$$\inf_{\|v\|=1}\left(\underset{\sim}{T}(\underset{\sim}{S} + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*v, v\right) = \frac{1}{1 + c_0} \tag{B.1.14}$$

where

$$c_0 = \sup_{\|v\|=1} \inf_{\underset{\sim}{T}\tilde{v}=v} (\underset{\approx}{S}\tilde{v}, \tilde{v}). \tag{B.1.15}$$

Let us now prove (B.1.14). To proceed, for any $w \in V$, let

$$\tilde{w} = (\underset{\approx}{S} + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^* w, \quad v = \underset{\sim}{T}\tilde{w}.$$

By a simple calculation, we have the following

$$\frac{(\underset{\sim}{T}(\underset{\approx}{S} + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^* w, w)}{(\underset{\sim}{T}(\underset{\approx}{S} + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^* w, \underset{\sim}{T}(\underset{\approx}{S} + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^* w)} = \frac{(\underset{\approx}{S}\tilde{w}, \tilde{w}) + \|v\|^2}{\|v\|^2}.$$

By writing $\underset{\approx}{S} = \underset{\approx}{S} + \underset{\sim}{T}^*\underset{\sim}{T} - \underset{\sim}{T}^*\underset{\sim}{T}$, it is easy to verify that

$$(\underset{\approx}{S}\tilde{w}, \tilde{\phi}) = 0, \quad \forall \tilde{\phi} \in \mathcal{N}(\underset{\sim}{T})$$

which implies that

$$(\underset{\approx}{S}\tilde{w}, \tilde{w}) = \inf_{\underset{\sim}{T}\tilde{v}=v} (\underset{\approx}{S}\tilde{v}, \tilde{v}).$$

Using the fact that $\underset{\sim}{T}$ is onto and therefore $\underset{\sim}{T}^*$ is one to one, we conclude that $\underset{\sim}{T}(\underset{\approx}{S} + \underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*$ is a symmetric, positive definite isomorphism. Hence

$$\begin{aligned}
&\left(\inf_{\|v\|=1}(\underset{\sim}{T}(\underset{\approx}{S}+\underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*v, v)\right)^{-1} \\
&\qquad\qquad (\underset{\sim}{T}(\underset{\approx}{S}+\underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*w, w) \\
&= \sup_{w\in V}\frac{}{(\underset{\sim}{T}(\underset{\approx}{S}+\underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*w, \underset{\sim}{T}(\underset{\approx}{S}+\underset{\sim}{T}^*\underset{\sim}{T})^{-1}\underset{\sim}{T}^*w)} \\
&= \sup_{w\in V}\inf_{\underset{\sim}{T}\tilde{v}=v}\frac{(\underset{\approx}{S}\tilde{v},\tilde{v})}{\|v\|^2}+1 = \sup_{v\in V}\inf_{\underset{\sim}{T}\tilde{v}=v}\frac{(\underset{\approx}{S}\tilde{v},\tilde{v})}{\|v\|^2}+1 = c_0+1.
\end{aligned}$$

This proves the identity (B.1.14).

At this point, we can conclude that the relation (B.1.1) can be obtained by combining (B.1.13), (B.1.14) and (B.1.9) together with a simple change of variable $T_i v_i \leftrightarrow v_i$. This completes the proof. □

## B.2 Some new framework on the analysis of multigrid method.

In this section, we shall provide another point of view on the multigrid method due to M. Griebel, [34] and try to use the singular framework to prove the convergence of two grid method with exact solver on the coarse grid. We shall restrict our concern on the finite dimensional Hilbert space $V \subset H_0^1(\Omega)$ with $\Omega \subset \mathbb{R}^d$ being a bounded Lipschitz domain.

We shall now consider the following abstract variational problem. Find $u_h \in V$ such that

$$a(u_h, v_h) = \langle f, v_h\rangle \quad \forall v_h \in V, \tag{B.2.1}$$

where $a(\cdot,\cdot)$ is symmetric and positive definite. We shall introduce another subspace $W$ of $H_0^1(\Omega)$ and assume that

$$W \subset V. \tag{B.2.2}$$

Now let us choose basis for both $W$ and $V$ as follows.

$$W = \text{span}\{\psi_1, \cdots, \psi_m\} \quad \text{and} \quad V = \text{span}\{\phi_1, \cdots, \phi_n\}, \tag{B.2.3}$$

where $n = \dim V$ and $m = \dim W$.

We note that any basis function $\psi_j \in W$ can be represented by the basis functions $\{\phi_i\}_{1\le i\le n}$ because of (B.2.2). Namely, for each $1 \le j \le m$, there exists $c_j = (c_j^1, \cdots, c_j^n)^T \in \mathbb{R}^n$ such that

$$\psi_j = \sum_{i=1}^{n} c_j^i \phi_i. \tag{B.2.4}$$

The main idea in interpreting the classical two grid method in augmented framework is in representing the given function $v_h \in V$ as follows:

$$v_h = \sum_{i=1}^{n} \hat{v}^i \phi_i + \sum_{i=1}^{m} \bar{v}^i \psi_i. \tag{B.2.5}$$

With the above representations, the following holds true.

**Lemma B.2.1.** *Let* $\mathcal{V}_H = (c_1, \cdots, c_m) \in \mathbb{R}^{n\times m}$*, then the problem (B.2.1) can be represented by the following matrix equation :*

$$\tilde{\mathcal{A}} = \begin{pmatrix} \mathcal{V}_H^T \mathcal{A} \mathcal{V}_H & \mathcal{V}_H^T \mathcal{A} \\ \mathcal{A}\mathcal{V}_H & \mathcal{A} \end{pmatrix} \tilde{v} = \tilde{f}, \tag{B.2.6}$$

*where* $\mathcal{A} = (\mathcal{A}_{ij}) = (a(\phi_i, \phi_j))$*,* $\tilde{u} = (\tilde{w}, \tilde{v})^T$ *and* $\tilde{f} = (f^T \mathcal{V}_H, f^T)^T$*.*

*Proof.* Given a continuous symmetric and positive definite bilinear form $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$, from the fact that $W \subset V$, we can deduce that

$$\psi_i = \sum_{j=1}^{n} c_i^j \phi_j, \quad \forall 1 \le i \le m. \tag{B.2.7}$$

Since

$$\mathcal{V}_H = (c_1, \cdots, c_m) \quad c_i = (c_i^1, \cdots, c_i^n)^T \in \mathbb{R}^n \quad \forall 1 \le i \le m, \tag{B.2.8}$$

we can deduce that

$$a(\psi_i, \psi_j) = (\mathcal{V}_H^T \mathcal{A} \mathcal{V}_H)_{ij}, \tag{B.2.9}$$

and

$$a(\psi_i, \phi_j) = a(\sum_{k=1}^{n} c_i^k \phi_k, \phi_j) = \sum_{k=1}^{n} c_i^k a(\phi_k, \phi_j) = (\mathcal{V}_H^T \mathcal{A})_{ij}. \tag{B.2.10}$$

This completes the proof. □

We shall list some properties of the augmented matrix $\tilde{\mathcal{A}}$ given in (B.2.6). First of all, it is not surprising to note that $\tilde{\mathcal{A}}$ is a singular matrix with positive diagonal.

Moreover, $\mathcal{R}(\tilde{\mathcal{A}})$ and $\mathcal{N}(\tilde{\mathcal{A}})$ can be characterized as follows:

$$\mathcal{R}(\tilde{\mathcal{A}}) = \left\{ \begin{pmatrix} \mathcal{V}_H^T v_h \\ v_h \end{pmatrix} : v_h \in V \right\} \quad \text{and} \quad \mathcal{N}(\tilde{A}) = \left\{ \begin{pmatrix} c \\ -V_H c \end{pmatrix} : c \in \mathcal{W} \right\} \tag{B.2.11}$$

We shall consider the following space decomposition for the above matrix problem. Let us identify the function space whose representations are in B.2.5 by the space $\tilde{\mathcal{V}}_0$. Namely,

$$\tilde{\mathcal{V}}_0 = \{\tilde{v} = (\bar{v}^T, \hat{v}^T)^T \in \mathbb{R}^{m+n} : v_h = \sum_{i=1}^{n} \hat{v}^i \phi_i + \sum_{i=1}^{m} \bar{v}^i \psi_i, \quad \forall\, v_h \in V\}. \tag{B.2.12}$$

We shall consider the following space decomposition for $\tilde{\mathcal{V}}_0$:

$$\tilde{\mathcal{V}}_0 = \text{span}\{\tilde{e}_1, \cdots, \tilde{e}_m\} + \tilde{\mathcal{V}} = \tilde{\mathcal{W}} + \sum_{i=1}^{n} \text{span}\{\tilde{e}_{m+i}\}, \tag{B.2.13}$$

where $\{\tilde{e}_1, \cdots, \tilde{e}_{m+n}\}$ is the canonical basis for $\mathbb{R}^{m+n}$ and $\tilde{\mathcal{V}}$ and $\tilde{\mathcal{W}}$ are the representations of the spaces $V$ and $W$ respectively.

Now we shall consider the successive subspace correction for the problem (B.2.6), whose error transfer matrix can be written as follows :

$$\tilde{\mathcal{E}} = (I - \tilde{P}_{\tilde{\mathcal{W}}})(I - \tilde{P}_{\tilde{e}_{m+1}}) \cdots (I - \tilde{P}_{\tilde{e}_{m+n}}). \tag{B.2.14}$$

**Remark B.2.1.** *It is easy to show that the MSSC (B.2.14) is equivalent to the multigrid method for the system $\mathcal{A}\tilde{u} = f$ with Gauss-Seidel smoothing on the fine grid $\tilde{\mathcal{V}}$ and exact solver for coarse grid $\tilde{\mathcal{W}}$.*

**Theorem B.2.1.** *The convergence rate for the MSSC (B.2.14) shall be given as followings.*

$$|\tilde{\mathcal{E}}|^2_{\tilde{\mathcal{A}}} = \frac{c_0(\tilde{\mathcal{A}})}{1 + c_0(\tilde{\mathcal{A}})}, \tag{B.2.15}$$

*where*

$$c_0(\tilde{\mathcal{A}}) = \sup_{\tilde{u}\in\mathcal{R}(\tilde{\mathcal{A}})} \inf_{\tilde{c}\in\mathcal{N}(\tilde{\mathcal{A}})} \sum_{i=0}^n \inf_{\tilde{v}_i=\tilde{u}+\tilde{c}} \frac{\sum_{i=0}^n |\tilde{P}_i(\sum_{j=i+1}^n \tilde{u}_j)|^2_{\tilde{\mathcal{A}}}}{(\tilde{u},\tilde{u})_{\tilde{\mathcal{A}}}} \tag{B.2.16}$$

*and moreover, we have*

$$c_0(\tilde{\mathcal{A}}) \lesssim 1. \tag{B.2.17}$$

*Proof.* Recall for any $\tilde{u} \in \mathcal{R}(\tilde{\mathcal{A}})$ and $\tilde{c} \in \mathcal{N}(\tilde{\mathcal{A}})$,

$$\tilde{u} = \begin{pmatrix} \mathcal{V}_H^T\hat{v} \\ \hat{v} \end{pmatrix} \quad \text{and} \quad \tilde{c} = \begin{pmatrix} \bar{c} \\ -\mathcal{V}_H\bar{c} \end{pmatrix} \quad \hat{v}\in\mathbb{R}^n, \quad \bar{c}\in\mathbb{R}^m. \tag{B.2.18}$$

$$\begin{aligned}
c_0(\tilde{\mathcal{A}}) &= \sup_{\tilde{u}\in\mathcal{R}(\tilde{\mathcal{A}})} \inf_{\tilde{c}\in\mathcal{N}(\tilde{\mathcal{A}})} \sum_{i=0}^n \inf_{\tilde{u}_i=\tilde{u}+\tilde{c}} \frac{\sum_{i=1}^n |\tilde{P}_i(\sum_{j=i+1}^n \tilde{u}_j)|^2_{\tilde{\mathcal{A}}} + |\tilde{P}_0(\tilde{u}+\tilde{c})|^2_{\tilde{\mathcal{A}}}}{(\tilde{v},\tilde{v})_{\tilde{\mathcal{A}}}} \\
&= \sup_{\tilde{v}\in\mathcal{R}(\tilde{\mathcal{A}})} \inf_{\tilde{c}\in\mathcal{N}(\tilde{\mathcal{A}})} \sum_{i=0}^n \inf_{\tilde{u}_i=\tilde{u}+\tilde{c}} \frac{\sum_{i=1}^n |\tilde{P}_i(\sum_{j=i+1}^n \tilde{u}_j)|^2_{\tilde{\mathcal{A}}} + |\tilde{P}_0(\tilde{u}+\tilde{c})|^2_{\tilde{\mathcal{A}}}}{\|(\mathcal{I}+\mathcal{V}_H\mathcal{V}_H^T)\hat{v}\|^2_{\mathcal{A}}} \\
&= \sup_{\tilde{u}\in\mathcal{R}(\tilde{\mathcal{A}})} \inf_{\tilde{c}\in\mathcal{N}(\tilde{\mathcal{A}})} \sum_{i=0}^n \inf_{\tilde{u}_i=\tilde{u}+\tilde{c}} \frac{\sum_{i=1}^n |\tilde{P}_i(\sum_{j=i+1}^n \tilde{v}_j)|^2_{\tilde{\mathcal{A}}} + \|P_H(\hat{v}+\mathcal{V}_H\mathcal{V}_H^T\hat{v})\|^2_{\mathcal{A}}}{\|(\mathcal{I}+\mathcal{V}_H\mathcal{V}_H^T)\hat{v}\|^2_{\mathcal{A}}} \\
&= \sup_{\hat{v}\in\mathbb{R}^n} \inf_{\bar{c}\in\mathbb{R}^m} \sum_{i=1}^n \inf_{\hat{v}_i=\hat{v}-\mathcal{V}_H\bar{c}} \frac{\sum_{i=1}^n \|P_i(\sum_{j=i+1}^n \hat{v}_j)\|^2_{\mathcal{A}} + \|P_H(\hat{v}+\mathcal{V}_H\mathcal{V}_H^T\hat{v})\|^2_{\mathcal{A}}}{\|(\mathcal{I}+\mathcal{V}_H\mathcal{V}_H^T)\hat{v}\|^2_{\mathcal{A}}}
\end{aligned}$$

We note that the second term shall be bounded independently of the mesh size $h$. Now let us look closely at the term noticing that $\mathcal{A}_{ij} = a(\phi_i, \phi_j)$ and putting $\Omega_i = \text{supp}\phi_i$,

$$\begin{aligned}
\sum_{i=1}^{n} \|P_i(\sum_{j=i+1}^{n} \hat{v}_j)\|_{\mathcal{A}}^2 &= \sum_{i=1}^{n} ((e_i^T \mathcal{A} e_i)^{-1} e_i^T \mathcal{A}(\sum_{j=i+1}^{n} \hat{v}_j), e_i^T \mathcal{A}(\sum_{j=i+1}^{n} \hat{v}_j)) \\
&= \sum_{i=1}^{n} a(\phi_i, \phi_i)^{-1} a(\phi_i, \sum_{j=i+1}^{n} \hat{v}_j \phi_j)^2 \\
&= \sum_{i=1}^{n} a(\phi_i, \phi_i)^{-1} \left\{ \int_{\Omega_i} \phi_i (\sum_{j=i+1}^{n} \hat{v}_j \phi_j) dx \right\}^2 \\
&\leq \sum_{i=1}^{n} \int_{\Omega_i} (\sum_{j=i+1}^{n} \hat{v}_j \phi_j)^2 dx \\
&\leq \sum_{i=1}^{n} |v - w|_{1,\Omega_i}^2 = |v - w|_1^2, \quad \forall v \in \mathcal{V}, \forall w \in \mathcal{W}
\end{aligned}$$

From this relation, we obtain that with some constant $c$ independent of $h$

$$c_0(\tilde{\mathcal{A}}) \leq \sup_{v \in \mathcal{V}} \inf_{w \in \mathcal{W}} \frac{\|v - w\|_1^2}{\|v\|_1^2} + c \lesssim 1. \tag{B.2.19}$$

This completes the proof. □

# Appendix C

## On the non-dimensionalization of FENE-PM model

In the chapter 2, we introduced the FENE-PM and study the corresponding energy law. This model is also suggested to predict a continual oscillation of the falling sphere in [21]. In this section, for supplementing the chapter 2 and also for the future reference of computational works, we shall briefly review this model for more details together with the non-dimensionalization. This model is first made by Wedgewood et al. in [86]. As usual, the total stress $\sigma$ is decomposed into the sum of two contributions, one from the Newtonian contribution and the other from the polymeric contribution.

$$\sigma = \mu_s \dot{\gamma} + \tau, \tag{C.0.1}$$

where $\dot{\gamma} = \nabla \mathbf{u} + \nabla \mathbf{u}^T$ and $\tau$ is the polymer contribution to stress tensor. The Kramers equation relates the stress tensor and the dyadic product of end-to-end vector, or the components of the second moment. Namely for the FENE-PM chain, the Kramers equation is given by

$$\tau = nHZ \sum_{j=1}^{N-1} \langle Q_j Q_j \rangle - (N-1)nkT\, I, \tag{C.0.2}$$

where $I$ is the unit tensor and

$$Z = 1 + (3/b)\left\{1 + \frac{\mathrm{tr}\tau}{3(N-1)nkT}\right\}. \tag{C.0.3}$$

By combining the second-moment equation for the conformation tensor and Kramers equation, one obtains the constitutive equation for the FENE-PM chain :

$$\begin{aligned}
\tau &= \sum_{j=1}^{N-1} \tau_j \\
Z\tau_j &+ \lambda_j \frac{\delta_F \tau_j}{\delta_F t} - \lambda_j\{\tau_j + nkT\, I\}\frac{D\ln Z}{Dt} = nkT\lambda_j\dot{\gamma} \\
Z &= 1 + (3/b)\left\{1 + \frac{\mathrm{tr}\tau_p}{3(N-1)nkT}\right\},
\end{aligned}$$

where $b = HQ_0^2/kT$ is the FENE parameter, $\lambda_j = \zeta/2Ha_j$, $\zeta$ is the bead friction coefficient and $a_j$ is the eigenvalues of the Rouse matrix.

Now we shall set $\mu_p = \lambda_j nkT$ and obtain that

$$\tau = \sum_{j=1}^{N-1} \tau_j \tag{C.0.4}$$

$$Z\tau_j + \lambda_j \frac{\delta_F \tau_j}{\delta_F t} - \lambda_j\{\tau_j + \frac{\mu_p}{\lambda_j} I\}\frac{D\ln Z}{Dt} = \mu_p\dot{\gamma} \tag{C.0.5}$$

$$Z = 1 + (3/b)\left\{1 + \frac{\lambda_j}{\mu_p}\left(\frac{\mathrm{tr}\tau}{3(N-1)}\right)\right\}. \tag{C.0.6}$$

We are now in a position to nondimensionalize the equations (C.0.4), (C.0.5) and (C.0.4). The same scales that are used to non-dimensionalize the Johnson-Segalman model in the chapter 5 shall also be used for the FENE-PM model. Namely,

- Length scale : Radius of Sphere $r$.
- Velocity scale : Stokes terminal velocity

$$V_N = \frac{2}{9}\frac{r^2(\rho_s - \rho_f)g}{\mu K_N(r/R)} \quad \text{with} \quad \mu = \mu_s + \mu_p.$$

- Time scale : $\dfrac{r}{V_N}$.
- Stress and Pressure scale : $\dfrac{\mu V_N}{r}$.

Let us then set

$$\tau = \frac{\mu V_N}{r}\tilde{\tau}, \tag{C.0.7}$$

$$\eta = \frac{\mu_s}{\mu} \quad \text{and} \quad \mathrm{We}_j = \frac{\lambda_j V_N}{r}. \tag{C.0.8}$$

Let us first look at the non-dimensional representation of $Z(\tau)$.

$$\begin{aligned} Z &= 1 + (3/b)\left\{1 + \frac{\lambda_j}{\mu_p}\frac{1}{3(N-1)}\mathrm{tr}\tilde{\tau}_p\frac{\mu V_N}{r}\right\} \\ &= 1 + (3/b)\left\{1 + \frac{\mathrm{We}_j}{1-\eta}\frac{\mathrm{tr}\tilde{\tau}_p}{3(N-1)}\right\} \end{aligned} \tag{C.0.9}$$

For $\tau_{j,a}$, we have

$$\begin{aligned} \tau_{j,a} &= \tau_j + \frac{\mu_p}{\lambda_j}I = \frac{\mu V_N}{r}\tilde{\tau}_j + \frac{\mu_p}{\lambda_j}I = \frac{\mu V_N}{r}\left(\tilde{\tau}_j + \frac{\mu_p}{\lambda_j \mu V_N/r}I\right) \\ &= \frac{\mu V_N}{r}\left(\tilde{\tau}_j + \frac{\mu_p}{\mu}\frac{1}{\mathrm{We}_j}I\right) = \frac{\mu V_N}{r}\left(\tilde{\tau}_j + \frac{1-\eta}{\mathrm{We}_j}I\right), \end{aligned} \tag{C.0.10}$$

where the unit tensor has not been normalized since $\mu_p$ has the stress unit. Reformulation of the constitutive equation in terms of the conformation tensor

$$\tau_{j,a} = \tau_j + \frac{\mu_p}{\lambda_j} I \tag{C.0.11}$$

leads (C.0.5) to

$$\begin{aligned} \left( Z - \lambda_j \frac{D \ln Z}{Dt} \right) \tau_{j,a} \quad &+ \quad \lambda_j \frac{\delta_F \tau_{j,a}}{\delta_F t} = \frac{\mu_p}{\lambda_j} Z \\ Z \quad &= \quad 1 + (3/b) \left\{ 1 + \frac{\lambda_j}{\mu_p} \left( \frac{\mathrm{tr}\tau}{3(N-1)} \right) \right\}. \end{aligned} \tag{C.0.12}$$

Combined with the non-dimensionalizations (C.0.9) and (C.0.10), the equation (C.0.12) becomes

$$\left( Z - \mathrm{We}_j \frac{D \ln Z}{Dt} \right) \tilde{\tau}_{j,a} + \mathrm{We}_j \frac{\delta \tilde{\tau}_{j,a}}{\delta t} = \frac{1-\eta}{\mathrm{We}_j} Z. \tag{C.0.13}$$

As has been discussed in the chapter 3, this model is somewhat different from the models like Oldroyd-B, PTT or Johnson-Segalman in that the coefficient function, namely

$$Z - \mathrm{We}_j \frac{D \ln Z}{Dt} \tag{C.0.14}$$

for $\tilde{\tau}_{j,a}$ is not necessarily positive. Namely, the usual energy law (see the chapter 3) does not hold for such a model as indicated. Furthermore, in view of an oscillation, a possible oscillation in $Z$, namely, the change in the sign of the function (C.0.14) may hinder the damping as in the other models which result in at most transient oscillations and can cause an oscillation which might not be transient.

In addition to such a exceptional behavior, one might make a simple modification of model ,for example, by replacing the derivative $\frac{\delta_F}{\delta_F t}$ into $\frac{\delta_E}{\delta_E t}$. This then shall result in the interesting results since the oscillation of sphere apparently depends on the slip parameter "$a$". Some extensive numerical works and also studies for such models are in progress in search for the model responsible for the continual and irregular oscillation of a falling sphere.

# References

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank. *Matrix Riccati equations : in control and systems theory*. Systems and Control. Boston : Birkhaser Verlag, 2003.

[2] M.T. Arigo and G.H. McKinley. An experimental investigation of negative wakes behind spheres settling in a shear-thinning viscoelastic fluid. *Rheol. Acta*, 37:307–327, 1998.

[3] F.P.T. Baaijens. An u-ale formulation of 3-d unsteady viscoelastic flow. *Int. J. Numer. Meth. Eng.*, 36:1115–1143, 1993.

[4] A. Belmonte. Self-oscillations of a cusped bubble rising through a micellar solution. *Rhel. Acta*, 39:554–559, 2000.

[5] M. Benzi and D.B. Szyld. Existence and uniqueness of splittings for stationary iterative methods with applications to alternating methods. *Numer. Math.*, 76:309–321, 1997.

[6] A.N. Beris and B.J. Edwards. *Thermodynamics of Flowing Systems : with Internal Microstructure*, volume 36 of *Oxford engineering science series*. New York : Oxford Science Publications, 1994.

[7] A.B. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. SIAM classics in Applied Mathematics, Philadelphia. SIAM, 1994.

[8] C. Bernardi, M. Dauge, and Y. Maday. *Spectral Methods for Axisymmetric Domains*. Series in Applied Mathematics. Elsevier Science Ltd, 1999.

[9] R.B. Bird, R.C. Armstrong, and O. Hassager. *Dynamics of Polymeric Liquids*, volume 1-2. A Wiley-Interscience Publication, 1987.

[10] C. Bodart and M.J. Crochet. The time-dependent flow of a viscoelastic fluid around a sphere. *J. Non. Newt. Fluid. Mech.*, 54:303–329, 1994.

[11] R.V. Boger and K. Walters. *Rheological Phenomenoa in Focus*. Elsevier, Amsterdam, 1993.

[12] J. H. Bramble and X. Zhang. *The analysis of multigrid methods*. Handbook of numerical analysis, Vol. VII. North-Holland, Amsterdam, 2000.

[13] J.H. Bramble and J.E. Pasciak. Iterative techniques for time dependent stokes problems. *Comput. Math. with Appl.*, 33:13–30, 1997.

[14] S. Brenner and L.R. Scott. *The mathematical theory of finite element methods*. Number 15 in Texts in applied mathematics. Springer-Verlag, New York, 2nd ed edition, 2002.

[15] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from lagrange multipliers. *RAIRO Anal. Numer.*, R2:129–151, 1974.

[16] F. Brezzi and M. Fortin. *Mixed and Hybrid finite elements methods*, volume 15 of *Springer Series in Computational Mathematics*. New York : Springer-Berlag, 1991.

[17] Z. H. Cao. A note on properties of splittings of singular symmetric positive semidefinite matrices. *Numer. Math.*, 88:603–606, 2001.

[18] T.F. Chan, J. Xu, and L. Zikatanov. An agglomeration multigrid method for unstructured grids. *Proceedings of the 10th Int. Conf. on Domain Decomposition Methods, AMS*, 1998.

[19] Q. Chang and W. Sun. Multigrid methods for singular systems. *Preprint*, 2002.

[20] J. Chemin and N. Masmoudi. About lifespan of regular solutions of equations related to viscoelastic fluids. *SIAM. J. Math. Anal*, 33:84–112, 2001.

[21] S. Chen and J.P. Rothstein. Flow of a wormlike micelle solution past a falling sphere. *J. Non. Newt. Fluid. Mech.*, 116:205–234, 2004.

[22] J.R. Clermont and J.M. Pierrard. Experimental study of a non-viscometric flow: kinematics of a viscoelastic fluid at the exit of a cylindrical tube. *J. Non-Newt. Fluid. Mech*, 1:175–182, 1976.

[23] J.M. Dealy and T.K.P. Vu. The weissenberg effect in molten polymers. *J. Non-Newt. Fluid. Mech*, 3:127–140, 1977.

[24] L. Dieci and T. Eirola. Positive definiteness in the numerical solution of riccati differential equations. *Numer. Math.*, 67:303–313, 1994.

[25] H.-S. Dou and N.Phan-Thien. The flow of an oldroyd-b fluid past a cylinder in a channel: adaptive viscosity vorticity formulation. *J. Non-Newt. Fluid Mech.*, 87:47–73, 1999.

[26] F. Dupret, J.M. Marchal, and M.J. Crochet. On the consequence of discretization errors in the numerical calculation of viscoelastic flow. *J. Non-Newt. Fluid. Mech.*, 18:173–186, 1985.

[27] P. Espanol, X.F. Yuan, and R.C. Ball. Shear banding flow in the johnson-segalman fluid. *J. Non-Newt. Fluid Mech.*, 65:93–109, 1996.

[28] K. Feng and Z. Shang. Volume-preserving algorithms for source-free dynamical systems. *Numer. Math.*, 71:451–463, 1995.

[29] G.C. Georgiou and D. Vlassopoulos. On the stability of the simple shear flow of a johnson-segalman fluid. *J. Non-Newt. Fluid Mech.*, 75:77–97, 1998.

[30] V. Girault and P.A. Raviart. *Finite Element Methods for Navier-Stokes equations Theory and algorithms.* Springer-Verlag, Berlin, 1986.

[31] J.D. Goddard. Material instability in complex fluids. *Annu. Rev. Fluid Mech.*, 35:113–133, 2003.

[32] R.J. Gordon and W.R. Schowalter. Anisotropic fluid theory : A different approach to the dumbbell theory of dilute polymer solutions. *Trans. of the Soc. Rheo.*, 16:79–97, 1972.

[33] C. Grand, J. Arrault, and M.E. Cates. Slow transient and metastability in wormlike micelle rheology. *J. Phys. II France*, 7:1071–1086, 1997.

[34] M. Griebel. Multilevel algorithms considered as iterative methods on semidefinite systems. *SIAM J. Sci. Comput.*, 15(3):547–565, 1994.

[35] W. Hackbusch. *Multigrid Methods and Applications*. Springer-Verlag, Berlin, 1985.

[36] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*. Number 95 in Applied Mathematical Sciences. New York : Springer-Verlag, 1994.

[37] J. Happel and H. Brenner. *Low Reynolds Number Hydrodynamics*. Martinus Nijhoff, Dordrecht, 1973.

[38] O.G. Harlen. The negative wake behind a sphere sedimenting through a viscoelastic fluid. *J. Non. Newt. Fluid. Mech.*, 108:411–430, 2002.

[39] O. Hassager. Negative wake behind bubbles in non-newtonian liquids. *Nature (London)*, 279:402–403, 1979.

[40] E. Hille and R.S. Phillips. *Functional Analysis and Semi-Groups*, volume 31 of *Americal Mathematical Society. Col. Publ.* New York : Amer. Math. Soc., 1948.

[41] T. Hughes. *Numerical Implementation of Constitutive Models: Rate-Independent Deviatoric Plasticity : Theoretical Foundation for Large-Scale Computations for Nonlinear Material Behavior*. Martinus Nijhoff Publisher, Dordrecht, The Netherlands, 1984.

[42] T. Hughes and J. Winget. Finite rotation effects in numerical integration of rate constitutive equations arising in large-deformation analysis. *Inter. J. Numer. Meth. Eng.*, 15(12):1862–1867, 1980.

[43] M.A. Hulsen. Some properties and analytical expressions for plane flow of leonov and giesekus models. *J. Non-Newt. Fluid. Mech.*, 30:85–92, 1988.

[44] M.A. Hulsen. A sufficient condition for a positive definite configuration tensor in differential models. *J. Non-Newt. Fluid. Mech.*, 38:93–100, 1990.

[45] A. Jayaraman and A. Belmonte. Oscillations of a solid sphere falling through a wormlike micellar fluid. *Phys. Rev. E*, 67:65301–4, 2003.

[46] D.D. Joseph. *Fluid Dynamics of Viscoelastic Liquids*, volume 84 of *Applied Mathematical Sciences*. New York : Springer-Verlag, 1990.

[47] D.D. Joseph, M. Renardy, and J.C.Saut. Hyperbolicity and change of type in the flow of viscoelastic fluid. *Arch. Rat. Mech. Anal.*, 87(3):213–251, 1985.

[48] D.D. Joseph and J.C. Saut. Change of type and loss of evolution in the flow of viscoelastic fluids. *J. Non-Newt. Fluid. Mech.*, 20:117–141, 1986.

[49] M.W. Johnson Jr and D. Segalman. A model for visco-elastic fluid behaviour which allows non-affine deformation. *J. Non-Newt. Fluid. Mech.*, 2:255–270, 1977.

[50] H. B. Keller. On the solution of singular and semidefinite linear systems by iteration. *J. Soc. Indust. Appl. Math. Ser. B Numer. Anal.*, 2:281–290, 1965.

[51] R. Keunings. A survey of computational rheology. *Plenary Lecture, Proc. 13th Int. Congr. on Rheology, D.M. Binding et al. (Eds), British Society of Rheology, Glasgow*, 1:7–14, 2000.

[52] A. Klawonn. An optimal preconditioner for a class of saddle point problems with penalty term. *SIAM J. Sci. Comput.*, 19(2):540–552, 1998.

[53] P. Lancaster and L. Rodman. Existence and uniqueness theorems for the algebraic riccati equation. *Int. J. Control.*, 32:285–309, 1980.

[54] R.G. Larson. *Constitutive Equations for Polymeric Melts and Solutions.* Butterworth Series in Chemical Engineering. Butterworth Publisher, 1988.

[55] Y.-J. Lee, J. Wu, J. Xu, and L. Zikatanov. A sharp convergence estimate of the method of subspace corrections for singular system of equations. *Submitted to Math. Comp.*, 2003.

[56] F. Lin, C. Liu, and P. Zhang. On hydrodynamics of viscoelastic materials. *Preprint*, 2004.

[57] P.L. Lions and N. Masmoudi. Global solutions for some oldroyd models of non-newtonian flows. *Chin. Ann. of Math.*, 21B:131–146, 2000.

[58] C. Liu and N.J. Walkington. An eulerian description of fluids containing viscohyperelastic particles. *Arch. Rat. Mech. Ana.*, 159:229–252, 2001.

[59] A. Lozinski and R. G. Owens. An energy estimate for the oldroyd b model : Theory and applications. *J. Non-Newt. Fluid. Mech.*, 112:161–176, 2003.

[60] C.-Y. David Lu, P.D. Olmsted, and R.C. Ball. Effects of nonlocal stress on the deformation of shear banding flow. *Phys. Rev. Lett.*, 24(4):642–645, 2000.

[61] I. Marek and D.B. Szyld. Comparison theorems for the convergence factor of iterative methods for singular matrices. *Linear Algebra and its Applications.*, 316:67–87, 2000.

[62] I. Marek and D.B. Szyld. Algebraic schwarz methods for the numerical solution of markov chains. *Preprint*, 2003.

[63] G.H. McKinley. *Steady and Transient Motion of a Sphere in an Elastic Fluid*, volume Transport Processes in Bubbles, Drops and Particles. Taylor and Francis, 2001.

[64] R. H. Nochetto and L. B. Wahlbin. Positivity preserving finite element approximation. *Math. Comp.*, 71(240):1405–1419, 2002.

[65] J.G. Oldroyd. On the formulation of rheological equations of state. *Proc. Roy. Soc.*, A200:523–541, 1950.

[66] R.G. Owens and T.N. Phillips. *Computational Rheology*. London : Imperial College Press, 2002.

[67] T.N. Phillips and A.J. Williams. Viscoelastic flow through a planar contraction using a semi-lagrangian finite volume method. *J. Non-Newt. Fluid. Mech.*, 87:215–246, 1999.

[68] O. Pironneau. On the transport-diffusion algorithm and its applications to the navier-stokes equations. *Numer. Math.*, 38, 1982.

[69] O. Radulescu and P.D. Olmsted. Matched asymptotic solutions for the steady banded flow of the diffusive johnson-segalman model in various geometries. *J. Non-Newt. Fluid Mech.*, 91:143–164, 2000.

[70] W.T. Reid. *Riccati differential equations*, volume 86 of *Mathematics in science and engineering*. New York : Academic Press, 1972.

[71] M. Renardy. *Mathematical Analysis of Viscoelastic Flows*, volume 73 of *CBMS-NSF regional conference series in applied mathematics*. Philadelphia : SIAM, 2000.

[72] Y. Renardy. Spurt and instability in a two-layer johnson-segalman liquid. *Theor. Comput. Fluid Dyn.*, 7:463–475, 1995.

[73] D. Rjagopalan, M.T. Arigo, and G.H. McKinley. The sedimentation of a sphere through an elastic fluid part 1. steady motion. *J. Non-Newtonian Fluid Mech.*, 60:225–257, 1995.

[74] D. Rjagopalan, M.T. Arigo, and G.H. McKinley. The sedimentation of a sphere through an elastic fluid part 2. transient motion. *J. Non-Newtonian Fluid Mech.*, 65:17–46, 1996.

[75] T. Rusten and R. Winther. A preconditioned iterative method for saddle point problems. *SIAM J. Matrix Anal. Appl.*, 13(3):887–904, 1992.

[76] I.E. Schochetman, R.L. Smith, and S. Tsui. On the closure of the sum of closed subspaces. *Int. J. Math. Math. Sci.*, 26(5):257–267, 2001.

[77] L.R. Scott and M. Vogelius. *Conforming finite element methods for incompressible and nearly incompressible continua*, volume Lectures in Appl. Math., 22-2 of *Large-scale computations in fluid mechanics, Part 2(La Jolla, Calif., 1983)*. Amer. Math. Soc., Providence, RI, 1985.

[78] L. Shen and J. Xu. On a schur complement operator arising from navier-stokes equations and its preconditioning. In Charles A. Micchelli Zhongying Chen, Yuesheng Li and Yuesheng Xu, editors, *Advances in Computational Mathematics, Proceedings of the Guangzhou International Symposium*, volume 202. Marcel Dekker, INC., 1998.

[79] D. Sigli and M. Coutanceau. Effect of finite boundaries on the slow laminar isothermal flow of a viscoelastic fluid around a spherical obstacle. *J. Non-Newt. Fluid Mech.*, 2(1):1–21, 1977.

[80] J.C. Simo and T.J.R. Hughes. *Computational inelasticity*, volume 7 of *Interdisciplinary applied mathematics*. New York : Springer, 1998.

[81] M. Tabata. Finite element analysis of axisymmetric flow problems. *ZAMM*, 76 Suppl.:171–174, 1996.

[82] R. Temam. *Navier-Stokes equations : theory and numerical analysis*. AMS Chelsea Pub., 2001.

[83] C. Temperton and A. Staniforth. An efficient two-time-level semi-lagrangian semi-implicit integration scheme. *Q.J.R. Meteorol. Soc.*, 113:1025–1039, 1987.

[84] N.P. Thien and R.I. Tanner. A new constitutive equation derived from network theory. *J. Non-Newt. Fluid. Mech.*, 2:353–365, 1977.

[85] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid With contributions by A. Brandt, P. Oswald and K. Stüben.* Academic Press Inc., San Diego, CA, 2001.

[86] L.E. Wedgewood, D.N. Ostrov, and R.B. Bird. A finitely extensible bead-spring chain model for dilute polymer solutions. *J. Non-Newt. Fluid Mech.*, 40:119–139, 1991.

[87] J.L. White and A. Metzner. Development of constitutive equations for polymer melts and solutions. *J. Appl. Polym. Sci.*, 7:1867–1889, 1963.

[88] D. Xiu and G.E. Karniadakis. A semi-lagrangian high-order method for navier-stokes equations. *J. Comp. Phy.*, 172:658–684, 2001.

[89] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34:581–613, 1992.

[90] J. Xu. Survey to the iterative methods for stokes equations. *preprint*, 2002.

[91] J. Xu and L. Zikatanov. The method of subspace corrections and the method of alternating projections in Hilbert space. *J. Amer. Math. Soc.*, 15(3):573–597, 2002.

[92] X.-F. Yuan. Interfacial dynamics of viscoelastic fluid flows. *Phys. Chem. Chem. Phys.*, 1:2177–2182, 1999.

[93] W. Zulehner. Analysis of iterative methods for saddle point problems : A unified approach. *Math. Comp.*, 71(238):479–505, 2001.

# Vita

Young-Ju Lee was born in Kang-won Province, Korea on June 14, 1971. He is the eldest son of Jaesu Lee, the painstaking businessman. In 1998, he received the B.A. degree in Mathematics, from the Chung-Ang University and received fellowships for three years for excellence in undergraduate G.P.A.. During 1992-1995, he served in the army and got married to Eun-Ju Jang in 1995. He had two children, Sung-Min Lee and Hannah Lee who are nicknamed smile. In 1998, he enrolled in the Ph. D. program in mathematics at the Pennsylvania State University. Since then, he has been employed in the Mathematics Department of the Pennsylvania State University as a teaching assistant. From 2000 to 2002, he was supported as a research assistant of Prof. Jinchao Xu.