Numerical Approximation of Shear-Thinning and Johnson-Segalman Viscoelastic Fluid Flows

Jason Howell
Clemson University, jason.s.howell@gmail.com

Follow this and additional works at: https://tigerprints.clemson.edu/all_dissertations

Recommended Citation
https://tigerprints.clemson.edu/all_dissertations/102

This Dissertation is brought to you for free and open access by the Dissertations at TigerPrints. It has been accepted for inclusion in All Dissertations by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.
NUMERICAL APPROXIMATION OF SHEAR-THINNING AND JOHNSON-SEGALMAN VISCOELASTIC FLUID FLOWS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

by
Jason S. Howell
August 2007

Accepted by:
Dr. Vincent J. Ervin, Committee Chair
Dr. Hyesuk Lee, Co-Chair
Dr. Eleanor W. Jenkins
Dr. Daniel D. Warner
ABSTRACT

In this work computational approaches to the numerical simulation of steady-state viscoelastic fluid flow are investigated. In particular, two aspects of computing viscoelastic flows are of interest: 1) the stable computation of high Weissenberg number Johnson-Segalman fluids and 2) low-order approaches to simulating the flow of fluids obeying a power law constitutive model.

The numerical simulation of viscoelastic fluid flow becomes more difficult as a physical parameter, the Weissenberg number, increases. Specifically, at a Weissenberg number larger than a critical value, the iterative nonlinear solver fails to converge. For the nonlinear Johnson-Segalman constitutive model, defect-correction and continuation methods are examined for computing steady-state viscoelastic flows at high Weissenberg numbers. A two-parameter defect-correction method for viscoelastic fluid flow is presented and analyzed. In the defect step the Weissenberg number is artificially reduced to solve a stable nonlinear problem. The approximation is then improved in the correction step using a linearized correction iteration. Numerical experiments support the theoretical results and demonstrate the effectiveness of the method. Continuation methods with natural and pseudo-arclength parametrizations are also examined. The implementation of these methods is discussed and computations with the methods are performed. Numerical results indicate that, for the discrete approximation of the flow equations, a limiting Weissenberg value exists and represents an elastic limit for stable simulation.

Some shear-thinning viscoelastic fluids (e.g. paint, blood) are modeled using a power law constitutive equation. Through the introduction of an auxiliary variable with relevant physical meaning, the variational formulation for the steady-state flow of these fluids can be written as a two-fold saddle point problem. This approach leads to a larger system than the usual finite element approximations, however, the regularity requirements for test and trial functions are relaxed, which leads to an approximation method that is ideally suited for adaptive computation. We analyze this problem in the Sobolev spaces $L^r$ and

ii
$W^{1,r}$, where $r$ is determined by the nonlinearity of the problem. We show existence and uniqueness of the continuous and discrete variational formulations by proving appropriate inf-sup conditions in the Sobolev spaces. Numerical approximations are computed using lowest-order polynomial and Raviart-Thomas elements.
DEDICATION

For Katrina.
ACKNOWLEDGMENTS

I would first like to thank Dr. Vince Ervin for serving as my advisor and sharing his knowledge of mathematics with me. I cannot begin to describe how my understanding of mathematics has grown under his watch. There were never any questions or topics beyond discussion, and I will constantly strive to maintain the level of understanding, thoroughness, and attention to detail that he exhibits. However, the most important thing I have learned from Dr. Ervin was not in any paper or textbook or class notes. It was simply a quote he had posted outside of his office about ten years ago. While the quote verbatim has been long forgotten, the message is still crystal clear: persistence is the key to attaining your goals. The existence of this work is, in no small part, due to that message.

I would like to thank Dr. Hyesuk Lee, who has also served as my advisor, and was the first to show me the elegant theory of the finite element method. Our insightful conversations in her office and her help in my dissertation work have been very important to me. As I prepare to move on, it pleases and comforts me that I am, in some manner, walking in her footsteps. I hope that I have many more opportunities to work with both Dr. Ervin and Dr. Lee throughout my career.

I would like to thank the other members of my committee, Dr. Dan Warner and Dr. Lea Jenkins, for the time, effort, and patience that they have given me over the last few years. I have learned extensively in the courses they have taught and am grateful for the knowledge and experience they have shared with me in many conversations in their offices.

I would also like to acknowledge the National Science Foundation, in particular the Program in Computational and Applied Mathematics within the Division of Mathematical Sciences. I have been able to dedicate almost all of my efforts over the last three years to this thesis because of the funding opportunity that they have provided. I will be forever grateful to the grant PIs (Dr. Ervin, Dr. Jenkins, and Dr. Lee) for their efforts in securing the funding and allowing me to participate.
I would like to thank Iuliana Stanculescu of the Department of Mathematics at the University of Pittsburgh, for her valuable contributions in collaboration with Dr. Ervin and myself on the dual-mixed approximation method for shear-thinning flows.

I would like to thank Carol Woodward for the guidance and experience she shared with me during my internships at Lawrence Livermore National Laboratory. My work there with Carol, Clint Dawson, and Todd Coffey enriched my education and my life.

I am thankful to have had the opportunity to share an office with John Chrispell over the last three years. Aside from restoring my faith in humanity from time to time, our ongoing discussions of research and a wide variety other topics were invaluable. I could not have asked for a better officemate, and I hope to continue to work with John for the rest of my life.

Many friends and family members have been very important to my life, far beyond the scope of this note. I will not list them here, but they are in my thoughts. I am incredibly thankful that I had the chance to know one particular friend, named Troy. Even though his life was far too short, in what little time I had to spend with him he taught me more about myself than I could ever learn on my own. Words are insufficient – his impact on my life transcends description.

Finally, I wish to thank my wife, Dawn. She has supported all aspects of my doctoral education without hesitation. Her love, compassion, encouragement, and desire to see me succeed at every turn of this winding path were immeasurable. I want nothing more than to continue to travel with her through this incredible journey, side by side.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE PAGE</td>
<td>i</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Notation and Assumptions</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Models of Fluid Flow</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1 Conservation of Mass</td>
<td>5</td>
</tr>
<tr>
<td>1.3.2 Conservation of Momentum</td>
<td>6</td>
</tr>
<tr>
<td>1.3.3 Steady-State Model</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Constitutive Laws</td>
<td>7</td>
</tr>
<tr>
<td>1.4.1 Newtonian Fluids and the Navier-Stokes Equations</td>
<td>8</td>
</tr>
<tr>
<td>1.4.2 Johnson-Segalman Fluids</td>
<td>8</td>
</tr>
<tr>
<td>1.4.3 Shear-Thinning Fluids</td>
<td>10</td>
</tr>
<tr>
<td>2. DEFECT-CORRECTION METHODS FOR JOHNSON-SEGALMAN FLUIDS</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Problem Description</td>
<td>13</td>
</tr>
<tr>
<td>2.2.1 Continuous Problem</td>
<td>13</td>
</tr>
<tr>
<td>2.2.2 Variational Formulation</td>
<td>14</td>
</tr>
<tr>
<td>2.3 Finite Element Approximation</td>
<td>15</td>
</tr>
<tr>
<td>2.4 Defect Correction Method</td>
<td>19</td>
</tr>
<tr>
<td>2.4.1 Algorithm</td>
<td>19</td>
</tr>
<tr>
<td>2.4.2 Analysis of the Defect Step</td>
<td>21</td>
</tr>
<tr>
<td>2.4.3 Analysis of the Correction Step</td>
<td>25</td>
</tr>
<tr>
<td>2.4.4 Implementation of the Method</td>
<td>45</td>
</tr>
<tr>
<td>2.4.5 Newton Corrector Iteration</td>
<td>46</td>
</tr>
<tr>
<td>2.4.6 Choosing the defected parameters</td>
<td>46</td>
</tr>
<tr>
<td>2.5 Numerical Experiments</td>
<td>47</td>
</tr>
<tr>
<td>2.5.1 Example 1</td>
<td>48</td>
</tr>
<tr>
<td>2.5.2 Example 2</td>
<td>51</td>
</tr>
</tbody>
</table>

vii
Table of Contents (Continued)

3. CONTINUATION METHODS FOR JOHNSON-SEGALMAN FLUIDS . . . 55
   3.1 Problem Description ........................................... 57
      3.1.1 Continuous Problem and Variational Formulation ........... 57
      3.1.2 Finite Element Approximation ............................. 58
      3.1.3 Newton Iteration ......................................... 59
   3.2 Continuation in Weissenberg Number ........................... 62
      3.2.1 Simple Continuation ....................................... 62
      3.2.2 Natural Continuation ...................................... 63
   3.3 Turning Points and Pseudo-arclength Continuation ............. 67
      3.3.1 Parametrization with respect to arc length ................. 67
      3.3.2 Pseudo-arclength constraints ................................ 70
      3.3.3 Combined Methods ......................................... 73
   3.4 Numerical experiments ......................................... 73
      3.4.1 Four-to-one Contraction Flow ............................. 74
      3.4.2 High Weissenberg Number Results .......................... 74
      3.4.3 Pseudo-Arclength Constraints ............................. 79

4. DUAL-MIXED APPROXIMATION OF A THREE-FIELD MODEL OF SHEAR-
   THINNING FLUIDS .................................................. 83
   4.1 Introduction .................................................... 83
   4.2 Mathematical Setting .......................................... 84
   4.3 Existence and Uniqueness ....................................... 87
      4.3.1 Inf-sup Condition for $B$ ................................ 90
      4.3.2 Inf-sup Condition for $C$ ................................ 92
   4.4 Finite Element Approximation .................................. 99
      4.4.1 A Priori Estimates ....................................... 100
      4.4.2 Discrete Approximation Spaces ........................... 109
      4.4.3 Higher Order Approximation ............................... 119
   4.5 Numerical Experiments ......................................... 129
      4.5.1 Example 1 .................................................. 130
      4.5.2 Example 2 .................................................. 130

5. SUMMARY AND FUTURE WORK ...................................... 135
   5.1 Defect-Correction Methods for Johnson-Segalman Fluids ........ 135
   5.2 Continuation Methods for Johnson-Segalman Fluids ............. 135
   5.3 Dual-Mixed Approximation of Shear-Thinning Fluids ............ 136

BIBLIOGRAPHY ....................................................... 137
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.1</td>
<td>Errors and convergence rates for $\lambda = 5.0$ and selected values of $\lambda$, $\tilde{\lambda}$, defect correction method (Picard corrector).</td>
<td>49</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Iteration counts for Picard and Newton correctors, $h = 1/8$, $\lambda = 5.0$, selected values of $\lambda$ and $\tilde{\lambda}$.</td>
<td>49</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Mesh characteristics</td>
<td>53</td>
</tr>
<tr>
<td>2.5.4</td>
<td>Solution norms, successive meshes, $\lambda = 0.7$, $a = 1$.</td>
<td>53</td>
</tr>
<tr>
<td>2.5.5</td>
<td>Critical $\lambda$ values for $(\sigma, u, p) = (0, 0, 0)$.</td>
<td>54</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Maximum values of $\lambda$ obtained by the different solution methods, selected values of $a$.</td>
<td>75</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Number of steps and steplength reductions required for $a = 1$ on mesh M1.</td>
<td>76</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Performance statistics for pseudo-arclength constraints to pass $\lambda = 11.0$, $\Delta s_0 = 1.0$.</td>
<td>79</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Performance statistics for pseudo-arclength constraints to pass $\lambda = 11.12838$, $\Delta s_0 = 0.01$.</td>
<td>80</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Performance statistics for pseudo-arclength constraints, $\lambda = 11.12884$, $\Delta s_0 = 0.0001$.</td>
<td>80</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Approximation errors and rates of convergence for Example 1.</td>
<td>133</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Approximation norms and relative differences for Example 2.</td>
<td>134</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.1 Iteration counts for fixed values of $\lambda$, $\lambda = 5.0$, $h = 1/8$</td>
<td>50</td>
</tr>
<tr>
<td>2.5.2 Iteration counts for fixed values of $\tilde{\lambda}$, $\lambda = 5.0$, $h = 1/8$</td>
<td>51</td>
</tr>
<tr>
<td>2.5.3 Geometry of 4:1 contraction domain</td>
<td>52</td>
</tr>
<tr>
<td>3.1.1 Some possibilities for the behavior of solution curves for high Weissenberg number</td>
<td>61</td>
</tr>
<tr>
<td>3.2.1 Representation of the simple continuation process</td>
<td>64</td>
</tr>
<tr>
<td>3.2.2 Illustration of one step of natural continuation in $\lambda$</td>
<td>66</td>
</tr>
<tr>
<td>3.3.1 Illustration of the pseudo-arclength condition $N_1$</td>
<td>71</td>
</tr>
<tr>
<td>3.3.2 Illustration of the pseudo-arclength condition $N_2$</td>
<td>72</td>
</tr>
<tr>
<td>3.4.1 Solution norm computed by pseudo-arclength continuation</td>
<td>76</td>
</tr>
<tr>
<td>3.4.2 Horizontal velocity $u_1$ along the line of symmetry for increasing $\lambda$</td>
<td>77</td>
</tr>
<tr>
<td>3.4.3 Stress component $\sigma_{11}$ along a line near the reentrant corner for increasing $\lambda$</td>
<td>78</td>
</tr>
<tr>
<td>3.4.4 Velocity field and speed contour, $\lambda = 9.3466$, $a = 1$, mesh M2</td>
<td>78</td>
</tr>
<tr>
<td>3.4.5 Turning point computed by $N_2$ for mesh M1, $a = 1$, $\Delta s_0 = 0.0001$</td>
<td>81</td>
</tr>
<tr>
<td>3.4.6 Velocity field of upper solution branch minus lower solution branch, $\lambda = 11.1289745$, mesh M1, $a = 1$</td>
<td>82</td>
</tr>
<tr>
<td>4.5.1 Streamlines for $r = 2.0$, driven cavity</td>
<td>131</td>
</tr>
<tr>
<td>4.5.2 Streamlines for $r = 1.5$, driven cavity</td>
<td>132</td>
</tr>
<tr>
<td>4.5.3 Streamlines for $r = 1.1$, driven cavity</td>
<td>132</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1 Motivation

The study of viscoelastic fluid flow is a topic of interest on many fronts, with relevant applications. These applications are broad in scope, including industrial polymer and film processing and the biological modeling of fluids such as blood and mucus. To accompany theoretical and experimental observations, researchers have developed computational algorithms for simulating viscoelastic fluid flow. However, the numerical approximation of viscoelastic fluid flows is a difficult nonlinear problem. This is due mainly to (i) the number of unknowns required for an accurate computation over the problem domain, and (ii) the hyperbolic, nonlinear character of the constitutive equation for the stress.

The most common approaches to simulating viscoelastic flows utilize a finite element method. Of the many interesting avenues of research in numerical methods applied to these problems, two issues of particular interest are:

- For fluids with a differential-type constitutive law, a phenomenon known as the high Weissenberg number problem limits the parameter space for which current computational techniques can obtain a reliable approximation,
- For shear-thinning fluids, the development of computational approaches that utilize structure and low order elements that are well-suited for adaptive computation.

These topics motivate the research that is performed in this work.

The high Weissenberg number problem manifests itself in the inability for nonlinear solvers to converge for Weissenberg numbers beyond a critical value. As the Weissenberg number increases, boundary layers for the stress develop which add to the difficulty of computing accurate numerical approximations. There has been considerable interest by researchers over the years in developing stable numerical algorithms for high Weissenberg number flows. At present, the high Weissenberg number limit encountered in numerical studies is generally regarded as numerical artifact and not a property of the modeling equations [5, 65].
Many shear-thinning fluids are modeled by equations that are a generalized form of a nonlinear Stokes problem. These fluids, of which paints, lubricators, and biological fluids are examples, are governed by constitutive laws that specify the fluid’s extra stress as a nonlinear function of the deformation tensor. Through the introduction of auxiliary variables and the use of a Lagrange multiplier, the nonlinear generalized Stokes problem leads to a dual-mixed structure. In conjunction with low-order approximation spaces, this structure can be analyzed in the appropriate setting and exploited to produce an approximation method that is well-suited for adaptive computation.

1.2 Notation and Assumptions

In this section some notation and conventions are discussed. The information presented here is consistent throughout this paper, however it will occasionally be necessary to introduce context-specific notation and assumptions.

Let $n$ denote the number of spatial dimensions and $\mathbb{R}$ denote the set of real numbers. The domain is denoted by $\Omega \subset \mathbb{R}^n$ with boundary $\Gamma = \partial \Omega$ and let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ represent a spatial variable. Unless otherwise noted, functions $f = f(\mathbf{x})$ are assumed to be defined on $\mathbb{R}^n$, with vector and tensor quantities expressed in boldface, e.g.,

$$
\mathbf{u} = \begin{bmatrix}
u_1 \\
u_2 \\
u_3
\end{bmatrix}
$$

and

$$
\mathbf{\sigma} = \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} \\
\sigma_{31} & \sigma_{32} & \sigma_{33}
\end{bmatrix}
$$

being examples of a vector and tensor, respectively, in $\mathbb{R}^3$. The identity tensor is denoted by $I$ and the trace of a tensor is the usual definition

$$
tr(\mathbf{\sigma}) = \sum_{i=1}^{n} \sigma_{ii}.
$$

For vectors $\mathbf{u} = (u)_i, \mathbf{v} = (v)_i$ and tensors $\mathbf{\sigma} = (\sigma)_{ij}, \mathbf{\tau} = (\tau)_{ij}$, the following products are defined:

$$
\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i
$$

and

$$
\mathbf{\sigma} : \mathbf{\tau} = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} \tau_{ij}
$$
and the products $\sigma u$, $u \sigma$, and $\sigma \tau$ are defined in the usual matrix-vector and matrix-matrix manner. The vector-tensor inner products are defined as

$$u \cdot \sigma = \begin{bmatrix} u_1 \sigma_{11} + u_2 \sigma_{21} \\ u_1 \sigma_{12} + u_2 \sigma_{22} \end{bmatrix}, \quad \text{and} \quad \sigma \cdot u = \begin{bmatrix} u_1 \sigma_{11} + u_2 \sigma_{12} \\ u_1 \sigma_{21} + u_2 \sigma_{22} \end{bmatrix},$$

in two dimensions, with a corresponding extension to three dimensions. Note that $u \cdot \sigma = \sigma \cdot u$ if $\sigma$ is symmetric. With the definitions above, note that $tr(\sigma) = \sigma : I = I : \sigma$.

Partial derivatives of a function $f$ may be denoted by either $\partial f/\partial x_j$, $f_{x_j}$, or, when $f = (f)_i$ is a vector function, $f_{i,x_j} = f_{i,j}$. The gradient of a function $f$ will be denoted by $\nabla f$, and where $f = (f)_i$ is a vector function, the gradient will represent

$$\nabla f = (\nabla f)_{i,j} = \partial f_i/\partial x_j. \quad (1.2.1)$$

While the gradient of a vector function is sometimes defined as $(\nabla f)_{i,j} = \partial f_j/\partial x_i$, the definition (1.2.1) is consistent with [8, 12, 69] and others. The divergence of a vector $u$ may be written as

$$\nabla \cdot u = div u = \sum_{i=1}^{n} \frac{\partial u_i}{\partial x_i}$$

and the divergence of a tensor $\sigma$ is the vector

$$\nabla \cdot \sigma = div \sigma = \begin{bmatrix} \sigma_{11,x_1} + \sigma_{12,x_2} + \sigma_{13,x_3} \\ \sigma_{21,x_1} + \sigma_{22,x_2} + \sigma_{23,x_3} \\ \sigma_{31,x_1} + \sigma_{32,x_2} + \sigma_{33,x_3} \end{bmatrix}.$$  

The product $u \cdot \nabla \sigma$ in two spatial dimensions is written as the tensor

$$u \cdot \nabla \sigma = \begin{bmatrix} u_1 \sigma_{11,x_1} + u_2 \sigma_{11,x_2} & u_1 \sigma_{12,x_1} + u_2 \sigma_{12,x_2} \\ u_1 \sigma_{21,x_1} + u_2 \sigma_{21,x_2} & u_1 \sigma_{22,x_1} + u_2 \sigma_{22,x_2} \end{bmatrix}$$

and is easily generalized to three dimensions. Given a velocity vector $u$, the rate of strain or deformation tensor is given by

$$D(u) = \frac{1}{2} \left( \nabla u + (\nabla u)^T \right).$$

where the $(\nabla u)^T$ denotes the transpose of the gradient tensor.
For $1 \leq r \leq \infty$, let $L^r(\Omega)$ represent the space of all functions $f$ defined on $\Omega$ that satisfy
\[
\int_{\Omega} |f(x)|^r \, d\Omega \leq \infty.
\]
Let $W^{m,r}(\Omega)$ denote the standard Sobolev space [1] of order $m$ with $W^{0,r}(\Omega) = L^r(\Omega)$, and when $r = 2$ the notation $H^m(\Omega) = W^{m,2}(\Omega)$ may be used. Spaces of vector and tensor functions may appear in bold, such as $\mathbf{H}^m$. Superscripts denoting dimension will be omitted from the notation of spaces when it is clear that the elements are vectors or tensors. The space $\mathbf{H}(\text{div}, \Omega)$ is defined by
\[
\mathbf{H}(\text{div}, \Omega) = \{ f \in L^2(\Omega) : \text{div} f \in L^2(\Omega) \}.
\]
On occasion it will be useful to denote with subscript 0 the corresponding sets of functions that vanish on the boundary, e.g.,
\[
W_0^{m,r}(\Omega) = \{ f \in W^{m,r}(\Omega) : f = 0 \text{ on } \Gamma \}.
\]
Let $r'$ denote the unitary conjugate of $r$, that is, for $1 < r < \infty$, $r'$ satisfies $1/r + 1/r' = 1$. For the cases $r = 1$ and $r = \infty$, $r' = \infty$ and $r' = 1$, respectively. The dual space of $W_0^{m,r}(\Omega)$ is then given by the space $W^{-m,r'}(\Omega)$.

Standard notation of $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$ will be used for inner products with specific descriptions given in the context that they appear. For $\mathcal{D} \subset \Omega$, the notation $\| \cdot \|_{m,r,\mathcal{D}}$ will be used to denote the norm functions in $W^{m,r}(\Omega)$. The $\mathcal{D}$ may be omitted from the subscript when $\mathcal{D} = \Omega$, as well as the $r$ when $r = 2$. The seminorm will be denoted by $| \cdot |_{m,r,\Omega}$ with the same rules of notation. For a given Banach space $B$, $\| \cdot \|_B$ may be used to denote its norm.

For the remainder of the paper, $n = 2$ or 3, the variables $p, q$ will usually be used to represent functions describing fluid pressure, $\mathbf{u}, \mathbf{v}, \mathbf{w}$ will be used to represent fluid velocity field vectors, and $\sigma, \tau, \psi, \phi$ will most often be used to represent fluid stress tensors of differing types. The variable $h$ will be used as a spatial mesh parameter.
1.3 Models of Fluid Flow

The equations of fluid flow are derived from physical conservation laws. For this study, the conservation of mass and the conservation of momentum will be used to arrive at a model of steady-state fluid flow. The fluid is assumed to have constant temperature.

An important concept is that of a material derivative or substantive derivative. The material derivative $D/Dt$ of a fluid property $\alpha = \alpha(x,t)$ is given by

$$\frac{D\alpha}{Dt} = \frac{\partial \alpha}{\partial t} + u \cdot \nabla \alpha,$$

where $u$ is the fluid’s velocity field. The $\partial \alpha/\partial t$ term is the rate of change of the property with respect to time, and the $u \cdot \nabla \alpha$ represents the convected portion of the derivative, i.e., how the property changes with the transport induced by the fluid’s velocity field. The fluid property $\alpha$ may represent quantities such as mass, momentum, or a stress component.

Another important tool that is required is the Reynolds Transport Theorem:

**Theorem 1.3.1 (Reynolds Transport Theorem [65])** Let $\kappa(x,t)$ represent a given quantity in an arbitrary fluid volume $V$, and let $u$ be the fluid velocity field. Then

$$\frac{D}{Dt} \int_V \kappa \, dV = \int_V \left( \frac{D}{Dt} \kappa + \kappa \nabla \cdot u \right) \, dV. \quad (1.3.1)$$

1.3.1 Conservation of Mass

Let $\rho = \rho(x,t)$ be the density of a fluid. Then we have that, within an arbitrary volume $V \subset \Omega$, the mass $M$ of the fluid is given by

$$M = \int_V \rho \, dV.$$

The law of conservation of mass stipulates that the mass of fluid in $V$ must remain constant with respect to time, i.e.,

$$\frac{DM}{Dt} = \frac{D}{Dt} \int_V \rho \, dV = 0.$$

Applying the definition of the material derivative and the Reynolds Transport Theorem, we have

$$\int_V \left( \frac{\partial \rho}{\partial t} + u \cdot \nabla \rho + \rho \nabla \cdot \mathbf{u} \right) \, dV = \int_V \left( \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) \right) \, dV = 0. \quad (1.3.2)$$
As the volume $V$ is arbitrary, (1.3.2) reduces to the pointwise equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \quad \text{in } \Omega.$$  

(1.3.3)

Throughout this work, the fluid is assumed to have a constant density in $\Omega$. Thus the condition (1.3.3) reduces to

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega,$$

(1.3.4)

and the fluid is said to be *incompressible*. The incompressibility condition (1.3.4) implies that the flow of fluid into any subset $V$ of $\Omega$ must be the same as the flow out of $V$.

1.3.2 Conservation of Momentum

The law of conservation of momentum stipulates that the rate of change of momentum of a volume of fluid must be equal to the sum of the forces acting on the fluid. First note that the momentum of a volume $V$ of fluid can be written as

$$\int_V \rho \mathbf{u} \, dV.$$

Also, assume $V$ has a closed and bounded surface $\partial V$ with outward unit normal vector $\mathbf{n}$. As done in [65], let $\mathbf{s}_n$ be the stress vector that describes the forces acting on the surface of the fluid volume, and let $\mathbf{b}$ denote the body forces (such as gravity) acting on the volume of fluid itself. The conservation law then implies that

$$D \frac{D}{Dt} \int_V \rho \mathbf{u} \, dV = \int_V \rho \mathbf{b} \, dV + \int_{\partial V} \mathbf{t} \, d(\partial V),$$

or, using Theorem 1.3.1 and (1.3.4), we can write

$$\int_V \rho \frac{D \mathbf{u}}{Dt} \, dV = \int_V \rho \mathbf{b} \, dV + \int_{\partial V} \mathbf{t} \, d(\partial V).$$  

(1.3.5)

Now, as described by Theorem 2.1 in [65], there is a symmetric rank 2 tensor $\mathbf{T}$ such that $\mathbf{s}_n = \mathbf{n} \cdot \mathbf{T} = \mathbf{T} \cdot \mathbf{n}$. Thus, using the divergence theorem and (1.3.5), the conservation of momentum equation reduces to the pointwise equation

$$\rho \frac{D \mathbf{u}}{Dt} = \rho \mathbf{b} + \nabla \cdot \mathbf{T} \quad \text{in } \Omega.$$  

(1.3.6)
Expanding the material derivative and letting $f = \rho \mathbf{b}$, we have

$$\rho \left( \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) = f + \nabla \cdot \mathbf{T}. \quad (1.3.7)$$

The stress tensor $\mathbf{T}$ can be written as the sum of a pressure piece and an extra stress tensor:

$$\mathbf{T} = -p\mathbf{I} + \sigma, \quad (1.3.8)$$

and thus (1.3.7) can be written as

$$\rho \left( \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) + \nabla p - \nabla \cdot \sigma = f. \quad (1.3.9)$$

### 1.3.3 Steady-State Model

In this work, only steady-state fluid flow models are considered. Thus, $\partial \mathbf{u}/\partial t = 0$ and together, the momentum and mass conservation equations can be written together as

$$\rho \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p - \nabla \cdot \sigma = f \quad \text{in} \ \Omega, \quad (1.3.10)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in} \ \Omega. \quad (1.3.11)$$

The equations (1.3.10) and (1.3.11) (together with a boundary condition for the velocity $\mathbf{u}$) contain three unknowns - in order for the system to be properly determined, additional information is necessary. This information is present in the form of a constitutive law.

### 1.4 Constitutive Laws

A constitutive law is a relationship between the fluid velocity (and perhaps its derivatives) and the extra stress tensor $\sigma$. Fluids are generally considered to be either Newtonian or non-Newtonian. Newtonian fluids have a stress tensor that is proportional to the deformation tensor. A non-Newtonian fluid has a much more complicated description of the stress tensor. For some fluids, the stress-deformation relationship may be described as algebraic or differential. Viscoelastic fluids are a type of non-Newtonian fluid in which the stress does not depend linearly on the strain. In many cases, the relationship between $\sigma$ and $\nabla \mathbf{u}$ is nonlinear, implicit, and hyperbolic in nature.
1.4.1 Newtonian Fluids and the Navier-Stokes Equations

For a Newtonian fluid, the constitutive relationship is given by

$$\sigma = 2\nu \mathbf{D}(u),$$  \hspace{1cm} (1.4.1)\]

for some constant $\nu$. Thus the variable $\sigma$ can be eliminated from (1.3.10). With the specification of $u$ on the boundary $\Gamma$, the steady-state Newtonian fluid model is given by

$$\rho u \cdot \nabla u + \nabla p - 2\nu \nabla \cdot \mathbf{D}(u) = f \quad \text{in } \Omega,$$  \hspace{1cm} (1.4.2)\]

$$\nabla \cdot u = 0 \quad \text{in } \Omega,$$  \hspace{1cm} (1.4.3)\]

$$u = u_\Gamma \quad \text{on } \Gamma.$$  \hspace{1cm} (1.4.4)\]

The incompressibility condition implies that $\nabla \cdot \mathbf{D}(u) = \Delta u$ and the reader will observe that (1.4.2)-(1.4.4) are the familiar steady-state Navier-Stokes equations.

1.4.2 Johnson-Segalman Fluids

A class of viscoelastic fluids which have differential-type constitutive equations are Johnson-Segalman fluids [47]. For these fluids, the extra stress tensor is assumed to have a Newtonian part and a viscoelastic part:

$$\sigma = \sigma_N + \sigma_V.$$

The Newtonian part is described by

$$\sigma_N = 2(1 - \alpha)\mathbf{D}(u),$$

for $0 \leq \alpha \leq 1$, where $\alpha$ may be physically interpreted as the portion of the fluid viscosity that is polymeric. For notational simplicity let $\sigma$ represent the viscoelastic part of the extra stress.

It is necessary to first introduce the function $g_\alpha$ defined by

$$g_\alpha(\sigma, \nabla u) := \frac{1-a}{2}(\sigma \nabla u + \nabla u^T \sigma) - \frac{1+a}{2}(\nabla u \sigma + \sigma \nabla u^T),$$  \hspace{1cm} (1.4.5)\]
for $a \in [-1, 1]$. This function represents a linear combination of the stress-velocity gradient products in both the upper-convected derivative and lower-convected derivative. We have that (1.4.5) can be written as

$$g_a(\sigma, \nabla u) := \sigma W(u) - W(u) \sigma - a(D(u) \sigma + \sigma D(u)),$$

(1.4.6)

where

$$W(\nabla u) = \frac{1}{2}(\nabla u - \nabla u^T)$$

is the vorticity tensor. With these definitions, the steady-state Johnson-Segalman constitutive law can be written as

$$\sigma + \lambda ((u \cdot \nabla) \sigma + g_a(\sigma, \nabla u)) = 2\alpha D(u),$$

(1.4.7)

where $\lambda$ is a dimensionless physical parameter known as the Weissenberg number. The Weissenberg number is a characteristic relaxation time of the fluid and can be thought of as a measure of the fluid’s elasticity.

Let $\Gamma_{in}$ denote the subset of $\Gamma$ on which fluid flows into $\Omega$. The velocity on the boundary is specified to be $u_{\Gamma}$, and the extra stress on $\Gamma_{in}$ is given as $\sigma_{\Gamma_{in}}$. The flow is assumed to be slow enough (creeping) such that the convective term $u \cdot \nabla u$ can be neglected. Together with the momentum and mass equations (1.3.10) and (1.3.11), the steady-state model of a Johnson-Segalman fluid is given by

$$\sigma + \lambda (u \cdot \nabla) \sigma + \lambda g_a(\sigma, \nabla u) - 2\alpha D(u) = 0 \text{ in } \Omega,$$

(1.4.8)

$$-\nabla \cdot \sigma - 2(1 - \alpha) \nabla \cdot D(u) + \nabla p = f \text{ in } \Omega,$$

(1.4.9)

$$\text{div } u = 0 \text{ in } \Omega,$$

(1.4.10)

$$u = u_{\Gamma} \text{ on } \Gamma,$$

(1.4.11)

$$\sigma = \sigma_{\Gamma_{in}} \text{ on } \Gamma_{in}.$$

(1.4.12)

**Remark 1.4.1** For the case $a = 1$ the Johnson-Segalman model reduces to the well-known Oldroyd-B model [8, 69].
1.4.3 Shear-Thinning Fluids

Some shear thinning fluids, such as blood and paint, obey constitutive laws in which the extra stress can be expressed explicitly as a nonlinear function of the velocity gradient. Several examples of such constitutive laws are:

**Power Law**

\[ \sigma = \nu_0 |\nabla u|^{r-2} \nabla u, \quad \nu_0 > 0, \quad 1 < r < 2. \quad (1.4.13) \]

The power law model has been used to model the viscosity of many polymeric solutions and melts over a considerable range of shear rates \[44\].

**Ladyzhensky Law**\[54\]:

\[ \sigma = (\nu_0 + \nu_1 |\nabla u|^{r-2}) \nabla u, \quad \nu_0 \geq 0, \quad \nu_1 > 0, \quad r > 1, \quad (1.4.14) \]

which has been used in modeling fluids with large stresses.

**Carreau Law**:

\[ \sigma = \nu_0 \left(1 + |\nabla u|^2\right)^{(r-2)/2} \nabla u, \quad \nu_0 > 0, \quad r \geq 1, \quad (1.4.15) \]

used in modeling visco-plastic flows and creeping flow of metals.

In a general sense, the constitutive laws (1.4.13),(1.4.14),(1.4.15), can be written as

\[ \sigma := g(\nabla u) \text{ in } \Omega, \quad (1.4.16) \]

where \( g \) is a nonlinear operator with specific properties (further discussion of such properties will be postponed until Chapter 4). As with Johnson-Segalman fluids, the flow is assumed to be slow so that the convective term \( u \cdot \nabla u \) can be neglected in the momentum equations.
The equations (1.3.10) and (1.3.11), together with a boundary condition on $\mathbf{u}$, the steady-state model of a shear-thinning fluid can be written as

$$\mathbf{\sigma} = g(\nabla \mathbf{u}) \quad \text{in } \Omega, \quad (1.4.17)$$

$$-\nabla \cdot \mathbf{\sigma} + \nabla p = f \quad \text{in } \Omega, \quad (1.4.18)$$

$$\text{div} \mathbf{u} = 0 \quad \text{in } \Omega, \quad (1.4.19)$$

$$\mathbf{u} = \mathbf{u}_\Gamma \quad \text{on } \Gamma. \quad (1.4.20)$$
CHAPTER 2
DEFECT-CORRECTION METHODS FOR JOHNSON-SEGALMAN FLUIDS

2.1 Introduction

As mentioned in Section 1.1, the numerical approximation of viscoelastic fluid flows obeying a Johnson-Segalman constitutive law becomes more difficult as the Weissenberg number increases. One approach for the numerical approximation of high Weissenberg number flows is defect-correction methods [10]. A defect correction method has two basic steps. The first step uses a stable, nearby problem to form an initial approximation to the solution of the original given problem. The second step (correction step) iteratively improves the defected approximation using residual corrections. Defect correction methods have been applied to convection-diffusion problems [4, 28, 46] as well as to handle difficulties in computing solutions to the Navier-Stokes equations for high Reynolds number [55, 56].

For viscoelastic fluid flows, in [57] Lee presented a method in which the defect step consisted of solving the nonlinear problem for a reduced Weissenberg number on the convective term in the constitutive equation. In the corrector step an iterative, linear residual correction algorithm was used. The error estimate showed that the defect correction method preserved the optimal order of convergence for the discretization scheme. In [29], a defect-correction method was applied to the linear Oseen-viscoelastic model problem. This approach, in the defect step, reduced the Weissenberg number independently on both the convective and stress-deformation interaction terms in the constitutive equation. Again an iterative linear corrector was used.

This chapter extends the work of [57] and [29]. Specifically, a two-parameter defect-correction algorithm for steady-state viscoelastic fluid flows obeying the nonlinear Johnson-Segalman constitutive model is investigated. To compute solutions for a Weissenberg number larger than the critical value, the original problem is defected by reducing the Weissenberg number to form a nearby, stable problem. These defect parameters can be chosen
independently of each other. The initial approximation is then corrected with a linear iteration which, upon convergence, solves the original undefected problem.

This chapter is organized as follows. In Section 2.2 the continuous problem and a variational formulation are described in an appropriate setting. A corresponding finite element approximation using the discontinuous Galerkin method is described in Section 2.3. In Section 2.4 the defect-correction algorithm is presented and convergence properties of the method are shown. An alternate corrector iteration is also discussed. In Section 2.5 numerical results are presented to demonstrate the accuracy and effectiveness of the method.

2.2 Problem Description

Consider the viscoelastic (inertialess) fluid flow problem subject to the Johnson-Segalman constitutive equation (1.4.8)-(1.4.12) with homogeneous boundary conditions:

\[ \sigma + \lambda (u \cdot \nabla)\sigma + \lambda g_a(\sigma, \nabla u) - 2\alpha D(u) = 0 \quad \text{in } \Omega, \]  
\[ -\nabla \cdot \sigma - 2(1 - \alpha) \nabla \cdot D(u) + \nabla p = f \quad \text{in } \Omega, \]  
\[ \text{div } u = 0 \quad \text{in } \Omega, \]  
\[ u = 0 \quad \text{on } \Gamma. \]

Remark 2.2.1 As there is no inflow boundary \( \Gamma_{in} \), no boundary condition is specified for \( \sigma \).

2.2.1 Continuous Problem

Existence of a solution to the problem (2.2.1)-(2.2.4) was shown by Renardy [68] under a small data assumption. (See also [41] and [33].) Specifically, if \( \Omega \) has a \( C^\infty \)-smooth boundary and \( f \) is sufficiently regular and small, the problem (2.2.1)-(2.2.4) admits a unique bounded solution \( (\sigma, u, p) \in H^2(\Omega) \times H^3(\Omega) \times H^2(\Omega) \).
2.2.2 Variational Formulation

Define the function spaces for the velocity $u$, the pressure $p$ and the stress $\sigma$. Let

$X := H^1_0(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma\},$

$S := L^2_0(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q \, d\Omega = 0\},$

$\Sigma := (L^2(\Omega))^{d \times d} \cap \{\tau = (\tau_{ij}) : \tau_{ij} = \tau_{ji}, \ u \cdot \nabla \tau \in (L^2(\Omega))^{d \times d}\}.$

Introduce the divergence free space

$V = \{v \in X : \int_\Omega q \text{div} v \, d\Omega = 0 \quad \forall q \in S\}.$

The corresponding variational formulation of (2.2.1)-(2.2.4) is obtained by taking the inner product of (2.2.1)-(2.2.3) with stress, velocity, and pressure test functions $\tau, v, \text{ and } q$ respectively. Specifically the variational formulation is: Given $f \in H^{-1}(\Omega)$, find $(\sigma, u, p) \in \Sigma \times X \times S$ such that

$(\sigma, \tau) + \lambda((u \cdot \nabla)\sigma, \tau) + \lambda(g_a(\sigma, \nabla u), \tau) - 2\alpha(D(u), \tau) = 0, \quad \forall \tau \in \Sigma; \quad (2.2.1)$

$(\sigma, D(v)) + 2(1 - \alpha)(D(u), D(v)) - (p, \nabla \cdot v) = (f, v), \quad \forall v \in X; \quad (2.2.2)$

$(q, \nabla \cdot u) = 0, \quad \forall q \in S. \quad (2.2.3)$

Using the divergence free space $V$, the variational formulation (2.2.1)-(2.2.3) is equivalent to:

$(\sigma, \tau) + \lambda((u \cdot \nabla)\sigma, \tau) + \lambda(g_a(\sigma, \nabla u), \tau) - 2\alpha(D(u), \tau) = 0, \quad \forall \tau \in \Sigma; \quad (2.2.4)$

$(\sigma, D(v)) + 2(1 - \alpha)(D(u), D(v)) = (f, v), \quad \forall v \in V. \quad (2.2.5)$

Introduce the bilinear form $A : (\Sigma \times X) \times (\Sigma \times X) \to \mathbb{R}$ by

$A((\sigma, u), (\tau, v)) := (\sigma, \tau) - 2\alpha(D(u), \tau) + 2\alpha(\sigma, D(v)) + 4\alpha(1 - \alpha)(D(u), D(v)). \quad (2.2.6)$

The continuity and coercivity of $A$ is shown in the following lemma.
Lemma 2.2.1 The bilinear form $A$ defined in (2.2.6) is continuous and coercive, i.e.,
a) (continuity) there exists a positive constant $C$ such that

$$A((\sigma, u), (\tau, v)) \leq C\|\sigma\|_{\Sigma \times X}\|\tau\|_{\Sigma \times X}, \quad \forall (\tau, v) \in \Sigma \times X,$$

b) (coercivity) there exists a positive constant $C$ such that

$$A((\sigma, u), (\sigma, u)) \geq C\|\sigma\|_{\Sigma \times X}^2, \quad \forall (\sigma, u) \in \Sigma \times X.$$

Proof: The reader is referred to Lemma 1.1 of [57].

\[\square\]

2.3 Finite Element Approximation

Of the many different numerical approximation approaches for viscoelastic fluid flows, methods which utilize a finite element approach are particularly prominent, and they usually incorporate some stabilization to handle the hyperbolic behavior of the constitutive model of the fluid. Some examples of such stabilization methods include the streamline upwind Petrov-Galerkin (SUPG) method, the discontinuous Galerkin (DG) method, and elastic-viscous split-stress (EVSS) method (see [5]). In this work, stabilization is accomplished by using the discontinuous Galerkin method. The extra stress tensor is approximated with piecewise linear functions that are allowed to be discontinuous from one element to the next. Stabilization is incorporated into the variational formulation through the inclusion of a term that accounts for the jump in the stress approximation across element boundaries.

Let $T_h$ denote a triangulation of $\Omega$ such that $\overline{\Omega} = \{\bigcup K : K \in T_h\}$. Assume that there exist positive constants $c_1, c_2$ such that

$$c_1 h \leq h_K \leq c_2 \rho_K,$$

where $h_K$ is the diameter of $K$, $\rho_K$ is the diameter of the greatest ball included in $K$, and $h = \max_{K \in T_h} h_K$. Let $P_k(K)$ denote the space of polynomials of degree less than or equal to $k$ on $K \in T_h$. We define the following finite element spaces, (Taylor-Hood) for the
approximation of \((u, p)\):

\[
X^h := \{ v \in X \cap (C^0(\Omega))^d : v|_K \in \mathcal{P}_2(K)^d, \forall K \in T_h \},
\]

\[
S^h := \{ q \in S \cap C^0(\Omega) : q|_K \in \mathcal{P}_1(K), \forall K \in T_h \},
\]

\[
V^h := \{ v \in X^h : (q, \nabla \cdot v) = 0, \forall q \in S^h \}.
\]

For the discontinuous Galerkin approximation of the stress, \(\sigma\) is approximated in the discontinuous finite element space of piecewise linears:

\[
\Sigma^h := \{ \tau \in \Sigma : \tau|_K \in \mathcal{P}_1(K)^{d \times d}, \forall K \in T_h \}.
\]

The finite element spaces defined above satisfy the standard approximation properties (see [12] or [38]), i.e., there exists a constant \(C\) such that

\[
\inf_{v^h \in X^h} \| v - v^h \|_1 \leq Ch^2 \| v \|_3, \quad \forall v \in H^3(\Omega), \quad (2.3.1)
\]

\[
\inf_{q^h \in S^h} \| q - q^h \|_0 \leq Ch^2 \| q \|_2, \quad \forall q \in H^2(\Omega), \quad (2.3.2)
\]

and

\[
\inf_{\tau^h \in \Sigma^h} \| \tau - \tau^h \|_0 \leq Ch^2 \| \tau \|_2, \quad \forall \tau \in H^2(\Omega). \quad (2.3.3)
\]

It is also well known that the Taylor-Hood pair \((X^h, S^h)\) satisfies the \(\text{inf-sup}\) (or \(LBB\)) condition ([12]),

\[
\inf_{0 \neq q^h \in S^h} \sup_{0 \neq v^h \in X^h} \frac{(q^h, \nabla \cdot v^h)}{\| v^h \|_1 \| q^h \|_0} \geq C, \quad (2.3.4)
\]

where \(C\) is a positive constant independent of \(h\).

Below some notation used in [8] is introduced to describe and analyze an approximate solution obtained using the discontinuous Galerkin method. Let

\[
\Gamma^h = \{ \cup \partial K, K \in T_h \} \setminus \Gamma, \quad (2.3.5)
\]

\[
\partial K^-(u) := \{ x \in \partial K, \ u \cdot n < 0 \},
\]

where \(\partial K\) is the boundary of \(K \in T_h\) and \(n\) is the outward unit normal vector on \(\partial K\), and

\[
\tau^\pm(u) := \lim_{\epsilon \to 0^\pm} \tau(x + \epsilon u(x)).
\]
Also, let
\[
(\sigma, \tau)_h := \sum_{K \in T_h} (\sigma, \tau)_K ,
\]
\[
\langle \sigma^\pm, \tau^\pm \rangle_{h,u} := \sum_{K \in T_h} \int_{\partial K^- (u)} (\sigma^\pm (u) : \tau^\pm (u)) |n \cdot u| \, ds ,
\]
\[
\|\tau\|_{0, \Gamma^h} := \left( \sum_{K \in T_h} \|\tau\|_{0, \partial K}^2 \right)^{1/2} ,
\]
for \( \sigma, \tau \in \prod_{K \in T_h} (L^2 (K))^{d \times d} \), and
\[
\|\xi\|_{m, h} := \left( \sum_{K \in T_h} \|\xi\|_{m,K}^2 \right)^{1/2} ,
\]
for \( \xi \in \prod_{K \in T_h} (W^{m,2} (K))^{d \times d} \).

Introduce the operator \( B_h \) on \( X_h \times \Sigma_h \times \Sigma_h \) defined by
\[
B_h (u^h, \sigma^h, \tau^h) := ((u^h \cdot \nabla) \sigma^h, \tau^h)_h + \frac{1}{2} (\nabla \cdot u^h \sigma^h, \tau^h)_h + (\sigma^{h+} - \sigma^{h-}, \tau^{h+})_{h,u^h} .
\]  

(2.3.7)

Note that the second term vanishes when \( \nabla \cdot u^h = 0 \) (this is not necessarily the case for any \( u^h \in X^h \)). This extra term is used to obtain coercivity of \( B_h (u^h, \cdot, \cdot) \). Using integration by parts, \( B_h \) may be written as
\[
B_h (u^h, \sigma^h, \tau^h) = -((u^h \cdot \nabla) \tau^h, \sigma^h)_h - \frac{1}{2} (\nabla \cdot u^h \tau^h, \sigma^h)_h + (\sigma^{h-}, \tau^{h+} - \tau^{h-})_{h,u^h} .
\]  

(2.3.8)

Combining (2.3.7) and (2.3.8), we obtain
\[
B_h (u^h, \sigma^h, \sigma^h) = \frac{1}{2} (\sigma^{h+} - \sigma^{h-}, \sigma^{h+} - \sigma^{h-})_{h,u^h} = \frac{1}{2} (\langle \sigma^{h+} - \sigma^{h-} \rangle_{h,u^h})^2 \geq 0 .
\]  

(2.3.9)
The discontinuous Galerkin finite element approximation of (2.2.1)–(2.2.3) is then as follows. Given \( \mathbf{f} \in H^{-1}(\Omega) \), find \((\mathbf{\sigma}^h, \mathbf{u}^h, p^h) \in \Sigma^h \times \mathbf{X}^h \times \mathbf{S}^h\) such that

\[
(\mathbf{\sigma}^h, \mathbf{\tau}^h) + \lambda B^h(\mathbf{u}^h, \mathbf{\sigma}^h, \mathbf{\tau}^h) + \lambda (g_a(\mathbf{\sigma}^h, \nabla \mathbf{u}^h), \mathbf{\tau}^h) - 2\alpha(D(\mathbf{u}^h), \mathbf{\tau}^h) = 0, \quad \forall \mathbf{\tau}^h \in \Sigma^h, \tag{2.3.10}
\]

\[
(\mathbf{\sigma}^h, D(\mathbf{v}^h)) + 2(1 - \alpha)(D(\mathbf{u}^h), D(\mathbf{v}^h)) - (p^h, \nabla \cdot \mathbf{v}^h) = (\mathbf{f}, \mathbf{v}^h), \quad \forall \mathbf{v}^h \in \mathbf{X}^h, \tag{2.3.11}
\]

\[
(q^h, \nabla \cdot \mathbf{u}^h) = 0, \quad \forall q^h \in \mathbf{S}^h. \tag{2.3.12}
\]

Existence of a solution to the discrete problem (2.3.10)-(2.3.12) has been shown by Baranger and Sandri [8] under the assumption that the continuous problem (2.2.1)-(2.2.4) yields a bounded solution \((\mathbf{u}, \mathbf{\sigma}, p) \in H^3(\Omega) \times H^2(\Omega) \times H^2(\Omega)\). The error estimates

\[
\|\mathbf{\sigma} - \mathbf{\sigma}^h\|_0 + \|\nabla(\mathbf{u} - \mathbf{u}^h)\|_0 \leq C h^{3/2}, \quad \|p - p^h\|_0 \leq C h^{3/2}
\]

for constant \(C > 0\), are also proven in [8].

Notice that, in view of (2.3.4), (2.3.10)-(2.3.12) is equivalent to: Given \( \mathbf{f} \in H^{-1}(\Omega) \), find \((\mathbf{\sigma}^h, \mathbf{u}^h) \in \Sigma^h \times \mathbf{V}^h\) such that

\[
(\mathbf{\sigma}^h, \mathbf{\tau}^h) + \lambda B^h(\mathbf{u}^h, \mathbf{\sigma}^h, \mathbf{\tau}^h) + \lambda (g_a(\mathbf{\sigma}^h, \nabla \mathbf{u}^h), \mathbf{\tau}^h) - 2\alpha(D(\mathbf{u}^h), \mathbf{\tau}^h) = 0, \quad \forall \mathbf{\tau}^h \in \Sigma^h, \tag{2.3.13}
\]

\[
(\mathbf{\sigma}^h, D(\mathbf{v}^h)) + 2(1 - \alpha)(D(\mathbf{u}^h), D(\mathbf{v}^h)) = (\mathbf{f}, \mathbf{v}^h), \quad \forall \mathbf{v}^h \in \mathbf{V}^h. \tag{2.3.14}
\]

Using the bilinear form \( A \) defined by (2.2.6), (2.3.13)-(2.3.14) can equivalently be written as

\[
A((\mathbf{\sigma}^h, \mathbf{u}^h), (\mathbf{\tau}^h, \mathbf{v}^h)) + \lambda B^h(\mathbf{u}^h, \mathbf{\sigma}^h, \mathbf{\tau}^h) + \lambda (g_a(\mathbf{\sigma}^h, \nabla \mathbf{u}^h), \mathbf{\tau}^h) = 2\alpha(\mathbf{f}, \mathbf{v}^h) \quad \forall (\mathbf{\tau}^h, \mathbf{v}^h) \in \Sigma^h \times \mathbf{V}^h. \tag{2.3.15}
\]

In the following analysis several inverse estimates are used. For convenience, these results are summarized here. (See [12], [38]). Assume \( n = 2 \). For \( \mathbf{u}^h \in \mathbf{V}^h \) and \( \mathbf{\sigma}^h \in \Sigma^h \),
we have
\[ \| u^h \|_\infty \leq Ch^{-1/2}\| u^h \|_{0,4}, \] (2.3.16)
\[ \| \sigma^h \|_\infty \leq Ch^{-1}\| \sigma^h \|_0, \] (2.3.17)
\[ \| \nabla \sigma^h \|_{0,4,h} \leq Ch^{-3/2}\| \sigma^h \|_0, \] (2.3.18)
\[ \| \nabla \sigma^h \|_{0,2,h} \leq Ch^{-1}\| \sigma^h \|_0. \] (2.3.19)

The local inverse inequality ([76]),
\[ \| \sigma \|_{0,\partial K}^2 \leq C\frac{p_K^2}{h_K}\| \sigma \|_{0,K}^2, \] (2.3.20)
is used to bound the jump term of $B^h$, where $p_K$ denotes the polynomial degree on mesh element $K$ and $h_K$ the local mesh parameter.

### 2.4 Defect Correction Method

In this section the defect correction method used in computing the solution to (2.3.10)-(2.3.12) is described. The idea behind the method is to avoid the approximation difficulties associated with a high Weissenberg number, $\lambda$-difficult, by computing an initial approximation using a $\lambda$-easy value, and then iteratively improving the approximation until the desired solution at $\lambda = \lambda$-difficult is obtained.

#### 2.4.1 Algorithm

To avoid the computational difficulties associated with the high Weissenberg number problem, replace $\lambda$ in the second and third terms of (2.3.10) with the parameters $\bar{\lambda}, \tilde{\lambda}$, respectively. These defect parameters are chosen so that $0 \leq \bar{\lambda}, \tilde{\lambda} \leq \lambda$. Let $(\sigma^h_0, u^h_0, p^h_0) \in \Sigma^h \times X^h \times S^h$ satisfy
\[
(\sigma^h_0, \tau^h) + \bar{\lambda}B^h(u^h_0, \sigma^h_0, \tau^h) + \tilde{\lambda}(g_a(\sigma^h_0, \nabla u^h_0), \tau^h) - 2\alpha(D(u^h_0), \tau^h) = 0, \quad \forall \tau^h \in \Sigma^h,
\]
\[
(\sigma^h_0, D(v^h)) + 2(1 - \alpha)(D(u^h_0), D(v^h)) - (p^h_0, \nabla \cdot v^h) = (f, v^h), \quad \forall v^h \in X^h,
\]
\[
(q^h, \nabla \cdot u^h_0) = 0, \quad \forall q^h \in S^h.
\]
Then the residuals \( R_1, R_2, R_3 \in \Sigma^h \times X^h \times S^h \) for equations (2.3.10)-(2.3.12) are defined by

\[
\left( R_1(u_0^h, \sigma_0^h, \tau^h) \right) := - (\sigma_0^h, \tau^h) - \lambda B^h(u_0^h, \sigma_0^h, \tau^h) - \lambda (g_a(\sigma_0^h, \nabla u_0^h), \tau^h) + 2\alpha(D(u_0^h), \tau^h), \quad \forall \tau^h \in \Sigma^h,
\]

(2.4.1)

\[
\left( R_2(u_0^h, \sigma_0^h, p_0^h), \nu^h \right) := (f, \nu^h) - (\sigma_0^h, D(\nu^h)) - 2(1 - \alpha)(D(u_0^h), D(\nu^h)) + (p_0^h, \nabla \cdot \nu^h), \quad \forall \nu^h \in X^h,
\]

(2.4.2)

\[
\left( R_3(u_0^h), q^h \right) := -(q^h, \nabla \cdot u_0^h), \quad \forall q^h \in S^h.
\]

(2.4.3)

Define the correction \( (\xi_0^h, \epsilon_0^h, \rho_0^h) \in \Sigma^h \times X^h \times S^h \) to the approximation \( (\sigma_0^h, u_0^h, p_0^h) \) via

\[
(\xi_0^h, \tau^h) + \bar{\lambda} B^h(u_0^h, \xi_0^h, \tau^h) + \bar{\lambda}(g_a(\xi_0^h, \nabla u_0^h), \tau^h) - 2\alpha(D(\epsilon_0^h), \tau^h))
\]

\[
= \left( R_1(u_0^h, \sigma_0^h), \tau^h \right), \quad \forall \tau^h \in \Sigma^h,
\]

(2.4.4)

\[
(\xi_0^h, D(\nu^h)) + 2(1 - \alpha)(D(\epsilon_0^h), D(\nu^h)) - (\rho_0^h, \nabla \cdot \nu^h) = \left( R_2(u_0^h, \sigma_0^h, p_0^h), \nu^h \right), \quad \forall \nu^h \in X^h,
\]

(2.4.5)

\[
(q^h, \nabla \cdot \epsilon_0^h) = \left( R_3(u_0^h), q^h \right), \quad \forall q^h \in S^h.
\]

(2.4.6)

The update \( (\sigma_1^h, u_1^h, p_1^h) := (\sigma_0^h + \xi_0^h, u_0^h + \epsilon_0^h, p_0^h + \rho_0^h) \) is expected to be a better approximation to \( (\sigma_0^h, u_0^h, p_0^h) \) than \( (\sigma_0^h, u_0^h, p_0^h) \). Note that combining (2.4.1)-(2.4.3) with (2.4.4)-(2.4.6), \( (\sigma_1^h, u_1^h, p_1^h) \) satisfies:

\[
(\sigma_1^h, \tau^h) + \bar{\lambda} B^h(u_1^h, \sigma_1^h, \tau^h) + \bar{\lambda}(g_a(\sigma_1^h, \nabla u_1^h), \tau^h) - 2\alpha(D(u_1^h), \tau^h))
\]

\[
= -(\lambda - \bar{\lambda})B^h(u_0^h, \sigma_0^h, \tau^h) - (\lambda - \bar{\lambda})(g_a(\sigma_0^h, \nabla u_0^h), \tau^h), \quad \forall \tau^h \in \Sigma^h
\]

(2.4.7)

\[
(\sigma_1^h, D(\nu^h)) + 2(1 - \alpha)(D(u_1^h), D(\nu^h)) - (p_1^h, \nabla \cdot \nu^h) = (f, \nu^h), \quad \forall \nu^h \in X^h
\]

(2.4.8)

\[
(q^h, \nabla \cdot u_1^h) = 0, \quad \forall q^h \in S^h.
\]

(2.4.9)
The second step can be repeated with \((\sigma_0^h, u_0^h, p_0^h)\) replaced by \((\sigma_i^h, u_i^h, p_i^h)\) to further improve the approximation. We summarize the defect-correction method in Algorithm 2.4.1.

**Algorithm 2.4.1 (Defect-Correction Method for the Johnson-Segalman model)**

**Step 1:** Solve the nonlinear defected problem: Find \((\sigma_0^h, u_0^h, p_0^h) \in \Sigma^h \times X^h \times S^h\) such that

\[
(\sigma_0^h, \tau^h) + \bar{\lambda}B^h(u_0^h, \sigma_0^h, \tau^h) + \bar{\lambda}(g_0(\sigma_0^h, \nabla u_0^h), \tau^h) - 2\alpha(D(u_0^h), \tau^h) = 0, \quad \forall \tau^h \in \Sigma^h,
\]

\[
(\sigma_0^h, D(v^h)) + 2(1 - \alpha)(D(u_0^h), D(v^h)) - (p_0^h, \nabla \cdot v^h) = (f, v^h), \quad \forall v^h \in X^h,
\]

\[
(q^h, \nabla \cdot u_0^h) = 0, \quad \forall q^h \in S^h,
\]

where \(\bar{\lambda}\) and \(\bar{\lambda}\) are chosen to be less than or equal to \(\lambda\).

**Step 2:** For \(i = 0, 1, 2, \ldots\), solve the following problem for the correction: Find \((\sigma_i^h, u_i^h, p_i^h) \in \Sigma^h \times X^h \times S^h\) such that

\[
(\sigma_i^h, \tau^h) + \bar{\lambda}B^h(u_i^h, \sigma_i^h, \tau^h) + \bar{\lambda}(g_0(\sigma_i^h, \nabla u_i^h), \tau^h) - 2\alpha(D(u_i^h), \tau^h))
\]

\[
= -\lambda - \bar{\lambda}B^h(u_i^h, \sigma_i^h, \tau^h) - (\lambda - \bar{\lambda})(g_0(\sigma_i^h, \nabla u_i^h), \tau^h), \quad \forall \tau^h \in \Sigma^h,
\]

\[
(\sigma_i^h, D(v^h)) + 2(1 - \alpha)(D(u_i^h), D(v^h)) - (p_i^h, \nabla \cdot v^h) = (f, v^h), \quad \forall v^h \in X^h,
\]

\[
(q^h, \nabla \cdot u_i^h) = 0, \quad \forall q^h \in S^h.
\]

### 2.4.2 Analysis of the Defect Step

Assume \((\sigma, u, p)\) is an exact solution of (2.2.1)-(2.2.3) (the undefected continuous problem). Also let \((\bar{\sigma}, \bar{u}, \bar{p})\) be an exact solution of (2.2.1)-(2.2.3) with the first \(\lambda\) replaced with \(\bar{\lambda}\) and the second \(\lambda\) replaced with \(\bar{\lambda}\) (the defected continuous problem). Let \(M\) be given by

\[
M := \max\{|\sigma|_2, |u|_2, |p|_2, |\bar{\sigma}|_3, |\bar{u}|_3, |\bar{p}|_2\}. \quad (2.4.7)
\]

**Lemma 2.4.1** Let \(\bar{\lambda} = \lambda - K_1 h^{3/2}\) and \(\bar{\lambda} = \lambda - K_2 h^{3/2}\) where \(0 \leq K_1 h^{3/2}, K_2 h^{3/2} \leq \lambda\). If there exists a solution \((\sigma_0^h, u_0^h, p_0^h) \in \Sigma^h \times X^h \times S^h\), then for \(M\) and \(h\) sufficiently small, Step 1 of Algorithm 2.4.1 admits a solution \((\sigma_i^h, u_i^h, p_i^h) \in \Sigma^h \times X^h \times S^h\) and there exists a constant \(C_{1.4.1}\) such that

\[
|\sigma - \sigma_i^h|_0 + |\nabla(u - u_i^h)|_0 \leq C_{1.4.1} h^{3/2}. \quad (2.4.8)
\]
Proof: From the embedding properties of Sobolev spaces [1] we have that there exists a constant $C_M$ such that

$$\|v\|_{0,4} \leq C_M \|\nabla v\|_{0,2} \quad \forall v \in H^1_0(\Omega),$$  \hspace{1cm} (2.4.9)

$$\|v\|_{0,\infty} \leq C_M \|v\|_{2,2} \quad \forall v \in H^2(\Omega).$$ \hspace{1cm} (2.4.10)

Existence of a solution to Step 1 is shown by Baranger and Sandri [8] for $M$ sufficiently small. For the proof of the error estimate, let $(\overline{\sigma}, \overline{u}, \overline{p})$ be a solution of (2.2.1)-(2.2.4) with the first $\lambda$ replaced with $\overline{\lambda}$ and the second $\lambda$ replaced with $\tilde{\lambda}$. Then, from [8], we have that

$$\|\sigma - \sigma^h_0\|_0 + \|\nabla(u - u^h_0)\|_0 \leq \tilde{C} h^{3/2},$$ \hspace{1cm} (2.4.11)

for some positive constant $\tilde{C}$. Note that the solution $(\sigma, u)$ satisfies the equations

$$(\sigma, \tau) + \overline{\lambda} (u \cdot \nabla)\sigma, \tau) + \tilde{\lambda} (g_a(\sigma, \nabla u), \tau) - 2\alpha (D(u), \tau) = 0, \quad \forall \tau \in \Sigma,$$ \hspace{1cm} (2.4.12)

$$(\sigma, D(v)) + 2(1 - \alpha)(D(u), D(v)) = (f, v), \quad \forall v \in V.$$ \hspace{1cm} (2.4.13)

To show the stated result, we obtain an estimation for $\|\nabla(u - \overline{u})\|_0$ and $\|\sigma - \sigma^h\|_0$. Subtracting (2.4.12)-(2.4.13) from (2.2.4)-(2.2.5) and using the bilinear form $A$, we have

$$A((\sigma - \sigma^h, u - \overline{u}), (\tau, v))$$

$$+ \lambda [((u \cdot \nabla)\sigma, \tau) - ((\overline{u} \cdot \nabla)\sigma, \tau)] + \tilde{\lambda} [(g_a(\sigma, \nabla u), \tau) - (g_a(\sigma^h, \nabla \overline{u}), \tau)]$$

$$+ (\lambda - \overline{\lambda})((u \cdot \nabla)\sigma, \tau) + (\lambda - \tilde{\lambda})(g_a(\sigma, \nabla u), \tau) = 0, \quad \forall (\tau, v) \in \Sigma \times V.$$ \hspace{1cm} (2.4.14)

Let $\tau = \sigma - \sigma^h$ and $v = u - \overline{u}$. Equation (2.4.14) becomes

$$\|\sigma - \sigma^h\|_0^2 + 4\alpha(1 - \alpha)\|\nabla(u - \overline{u})\|_0^2$$

$$+ \lambda [((u \cdot \nabla)(\sigma - \sigma^h + \sigma), \sigma - \sigma^h) - ((\overline{u} \cdot \nabla)\sigma, \sigma - \sigma^h)]$$

$$+ \tilde{\lambda} [(g_a(\sigma, \nabla(u - \overline{u} + \overline{u})), \sigma - \sigma^h) - (g_a(\sigma^h - \sigma^h + \sigma^h, \nabla \overline{u}), \sigma - \sigma^h)]$$

$$+ (\lambda - \overline{\lambda})((u \cdot \nabla)\sigma, \sigma - \sigma^h) + (\lambda - \tilde{\lambda})(g_a(\sigma, \nabla u), \sigma - \sigma^h) = 0. \hspace{1cm} (2.4.15)$$
Keeping the “positive terms” on the LHS and moving the other terms to the RHS we obtain

\[
\|\sigma - \sigma\|_0^2 + 4\alpha(1 - \alpha)\|\nabla(u - \overline{u})\|_0^2 + \lambda((u \cdot \nabla)(\sigma - \overline{\sigma}), \sigma - \overline{\sigma})
\]

\[
\leq \lambda((u - \overline{u} \cdot \nabla)\sigma, \sigma - \overline{\sigma}) + \lambda|g_a(\sigma, \nabla(u - \overline{u})), \sigma - \overline{\sigma}) + (g_a(\sigma - \overline{\sigma}, \nabla\overline{u}), \sigma - \overline{\sigma})| + (\lambda - \lambda)(u \cdot \nabla)\sigma, \sigma - \overline{\sigma}) + (\lambda - \lambda)(g_a(\sigma, \nabla\overline{u}), \sigma - \overline{\sigma})].
\]

(2.4.16)

Next we bound the terms on the RHS of (2.4.16). Using the estimates (2.4.9) and (2.4.10) we have

\[
|((u - \overline{u} \cdot \nabla)\sigma, \sigma - \overline{\sigma})| \leq \|(u - \overline{u} \cdot \nabla)\sigma\|_0\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq \|u - \overline{u}\|_{0,4}\|\nabla\sigma\|_0\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq C_M^2\|\nabla(u - \overline{u})\|_0\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq C_M^2 M\|\nabla(u - \overline{u})\|_0\|\sigma - \overline{\sigma}\|_0, \quad (2.4.17)
\]

\[
|g_a(\sigma, \nabla(u - \overline{u})), \sigma - \overline{\sigma}) + (g_a(\sigma - \overline{\sigma}, \nabla\overline{u}), \sigma - \overline{\sigma})|
\]

\[
\leq 4\|\nabla(u - \overline{u})\|_0\|\sigma - \overline{\sigma}\|_0 + 4\|(\sigma - \overline{\sigma})\nabla\overline{u}\|_0\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq 8\|\sigma\|_\infty\|\nabla(u - \overline{u})\|_0\|\sigma - \overline{\sigma}\|_0 + 8\|\sigma - \overline{\sigma}\|_0\|\nabla\overline{u}\|_\infty\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq 8C_M\|\sigma\|_2\|\nabla(u - \overline{u})\|_0\|\sigma - \overline{\sigma}\|_0 + 8C_M\|\sigma - \overline{\sigma}\|_0\|\nabla\overline{u}\|_3\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq 8C_M M\|\nabla(u - \overline{u})\|_0\|\sigma - \overline{\sigma}\|_0 + 8C_M M\|\sigma - \overline{\sigma}\|_0^2, \quad (2.4.18)
\]

\[
|((u \cdot \nabla)\sigma, \sigma - \overline{\sigma})| \leq \|u\|_{0,4}\|\nabla\sigma\|_{0,4}\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq C_M^2\|u\|_3\|\sigma\|_2\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq C_M^2 M^2\|\sigma - \overline{\sigma}\|_0, \quad (2.4.19)
\]

and

\[
|g_a(\sigma, \nabla u), \sigma - \overline{\sigma})| \leq 8\|\sigma\|_\infty\|\nabla u\|_0\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq 8C_M\|u\|_3\|\sigma\|_2\|\sigma - \overline{\sigma}\|_0
\]

\[
\leq 8C_M^2 M^2\|\sigma - \overline{\sigma}\|_0. \quad (2.4.20)
\]
Now, (2.4.16) and the estimates (2.4.17)-(2.4.20) imply that

$$\|\sigma - \sigma\|_0^2 + 4\alpha(1 - \alpha)\|\nabla(u - \overline{u})\|_0^2 + \overline{\lambda}((u \cdot \nabla)(\sigma - \sigma), \sigma - \sigma)$$

$$\leq \left(\overline{\lambda}C_M^2M + 8\overline{\lambda}C_M M\right)\|\nabla(u - \overline{u})\|_0\|\sigma - \sigma\|_0 + 8\overline{\lambda}C_M M\|\sigma - \sigma\|_0^2$$

$$+ \left((\lambda - \overline{\lambda})C_M^2M^2 + 8(\lambda - \overline{\lambda})C_M M^2\right)\|\sigma - \sigma\|_0$$

$$\leq \left(\overline{\lambda}C_M^2M + 8\overline{\lambda}C_M M\right)\left(\frac{1}{4\varepsilon_1}\|\nabla(u - \overline{u})\|_0^2 + \varepsilon_1\|\sigma - \sigma\|_0^2\right)$$

$$+ 8\overline{\lambda}C_M M\|\sigma - \sigma\|_0^2 + \varepsilon_2M^4\left((\lambda - \overline{\lambda})C_M^2 + 8(\lambda - \overline{\lambda})C_M\right)^2 + \frac{1}{4\varepsilon_2}\|\sigma - \sigma\|_0^2.$$  

(2.4.21)

Let $\varepsilon_1 = \frac{1}{2}$ and $\varepsilon_2 = 1$. Now, since $\sigma$ and $\overline{\sigma}$ are continuous, (2.3.9) implies that $((u \cdot \nabla)(\sigma - \overline{\sigma}), \sigma - \overline{\sigma}) = 0$. Thus (2.4.21) gives

$$\left[\frac{3}{4} - \frac{M}{2}\left(\overline{\lambda}C_M^2 + 24\overline{\lambda}C_M\right)\right]\|\sigma - \sigma\|_0^2$$

$$+ \left[4\alpha(1 - \alpha) - \frac{M}{2}\left(\overline{\lambda}C_M^2 + 8\overline{\lambda}C_M\right)\right]\|\nabla(u - \overline{u})\|_0^2$$

$$\leq M^4\left(K_1h^{3/2}C_M^2 + 8K_2h^{3/2}C_M\right)^2 = M^4C_M^2h^3(K_1C_M + 8K_2)^2.$$  

(2.4.22)

Therefore, if

$$M < \min\left\{\frac{3}{2\overline{\lambda}C_M + 8\overline{\lambda}C_M}, \frac{8\alpha(1 - \alpha)}{\overline{\lambda}C_M^2 + 8\overline{\lambda}C_M}\right\},$$  

(2.4.23)

then

$$\|\sigma - \sigma\|_0^2 + \|\nabla(u - \overline{u})\|_0^2 \leq C_0^2h^3,$$  

(2.4.24)

where

$$C_0^2 = M^4C_M^2(K_1C_M + 8K_2)^2/\min\left\{\frac{3}{4} - \frac{M}{2}\left(\overline{\lambda}C_M^2 + 24\overline{\lambda}C_M\right), \frac{4\alpha(1 - \alpha) - \frac{M}{2}\left(\overline{\lambda}C_M^2 + 8\overline{\lambda}C_M\right)}{2}\right\}.$$  

Thus, (2.4.8) follows from (2.4.11) and (2.4.24).
2.4.3 Analysis of the Correction Step

Step 1 of Algorithm 2.4.1 produces an initial approximation \((\sigma_0^h, u_0^h) \in \Sigma^h \times V^h\) that is the solution of the stable defected problem. This approximation is within a certain radius of the solution to the undefected problem. Step 2 of the algorithm provides an iterative procedure for correcting this initial iterate to an approximate solution of the original undefected problem. When convergent the computed approximation \((\sigma^h, u^h)\) will satisfy the same error estimates shown by Baranger and Sandri [8], namely (2.4.8) with \((\sigma_0^h, u_0^h)\) replaced by \((\sigma^h, u^h)\).

Let the nonempty ball \(B_h \subset \Sigma^h \times V^h\), centered at \((\sigma, u)\), be defined by

\[
B_h = \left\{(\tau^h, v^h) \in \Sigma^h \times V^h : \|\sigma - \tau^h\|_0, \|\nabla(u - v^h)\|_0 \leq Ch^{3/2}\right\}.
\] (2.4.25)

To show Step 2 of the algorithm converges, we construct a mapping \(\Phi : B_h \rightarrow \Sigma^h \times V^h\) such that \(\Phi(\sigma_i^h, u_i^h) = (\sigma_{i+1}^h, u_{i+1}^h)\) where \((\sigma_{i+1}^h, u_{i+1}^h)\) satisfies

\[
(\sigma_{i+1}^h, \tau^h) + \chi B^h(u_i^h, \sigma_{i+1}^h, \tau^h) + \lambda (g_a(\sigma_{i+1}^h, \nabla u_i^h), \tau^h) - 2\alpha (D(u_{i+1}^h), \tau^h)
\]

\[
= - \left(\lambda - \bar{\lambda}\right) B^h(u_i^h, \sigma_i^h, \tau^h) - \left(\lambda - \bar{\lambda}\right) (g_a(\sigma_i^h, \nabla u_i^h), \tau^h) \quad \forall \tau^h \in \Sigma^h,
\] (2.4.26)

\[
(\sigma_{i+1}^h, D(v^h)) + 2(1 - \alpha)(D(u_{i+1}^h), D(v^h)) = (f, v^h) \quad \forall v^h \in V^h.
\] (2.4.27)

The proof is similar in structure to that in [8] for a different iteration operator \(\Phi\) and consists of three parts,

1. Show \(\Phi\) is well defined and bounded on bounded sets.
2. Show \(\Phi\) is continuous.
3. Show that \(\Phi(B_h) \subset B_h\).

Then Schauder’s fixed-point theorem guarantees the existence of a fixed point \((\sigma^h, u^h) \in B_h\) of \(\Phi\) satisfying

\[
\|\sigma - \sigma^h\|_0 + \|\nabla(u - u^h)\|_0 \leq Ch^{3/2}.
\]

We begin with some preliminary bounds.
Lemma 2.4.2 Let $u_h^0$ satisfy

$$\|\nabla(u - u_h^0)\|_0 \leq C_0 h^{3/2}$$

for some constant $C_0$. Then the quantities $\|u_h^0\|_{\infty}$ and $\|\nabla u_h^0\|_{\infty}$ are bounded.

Proof: Note that if $\tilde{u}_h^0 \in V^h$ is defined by

$$(\nabla(u - \tilde{u}_h^0), \nabla v^h) = 0 \quad \forall v^h \in V^h,$$ (2.4.28)

then standard approximation results ([12]) imply that

$$\|\nabla(u - \tilde{u}_h^0)\|_0 \leq C h^2 \|u\|_3,$$ (2.4.29)

and

$$\|\nabla(u - \tilde{u}_h^0)\|_{1,4} \leq C h \|u\|_{2,4},$$ (2.4.30)

for some constant $C$. Using (2.3.16), (2.4.29), (2.4.30) and the embedding theorem of $W^{1,4}$ in $L^\infty$, $H^1$ in $L^4$, we have

$$\|u_h^0\|_{\infty} \leq \|u\|_{\infty} + \|u - \tilde{u}_h^0\|_{\infty} + \|\tilde{u}_h^0 - u_h^0\|_{\infty}$$

$$\leq M + C \left[ h \|u\|_{2,4} + h^{-1/2} \|\nabla(u - u_h^0)\|_0 \right]$$

$$\leq M + C \left[ Mh + Mh^{3/2} + C_0 h \right].$$ (2.4.31)

From (2.4.29), (2.4.10), (2.4.8), and [12] eq. (4.4.29), we have

$$\|\nabla u_h^0\|_{\infty} \leq \|\nabla u\|_{\infty} + \|\nabla(u - \tilde{u}_h^0)\|_{\infty} + \|\nabla(\tilde{u}_h^0 - u_h^0)\|_{\infty}$$

$$\leq C_M \|\nabla u\|_{2} + \|u - \tilde{u}_h^0\|_{1,\infty} + Ch^{-1} \|\nabla(\tilde{u}_h^0 - u_h^0)\|_0$$

$$\leq C_M \|\nabla u\|_{2} + C \|u\|_{2,2} + Ch^{-1} \left( \|\nabla(u - \tilde{u}_h^0)\|_0 + \|\nabla(u - u_h^0)\|_0 \right)$$

$$\leq C_M M + C \|u\|_{2} + Ch^{-1} \left( Ch^2 M + C_0 h^{3/2} \right)$$

$$\leq C \left[ M + Mh + C_0 h^{1/2} \right].$$ (2.4.32)
Note that we can write (2.4.26)-(2.4.27) as
\[ \tilde{A}(u^h_i, (\sigma^h_{i+1}, u^h_{i+1})), (\tau^h, v^h)) + \tilde{A}(u^h_i, (\sigma^h_{i}, u^h_{i}), (\tau^h, v^h)) = F((\sigma^h_{i}, u^h_{i}), (\tau^h, v^h)), \quad (2.4.33) \]
where

\[
\tilde{A}(u^h_i, (\sigma^h_{i+1}, u^h_{i+1})), (\tau^h, v^h)) = (\sigma^h_{i+1}, \tau^h) + \tilde{\lambda}(g_a(\sigma^h_{i+1}, \nabla u^h_{i}), \tau^h) - 2\alpha(D(u^h_{i+1}), \tau^h) \\
+ 2\alpha(\sigma^h_{i+1}, D(v^h)) + 4\alpha(1 - \alpha)(D(u^h_{i+1}), D(v^h)),
\]
\[
B^h(u^h_i, \sigma^h_{i+1}, \tau^h) = ((u^h_i \cdot \nabla)\sigma^h_{i+1}, \tau^h)_h + \frac{1}{2}(\nabla \cdot u^h_i \sigma^h_{i+1}, \tau^h) \\
+ (\sigma^h_{i+1} - \sigma^h_{i+1}, \tau^h^+)_h, u^h_i,
\]
\[
F((\sigma^h_{i}, u^h_{i}), (\tau^h, v^h)) = 2\alpha(f, v^h) - (\lambda - \tilde{\lambda})B^h(u^h_i, \sigma^h_{i}, \tau^h) \\
- (\lambda - \tilde{\lambda})(g_a(\sigma^h_{i}, \nabla u^h_i), \tau^h).
\]

Now we will discuss the three properties for \( \Phi \). For the proof of the three properties for all \( i = 0, 1, \ldots, \) it is sufficient to show each property holds for \( i = 0 \).

**Φ is well-defined and bounded on bounded sets**

**Lemma 2.4.3** Let \( \tilde{M} = \|\nabla u^h_0\|_\infty \). Note that, from (2.4.32), \( M \approx CM \) for \( h \) sufficiently small. Assume \( \tilde{M} \) satisfies
\[ 1 - 4\tilde{\lambda}\tilde{M} > 0. \quad (2.4.34) \]
Then \( Φ \) is well-defined and bounded on bounded sets.

**Proof:** \( \tilde{A} \) is coercive. We have

\[
\tilde{A}(u^h_0, (\sigma^h_1, u^h_1), (\sigma^h_1, u^h_1)) = \|\sigma^h_1\|^2_0 + \tilde{\lambda}(g_a(\sigma^h_1, \nabla u^h_0, \sigma^h_1) + 4\alpha(1 - \alpha)\|D(u^h_0)\|^2_0 \\
\geq \|\sigma^h_1\|^2_0 - 4\tilde{\lambda}\|\nabla u^h_0\|_\infty \|\sigma^h_1\|^2_0 + 4\alpha(1 - \alpha)\|D(u^h_0)\|^2_0 \\
\geq \|\sigma^h_1\|^2_0 - 4\tilde{\lambda}\tilde{M}\|\sigma^h_1\|^2_0 + 4\alpha(1 - \alpha)\|D(u^h_0)\|^2_0 \\
\geq (1 - 4\tilde{\lambda}\tilde{M})\|\sigma^h_1\|^2_0 + 4\alpha(1 - \alpha)\|D(u^h_0)\|^2_0 \\
\geq C\|\sigma^h_1, u^h_1\|^2_0 \Sigma^h \times \mathbb{V}_h \quad (2.4.35)
\]
if \( 1 - 4\tilde{\lambda}\tilde{M} > 0. \)
\( \lambda B^h \) is coercive. From (2.3.9) we have
\[
\lambda B^h(u_0^h, \sigma_0^h, \sigma_1^h) = \frac{1}{2} \lambda \langle (\sigma_1^+ - \sigma_1^-) \rangle_{h,u_0^h} \geq 0. \tag{2.4.36}
\]
The coercivity of \( \Phi \) follows directly from (2.4.35) and (2.4.36).

\( \Phi \) is bounded on bounded sets.

We have
\[
F((\sigma_0^h, u_0^h), (\sigma_1^h, u_1^h)) \leq 2\alpha |(f, u_1^h)| + (\lambda - \lambda)|B^h(u_0^h, \sigma_0^h, \sigma_1^h)|
+ (\lambda - \lambda)|(g_a(\sigma_0^h, \nabla u_0^h), \sigma_1^h)|
\]
with
\[
|(f, u_1^h)| \leq \|f\|_{1}\|u_1^h\|_1 \tag{2.4.37}
\]
and
\[
|B^h(u_0^h, \sigma_0^h, \sigma_1^h)| = |(u_0^h \cdot \nabla)\sigma_0^h, \sigma_1^h| + \frac{1}{2}|(\nabla \cdot u_0^h)(\sigma_0^h, \sigma_1^h) + (\sigma_0^+ - \sigma_0^-, \sigma_1^+)|_{h,u_0^h}
\leq \|u_0^h \cdot \nabla\|_{0,h} \|\sigma_0^h\|_0 \|\sigma_1^h\|_0 + \frac{1}{2}\|\nabla \cdot u_0^h\|_{0,h} \|\sigma_0^h\|_0 \|\sigma_1^h\|_0
+ \sqrt{2} \|u_0^h\|_{0,1} \|\sigma_0^h\|_{0,1} \|\sigma_1^h\|_0
\leq C\|u_0^h\|_{0} \|\nabla\sigma_0^h\|_{0,2} \|\sigma_0^h\|_0 + C\|\nabla u_0^h\|_{0} \|\sigma_0^h\|_{\infty} \|\sigma_1^h\|_0
+ \|u_0^h\|_{\infty} \|\sigma_0^h\|_{0} \|\sigma_1^h\|_0
\leq C h^{-1} \left( h^{-1} \|u_0^h\|_0 \|\sigma_0^h\|_0 + \|\nabla u_0^h\|_0 \|\sigma_0^h\|_0 \right) \|\sigma_1^h\|_0. \tag{2.4.38}
\]
The four terms in \( (g_a(\sigma_0^h, \nabla u_0^h), \sigma_1^h) \) are all of the form
\[
|(\sigma_0^h \nabla u_0^h, \sigma_1^h)| \leq \|\sigma_0^h \nabla u_0^h\|_0 \|\sigma_1^h\|_0 \leq C \|\sigma_0^h\|_{\infty} \|\nabla u_0^h\|_0 \|\sigma_1^h\|_0
\leq C h^{-1} (\|\nabla u_0^h\|_0 \|\sigma_0^h\|_0) \|\sigma_1^h\|_0.
\]
Therefore,
\[
|(g_a(\sigma_0^h, \nabla u_0^h), \sigma_1^h)| \leq C h^{-1} (\|\nabla u_0^h\|_0 \|\sigma_0^h\|_0) \|\sigma_1^h\|_0. \tag{2.4.39}
\]
Combining (2.4.33), (2.4.35), (2.4.36), (2.4.37)-(2.4.39) we have that
\[
\|(\sigma_1^h, u_1^h)\|_{\Sigma^h \times V^h} \leq C \left( \|f\|_{-1} + h^{-1} \left( h^{-1} \|u_0^h\|_0 \|\sigma_0^h\|_0 + \|\nabla u_0^h\|_0 \|\sigma_0^h\|_0 \right) \right). \tag{2.4.40}
\]
Thus, $\Phi$ is bounded on bounded sets.

$\Phi$ is continuous.

**Lemma 2.4.4** Assume $1 - 4\tilde{\lambda}\tilde{M} > 0$. Then $\Phi$ is continuous on $\Sigma^h \times V^h$.

**Proof:** To show $\Phi$ is continuous, we want to show that, for $h$ small enough, if $\Phi_1(\sigma^h_0, u^h_0) = (\sigma^h_1, u^h_1)$ and $\Phi_1(\tau^h_0, v^h_0) = (\tau^h_1, v^h_1)$, then

$$\| (\sigma^h_1 - \tau^h_1, u^h_1 - v^h_1) \|_{\Sigma^h \times V^h} \leq C\eta(h, \sigma^h_0, \tau^h_0, u^h_0, v^h_0),$$

where

$$\lim_{(\tau^h_0, v^h_0) \to (\sigma^h_0, u^h_0)} \eta(h, \sigma^h_0, \tau^h_0, u^h_0, v^h_0) = 0.$$  

For ease of notation we suppress the superscript $h$ on the variables $\sigma^h, \tau^h, u^h, v^h$ in the remainder of this proof. Now $\Phi_1(\sigma^0, u^0) = (\sigma^1, u^1)$ implies

$$\tilde{A}(u^0, (\sigma^1, u^1), (t, v)) + \tilde{\lambda}B^h(u^0, \sigma^1, t) = F((\sigma^0, u^0), (t, v)), \quad \forall (t, v) \in \Sigma^h \times V^h \tag{2.4.41}$$

and $\Phi_1(\tau^0, v^0) = (\tau^1, v^1)$ implies

$$\tilde{A}(v^0, (\tau^1, v^1), (t, v)) + \tilde{\lambda}B^h(v^0, \tau^1, t) = F((\tau^0, v^0), (t, v)), \quad \forall (t, v) \in \Sigma^h \times V^h. \tag{2.4.42}$$

Subtracting (2.4.42) from (2.4.41) and rearranging we obtain

$$\tilde{A}(u^0, (\sigma^1 - \tau^1, u^1 - v^1), (t, v)) + \tilde{\lambda}B^h(u^0, \sigma^1 - \tau^1, t)$$

$$= - (\lambda - \tilde{\lambda}) \left[ B^h(u^0, \sigma^0, t) - B^h(v^0, \tau^0, t) \right]$$

$$- (\lambda - \tilde{\lambda}) \left[ (g_a(\sigma^0, \nabla u_0), t) - (g_a(\tau^0, \nabla v_0), t) \right]$$

$$- \tilde{\lambda} \left[ B^h(u^0, \tau^1, t) - B^h(v^0, \tau^1, t) \right]$$

$$- \tilde{\lambda} (g_a(\tau^1, \nabla(u^0 - v^0)), t), \quad \forall (t, v) \in \Sigma^h \times V^h. \tag{2.4.43}$$
Let \( t = \sigma^1 - \tau^1 \) and \( v = u^1 - v^1 \). Then (2.4.43), (2.4.35) and the assumption on \( 1 - 4\tilde{\lambda}\tilde{M} \) implies

\[
\|(\sigma^1 - \tau^1, u^1 - v^1)\|_{\Sigma^h \times V_h}^2 \leq - (\lambda - \tilde{\lambda}) \left[ B^h(u^0, \sigma^0, \sigma^1 - \tau^1) - B^h(v^0, \tau^0, \sigma^1 - \tau^1) \right] \\
- (\lambda - \tilde{\lambda}) \left[ (g_a(\sigma^0, \nabla u_0), \sigma^1 - \tau^1) - (g_a(\tau^0, \nabla v_0), \sigma^1 - \tau^1) \right] \\
- \tilde{\lambda} \left[ B^h(u^0, \tau^1, \sigma^1 - \tau^1) - B^h(v^0, \tau^1, \sigma^1 - \tau^1) \right] \\
- \tilde{\lambda} (g_a(\tau^1, \nabla (u^0 - v^0)), \sigma^1 - \tau^1). \tag{2.4.44}
\]

For notational simplicity, we will let

\[
\hat{\sigma}_0 = \sigma^0 - \tau^0, \quad \hat{\sigma}_1 = \sigma^1 - \tau^1, \quad \hat{u}_0 = u^0 - v^0, \quad \hat{u}_1 = u^1 - v^1.
\]

Then (2.4.44) can be written as

\[
\| (\hat{\sigma}_1, \hat{u}_1) \|_{\Sigma^h \times V_h}^2 \leq - (\lambda - \tilde{\lambda}) \left[ B^h(u^0, \hat{\sigma}_0, \hat{\sigma}_1) + B^h(u^0, \tau^0, \hat{\sigma}_1) - B^h(v^0, \tau^0, \hat{\sigma}_1) \right] \\
- (\lambda - \tilde{\lambda}) \left[ (g_a(\sigma^0, \nabla u_0), \hat{\sigma}_1) - (g_a(\tau^0, \nabla v_0), \hat{\sigma}_1) \right] \\
- \tilde{\lambda} \left[ B^h(u^0, \tau^1, \hat{\sigma}_1) - B^h(v^0, \tau^1, \hat{\sigma}_1) \right] - \tilde{\lambda} (g_a(\tau^1, \nabla \hat{u}_0), \hat{\sigma}_1). \tag{2.4.45}
\]

Our goal is to show that each of the terms on the right-hand side of go to zero as \((\tau^0, v^0)\) goes to \((\sigma^0, u^0)\), for a fixed \( h \). This is accomplished when these terms can be shown to be dominated by products of bounded norms with \( \|\hat{\sigma}_0\|_0 \) or \( \|\hat{u}_0\|_0 \). We remark that Lemma 2.4.3 implies that the quantities \( \|\sigma^1\|_0, \|\tau^1\|_0, \) and \( \|\hat{\sigma}_1\|_0 \) are all bounded. Using (2.3.7), the first term on the RHS of (2.4.45) is written as

\[
B^h(u^0, \hat{\sigma}_0, \hat{\sigma}_1) = \left( u^0 \cdot \nabla \hat{\sigma}_0, \hat{\sigma}_1 \right)_h + \frac{1}{2} (\nabla \cdot u^0 \hat{\sigma}_0, \hat{\sigma}_1) + \langle \hat{\sigma}_0^+ - \hat{\sigma}_0^-, \hat{\sigma}_1^+ \rangle_{h,u^0}.
\]

With

\[
\|(u^0 \cdot \nabla \hat{\sigma}_0, \hat{\sigma}_1)_h \| \leq C\|u^0\|_\infty \|\nabla \hat{\sigma}_0\|_{0,2,h} \|\hat{\sigma}_1\|_0 \\
\leq (Ch^{-1}\|u^0\|_0)(Ch^{-1}\|\hat{\sigma}_0\|_0)\|\hat{\sigma}_1\|_0 \\
\leq Ch^{-2}\|u^0\|_0\|\hat{\sigma}_0\|_0\|\hat{\sigma}_1\|_0, \tag{2.4.46}
\]

30
\[(\nabla \cdot \mathbf{u}^0 \hat{\sigma}_0, \hat{\sigma}_1) \leq C\| \nabla \mathbf{u}_0 \|_0 \| \hat{\sigma}_0 \|_0 \| \hat{\sigma}_1 \|_0 \]
\[\leq (Ch^{-3/2}\| \mathbf{u}^0 \|_0)(Ch^{-1/2}\| \hat{\sigma}_0 \|_0)\| \hat{\sigma}_1 \|_0 \]
\[\leq Ch^{-2}\| \mathbf{u}^0 \|_0 \| \hat{\sigma}_0 \|_0 \| \hat{\sigma}_1 \|_0 , \quad (2.4.47)\]

and

\[|\langle \hat{\sigma}^+_0 - \hat{\sigma}^-_0, \hat{\sigma}^+_1 \rangle_{h, \mathbf{u}^0}| \leq C\| \mathbf{u}^0 \|_\infty \| \hat{\sigma}_0 \|_{0, \Gamma^h} \| \hat{\sigma}_1 \|_{0, \Gamma^h} \]
\[\leq (Ch^{-1}\| \mathbf{u}^0 \|_0)(Ch^{-1/2}\| \hat{\sigma}_0 \|_0)(Ch^{-1/2}\| \hat{\sigma}_1 \|_0) \]
\[\leq Ch^{-2}\| \mathbf{u}^0 \|_0 \| \hat{\sigma}_0 \|_0 \| \hat{\sigma}_1 \|_0 , \]

we have

\[(\lambda - \bar{\lambda})|B^h(\mathbf{u}^0, \hat{\sigma}_0, \hat{\sigma}_1)| \leq (C(\lambda - \bar{\lambda})h^{-2}\| \hat{\sigma}_1 \|_0)\| \mathbf{u}^0 \|_0 \| \hat{\sigma}_0 \|_0 = C\eta_1(h, \sigma^0, \mathbf{r}^0, \mathbf{u}^0, \mathbf{v}^0). \quad (2.4.48)\]

Next, as in (2.4.46) and (2.4.47),

\[B^h(\mathbf{u}^0, \mathbf{r}^0, \hat{\sigma}_1) - B^h(\mathbf{v}^0, \mathbf{r}^0, \hat{\sigma}_1) = (\hat{\mathbf{u}}_0 \cdot \nabla \mathbf{r}^0, \hat{\sigma}_1)_h + \frac{1}{2}(\nabla \cdot \hat{\mathbf{u}}_0 \mathbf{r}^0, \hat{\sigma}_1)_h \]
\[+ \langle \mathbf{r}^0 + - \mathbf{r}^0 -, \hat{\sigma}^+_1 \rangle_{h, \mathbf{u}^0} - \langle \mathbf{r}^0 + - \mathbf{r}^0 -, \hat{\sigma}^+_1 \rangle_{h, \mathbf{v}^0} \]
\[\leq Ch^{-2}\| \hat{\mathbf{u}}_0 \|_0 \| \mathbf{r}^0 \|_0 \| \hat{\sigma}_1 \|_0 \]
\[+ \langle \mathbf{r}^0 + - \mathbf{r}^0 -, \hat{\sigma}^+_1 \rangle_{h, \mathbf{u}^0} - \langle \mathbf{r}^0 + - \mathbf{r}^0 -, \hat{\sigma}^+_1 \rangle_{h, \mathbf{v}^0}. \quad (2.4.49)\]

The jump terms in (2.4.49) are bounded as in [8]. Let \( \mathbf{u}^0 \) be fixed and let \( \epsilon > 0 \) (we can assume that \( \epsilon \) is bounded above by 1). With \( \Gamma^h \) as in (2.3.5), define \( \theta^- \) as

\[\theta^- = \{ x \in \Gamma^h : |\mathbf{u}^h_0 \cdot \mathbf{n}(x)| \leq \epsilon \} \]

i.e., the parts of the interior edges of the triangulation with the magnitude of the flow normal to the edges less than or equal to \( \epsilon \) and define \( \theta^+ = \Gamma^h \setminus \theta^- \). Also define the norm

\[\| \sigma \|_{0,2,\theta^\pm}^2 = \sum_{K \in \Gamma^h} \int_{\partial K \cap \theta^\pm} (\sigma : \sigma) \, ds. \]
From the definition of the inner product (2.3.6) we have

\[
\left| \langle \tau_0^+ - \tau_0^-, \hat{\sigma}_1^+ \rangle_{h, u^0} - \langle \tau_0^+ - \tau_0^-, \hat{\sigma}_1^+ \rangle_{h, v^0} \right| \\
= \left| \sum_{K \in T_h} \int_{\partial K^-(u^0)} ((\tau_0^+(u^0) - \tau_0^-(u^0)) : \hat{\sigma}_1^+(u^0)) |u^0 \cdot n| \, ds \right. \\
- \left. \sum_{K \in T_h} \int_{\partial K^-(v^0)} ((\tau_0^+(v^0) - \tau_0^-(v^0)) : \hat{\sigma}_1^+(v^0)) |v^0 \cdot n| \, ds \right|
\]

\[
\leq \left| \sum_{K \in T_h} \int_{\theta^+} \left( ((\tau_0^+(u^0) - \tau_0^-(u^0)) : \hat{\sigma}_1^+(u^0)) |u^0 \cdot n| \\
- ((\tau_0^+(v^0) - \tau_0^-(v^0)) : \hat{\sigma}_1^+(v^0)) |v^0 \cdot n| \right) \, ds \right|
\]

\[
+ \left| \sum_{K \in T_h} \int_{\theta^-} \left( ((\tau_0^+(u^0) - \tau_0^-(u^0)) : \hat{\sigma}_1^+(u^0)) |u^0 \cdot n| \\
- ((\tau_0^+(v^0) - \tau_0^-(v^0)) : \hat{\sigma}_1^+(v^0)) |v^0 \cdot n| \right) \, ds \right|.
\]

(2.4.50)

Let \( v^0 \) be such that \( \| \hat{u}_0 \|_{\infty} = \| u^0 - v^0 \|_{\infty} \leq \epsilon/2 \). Then on the portion of \( \partial K \) in \( \theta^+ \), we have that \( u^0 \cdot n \) and \( v^0 \cdot n \) are of the same sign. Thus \( \tau_0^+(u^0) = \tau_0^+(v^0) \) and \( \hat{\sigma}_1^+(u^0) = \hat{\sigma}_1^+(v^0) \) and we have

\[
\left| \sum_{K \in T_h} \int_{\theta^+} \left( ((\tau_0^+(u^0) - \tau_0^-(u^0)) : \hat{\sigma}_1^+(u^0)) |u^0 \cdot n| \\
- ((\tau_0^+(v^0) - \tau_0^-(v^0)) : \hat{\sigma}_1^+(v^0)) |v^0 \cdot n| \right) \, ds \right|
\]

\[
\leq \left| \sum_{K \in T_h} \int_{\theta^+} \left( ((\tau_0^+(u^0) - \tau_0^-(u^0)) : \hat{\sigma}_1^+(u^0)) |u^0 - v^0| \cdot n \cdot n \right) \, ds \right|
\]

\[
\leq C \| u^0 \cdot n - v^0 \cdot n \|_{0, \infty, \theta^+} \| \tau_0 \|_{0, 2, \theta^+} \| \hat{\sigma}_1 \|_{0, 2, \theta^+}
\]

\[
\leq C h^{-1/2} \| u^0 - v^0 \|_{\infty} (C h^{-1/2} \| \tau_0 \|_0) (C h^{-1/2} \| \hat{\sigma}_1 \|_0)
\]

\[
\leq C h^{-3/2} \| \hat{u}_0 \|_{\infty} \| \tau_0 \|_0 \| \hat{\sigma}_1 \|_0
\]

\[
\leq C h^{-5/2} \| \hat{u}_0 \|_0 \| \tau_0 \|_0 \| \hat{\sigma}_1 \|_0.
\]

(2.4.51)
Now, by the definition of $\theta^-$ and the condition on $v^0$, we have that $|u^0 \cdot n| \leq \epsilon, |v^0 \cdot n| \leq \epsilon/2$ and thus

\[
\left| \sum_{K\in\mathcal{T}^h} \int_{\partial K^-} \left( ((\tau_0^+(u^0) - \tau_0^-(u^0)) : \hat{\sigma}_1^- (u^0)) |u^0 \cdot n| \\
- ((\tau_0^+(v^0) - \tau_0^-(v^0)) : \hat{\sigma}_1^+ (v^0)) |v^0 \cdot n| \right) \right| ds
\leq C\|\tau^0\|_{0,2,\theta^-} \|\hat{\sigma}_1\|_{0,2,\theta^-}(\epsilon)
\leq \epsilon Ch^{-1}\|\tau^0\|_0 \|\hat{\sigma}_1\|_0.
\]

(2.4.52)

Combining (2.4.49), (2.4.51), and (2.4.52) from (2.4.50) we have that

\[(\lambda - \bar{\lambda})|B^h(u^0, \tau^0, \hat{\sigma}_1) - B^h(v^0, \tau^0, \hat{\sigma}_1)| \leq C\|\hat{\sigma}_1\|_0\eta_2(h, \sigma^0, \tau^0, u^0, v^0)
\]

(2.4.53)

where

$$
\eta_2(h, \sigma^0, \tau^0, u^0, v^0) = (h^{-2} + h^{-5/2})\|\hat{u}_0\|_0 \|\tau^0\|_0.
$$

The second term in (2.4.45) can be written as

\[
(g_a(\sigma^0, \nabla u^0), \hat{\sigma}_1) - (g_a(\tau^0, \nabla v^0), \hat{\sigma}_1)
\]

\[
= (g_a(\sigma_0, \nabla u^0), \hat{\sigma}_1) + (g_a(\tau^0, \nabla \hat{u}_0), \hat{\sigma}_1)
\leq C(\|\sigma_0\|_{0,4} \|\nabla u^0\|_{0,4} \|\hat{\sigma}_1\|_0 + \|\tau^0\|_{\infty} \|\nabla \hat{u}_0\|_0 \|\hat{\sigma}_1\|_0)
\leq C\left((h^{-1/2} \|\sigma_0\|_0)(h^{-3/2} \|u^0\|_0) \|\hat{\sigma}_1\|_0 + (h^{-1} \|\tau^0\|_0)(h^{-1} \|\hat{u}_0\|_0) \|\hat{\sigma}_1\|_0 \right)
\leq C\|\hat{\sigma}_1\|_0 \eta_3(h, \sigma^0, \tau^0, u^0, v^0),
\]

(2.4.54)

where

$$
\eta_3(h, \sigma^0, \tau^0, u^0, v^0) = h^{-2} \left(\|\sigma_0\|_0 \|u^0\|_0 + \|\tau^0\|_0 \|\hat{u}_0\|_0 \right).
$$

The $\bar{\lambda}$ term in (2.4.45) can be estimated in a similar manner as (2.4.53) to obtain

\[
\bar{\lambda} |B^h(u^0, \tau^1, \hat{\sigma}_1) - B^h(v^0, \tau^1, \hat{\sigma}_1)| \leq C\|\tau^1\|_0 \|\hat{\sigma}_1\|_0 \eta_4(h, \sigma^0, \tau^0, u^0, v^0)
\]

(2.4.55)

where

$$
\eta_4(h, \sigma^0, \tau^0, u^0, v^0) = (h^{-2} + h^{-5/2})\|\hat{u}_0\|_0
$$

33
Finally the last term on the right-hand side of (2.4.45) is bounded
\[
(g_a(\tau^0, \nabla \hat{u}_0), \hat{\sigma}_1) \leq C \|\tau^0\|_\infty \|\nabla \hat{u}_0\|_0 \|\hat{\sigma}_1\|_0
\]
\[
\leq C \|\tau^0\|_0 \|\nabla \hat{u}_0\|_0 \|\hat{\sigma}_1\|_0 \eta_5(h, \sigma^0, \tau^0, u^0, v^0) \quad (2.4.56)
\]
with \(\eta_5(h, \sigma^0, \tau^0, u^0, v^0) = h^{-2}\|\hat{u}_0\|_0\). Thus (2.4.48) and (2.4.53)-(2.4.56) imply that \(\Phi\) is continuous on \(\Sigma^h \times V^h\).

\[\Phi(B_h) \subset B_h.\]

The proof that there exists a nonempty ball \(B_h\) in \(\Sigma^h \times V^h\) centered at \((\sigma, u)\) such that \(\Phi(B_h) \subset B_h\) is similar to the proof of Theorem 4.1 in [57]. Let \(\hat{\sigma}^h \in \Sigma^h\) be the orthogonal projection of \(\sigma\) on \(T^h\) in \(\Sigma\), and \(\hat{u}^h \in V^h\) given by (2.4.28). Then standard approximation results ([12]) imply that there exists a constant \(C'\) such that
\[
\|\nabla(u - \hat{u}^h)\|_0 \leq C' h^2 \|u\|_3, \quad (2.4.57)
\]
\[
\|\sigma - \hat{\sigma}^h\|_0 + h\|\nabla(\sigma - \hat{\sigma}^h)\|_0 \leq C' h^2 \|\sigma\|_2, \quad (2.4.58)
\]
\[
\|\sigma - \hat{\sigma}^h\|_{0,1^h} \leq C' h^{3/2} \|\sigma\|_2. \quad (2.4.59)
\]
Assuming \(h < 1\) and \(C'M \leq \bar{C}/4\) for some constant \(\bar{C}\), (2.4.7) guarantees that
\[
\|\sigma - \hat{\sigma}^h\|_0 + \|\nabla(u - \hat{u}^h)\|_0 \leq \frac{\bar{C}}{2} h^2 \leq \frac{\bar{C}}{2} h^{3/2}, \quad (2.4.60)
\]
and thus \((\hat{\sigma}^h, \hat{u}^h) \in B_h\).

**Lemma 2.4.5** There exist constants \(C\) and \(\delta\) such that if \(1 - 4\tilde{\lambda}\tilde{M} > 0, \sqrt{C}h^{1/2} \leq M,\)
\[
M \leq \min\{\bar{C}, \bar{C}/(4C')\}, \quad (2.4.61)
\]
and
\[
\frac{C}{\delta} \left( \frac{2(2\tilde{\lambda}M + \tilde{\lambda}M)^2}{(1 - 4\tilde{\lambda}M)} + \frac{h}{2\alpha(1-\alpha)} + \tilde{\lambda}M \right)^{1/2} \leq 1, \quad (2.4.62)
\]
then \(\Phi(B_h) \subset B_h\).
**Proof:** To prove $\Phi(B_h) \subset B_h$, we show that, given $(\sigma_0^h, u_0^h) \in B_h$, $\Phi(\sigma_0^h, u_0^h) = (\sigma_1^h, u_1^h) \in B_h$. Let $(\sigma_0^h, u_0^h) \in B_h$. Then,

$$\|\sigma - \sigma_0^h\|_0 + \|\nabla(u - u_0^h)\|_0 \leq \overline{C}h^{3/2},$$

and Lemma 2.4.2 implies that

$$\|u_0^h\|_\infty \leq C \left[ M + Mh + Mh^{3/2} + \overline{C}h \right]$$

and

$$\|\nabla u_0^h\|_\infty \leq C \left[ M + Mh + \overline{C}h^{1/2} \right].$$

The assumption of $\overline{C}h^{1/2} \leq M$ then implies the existence of a constant $C_u$ such that

$$\|u_0^h\|_\infty \leq C_u M \quad \text{and} \quad \|\nabla u_0^h\|_\infty \leq C_u M.$$

Since $(\sigma, u, p)$ satisfies (2.2.1)-(2.2.3), we have

$$A((\sigma, u), (\tau^h, v^h)) + \lambda(u \cdot \nabla \sigma, \tau^h) + \lambda(g_a(\sigma, \nabla u), \tau^h) - (p, \nabla \cdot v^h) = 2\alpha(f,v^h), \quad \forall (\tau^h, v^h) \in \Sigma^h \times V^h. \quad (2.4.63)$$

and, since $\nabla \cdot u = 0$ and $\sigma$ is continuous

$$B_h(u, \sigma, \tau^h) = ((u \cdot \nabla)\sigma, \tau^h)_h = ((u \cdot \nabla)\sigma, \tau^h).$$

Also, $B_h(\cdot, \sigma, \cdot)$ can be treated as a bilinear form (since the jump term vanishes when the second argument in $B_h(\cdot, \cdot, \cdot)$ is continuous). Now $(\sigma_1^h, u_1^h) = \Phi(\sigma_0^h, u_0^h)$ satisfies the relation

$$A((\sigma_1^h, u_1^h), (\tau^h, v^h)) + \lambda_h B_h(u_0^h, \sigma_1^h, \tau^h) + \lambda_h(g_a(\sigma_1^h, \nabla u_0^h), \tau^h)$$

$$- (\lambda - \overline{\lambda})(g_a(\sigma_0^h, \nabla u_0^h), \tau^h), \quad \forall (\tau^h, v^h) \in \Sigma^h \times V^h. \quad (2.4.64)$$
Subtracting (2.4.64) from (2.4.63) yields

\[ A((\sigma - \sigma_1^h, u - u_1^h), (\tau^h, v^h)) \]
\[ + \overline{\lambda} \left( B^h(u, \sigma, \tau^h) - B^h(u_0^h, \sigma_1^h, \tau^h) \right) + (\lambda - \overline{\lambda})B^h(u, \sigma, \tau^h) \]
\[ + \overline{\lambda} \left( (g_a(\sigma, \nabla u), \tau^h) - (g_a(\sigma_1^h, \nabla u_0^h), \tau^h) \right) \]
\[ + (\lambda - \overline{\lambda})(g_a(\sigma, \nabla u), \tau^h) - (p, \nabla \cdot v^h) \]
\[ = (\lambda - \overline{\lambda})B^h(u_0^h, \sigma_0^h, \tau^h) + (\lambda - \overline{\lambda})(g_a(\sigma_0^h, \nabla u_0^h), \tau^h), \quad \forall (\tau^h, v^h) \in \Sigma^h \times V^h. \]

i.e.,

\[ A((\sigma - \check{\sigma}^h, u - \check{u}^h), (\tau^h, v^h)) + A((\check{\sigma}^h - \sigma_1^h, \check{u}^h - u_1^h), (\tau^h, v^h)) \]
\[ + \overline{\lambda} \left( B^h(u, \sigma, \tau^h) + B^h(u_0^h, \check{\sigma}^h - \sigma_1^h, \tau^h) - B^h(u_0^h, \check{\sigma}^h, \tau^h) \right) + (\lambda - \overline{\lambda})B^h(u, \sigma, \tau^h) \]
\[ + \overline{\lambda} \left( (g_a(\sigma, \nabla u), \tau^h) + g_a(\check{\sigma}^h - \sigma_1^h, \nabla u_0^h), \tau^h) - (g_a(\check{\sigma}^h, \nabla u_0^h), \tau^h) \right) \]
\[ + (\lambda - \overline{\lambda})(g_a(\sigma, \nabla u), \tau^h) - (p, \nabla \cdot v^h) \]
\[ = (\lambda - \overline{\lambda})B^h(u_0^h, \sigma_0^h, \tau^h) + (\lambda - \overline{\lambda})(g_a(\sigma_0^h, \nabla u_0^h), \tau^h), \quad \forall (\tau^h, v^h) \in \Sigma^h \times V^h. \quad (2.4.65) \]

Rearranging (2.4.65) we obtain

\[ A((\check{\sigma}^h - \sigma_1^h, \check{u}^h - u_1^h), (\tau^h, v^h)) \]
\[ + \overline{\lambda}B^h(u_0^h, \check{\sigma}^h - \sigma_1^h, \tau^h) + \overline{\lambda}(g_a(\check{\sigma}^h - \sigma_1^h, \nabla u_0^h), \tau^h) \]
\[ = -A((\sigma - \check{\sigma}^h, u - \check{u}^h), (\tau^h, v^h)) \]
\[ + \overline{\lambda} \left( B^h(u_0^h, \check{\sigma}^h, \tau^h) - B^h(u, \sigma, \tau^h) \right) \]
\[ + \overline{\lambda} \left( (g_a(\check{\sigma}^h, \nabla u_0^h), \tau^h) - (g_a(\sigma, \nabla u), \tau^h) \right) \]
\[ + (\lambda - \overline{\lambda}) \left( B^h(u_0^h, \sigma_0^h, \tau^h) - B^h(u, \sigma, \tau^h) \right) \]
\[ + (\lambda - \overline{\lambda}) \left( (g_a(\sigma_0^h, \nabla u_0^h), \tau^h) - (g_a(\sigma, \nabla u), \tau^h) \right) \]
\[ + (p - q^h, \nabla \cdot v^h), \quad \forall (\tau^h, v^h, q^h) \in \Sigma^h \times V^h \times S^h. \quad (2.4.66) \]
Let \( \tau^h = \sigma^h - \sigma^h_1 \) and \( \nu^h = \bar{u}^h - u^h_1 \). For ease of notation, let \( \hat{\sigma} = \sigma^h - \sigma^h_1 \) and \( \hat{u} = \bar{u}^h - u^h_1 \).

Using (2.3.9) and Lemma 2.4.3, from (2.4.10) and (2.4.66) we have

\[
(1 - 4\lambda \tilde{M})\|\hat{\sigma}\|^2_0 + 4\alpha(1 - \alpha)\|\nabla \hat{u}\|^2_0 + \frac{\lambda}{2} \langle (\hat{\sigma}^+ - \hat{\sigma}^-) \rangle^2_{h, u^h_0} \\
\leq \tilde{A}(u^h_0, (\hat{\sigma}, \bar{u}, (\hat{\sigma}, \bar{u})) + \lambda B^h(u^h_0, \hat{\sigma}, \hat{\sigma}) \\
= -A((\sigma - \hat{\sigma}^h, u - \bar{u}^h), (\hat{\sigma}, \bar{u})) \\
+ \lambda \bigg( B^h(u^h_0, \hat{\sigma}^h, \hat{\sigma}) - B^h(u, \sigma, \sigma) \bigg) \\
+ \lambda \bigg( (g_a(\sigma^h, \nabla u^h_0), \hat{\sigma}) - (g_a(\sigma, \nabla u), \hat{\sigma}) \bigg) \\
+ (\lambda - \lambda) \bigg( B^h(u^h_0, \sigma^h_0, \hat{\sigma}) - B^h(u, \sigma, \hat{\sigma}) \bigg) \\
+ (\lambda - \lambda) \bigg( (g_a(\sigma^h_0, \nabla u^h_0), \hat{\sigma}) - (g_a(\sigma, \nabla u), \hat{\sigma}) \bigg) \\
+ (p - q^h, \nabla \cdot \bar{u}).
\]  

(2.4.72)

To show \((\sigma^h_0, u^h_0) \in B_h\), we bound each of the terms on the right-hand side of (2.4.72).

Using the definition of \( A \) and the interpolation properties (2.4.57) and (2.4.58), we have

\[
A((\sigma - \hat{\sigma}^h, u - \bar{u}^h), (\hat{\sigma}, \bar{u})) \\
\leq \|\sigma - \hat{\sigma}^h\|_0\|\hat{\sigma}\|_0 + 2\alpha\|\nabla(u - \bar{u}^h)\|_0\|\hat{\sigma}\|_0 + 2\alpha\|\sigma - \hat{\sigma}^h\|_0\|\nabla \hat{u}\|_0 \\
+ 4\alpha(1 - \alpha)\|\nabla(u - \bar{u}^h)\|_0\|\nabla \hat{u}\|_0 \\
\leq C'Mh^2\|\hat{\sigma}\|_0 + 2\alpha C'Mh^2\|\hat{\sigma}\|_0 + 2\alpha C'Mh^2\|\nabla \hat{u}\|_0 \\
+ 4\alpha(1 - \alpha)C'Mh^2\|\nabla \hat{u}\|_0 \\
\leq C_1Mh^2\|\hat{\sigma}\|_0 + C_2Mh^2\|\nabla \hat{u}\|_0.
\]  

(2.4.73)

We add and subtract \( \sigma \) in the second term on the RHS of (2.4.72) to get

\[
\lambda \bigg( B^h(u^h_0, \hat{\sigma}^h, \hat{\sigma}) - B^h(u, \sigma, \sigma) \bigg) \leq \lambda \bigg( |B^h(u^h_0, \sigma - \hat{\sigma}^h, \hat{\sigma})| + |B^h(u - u^h_0, \sigma, \hat{\sigma})| \bigg).
\]  

(2.4.74)
Using (2.3.8), the first term in (2.4.74) may be rewritten as

\[
B^h(u_0^h, \sigma - \tilde{\sigma}^h, \tilde{\sigma}) = \frac{1}{2}((\nabla \cdot u_0^h)\tilde{\sigma}, \sigma - \tilde{\sigma}^h)_h - (u_0^h \cdot \nabla)\tilde{\sigma}, \sigma - \tilde{\sigma}^h)_h + \langle (\sigma - \tilde{\sigma}^h)^-, \tilde{\sigma}^- - \tilde{\sigma}^+ \rangle_{h,u_0^h}. \tag{2.4.75}
\]

Let \( \tilde{u} \) be the \( P_1 \) continuous interpolate of \( u \) on \( V^h \). Then

\[
\| u_0^h - \tilde{u} \|_{0,4} \leq C \| \nabla (u_0^h - \tilde{u}) \|_0 \leq C(\| \nabla (u_0^h - u) \|_0 + \| \nabla (u - \tilde{u}) \|_0) \leq C(CH^{3/2} + Mh).
\tag{2.4.76}
\]

Since \( \nabla \tilde{\sigma} \) is \( P_0 \) on each \( K \), \( (\tilde{u} \cdot \nabla)\tilde{\sigma} \) is \( P_1 \). Also, since \( \tilde{\sigma}^h \) is the orthogonal projection of \( \sigma \) on \( T^h \) in \( \Sigma \),

\[
(((\tilde{u} \cdot \nabla)\tilde{\sigma}, \sigma - \tilde{\sigma}^h)_h = 0. \tag{2.4.77}
\]

Thus, by (2.3.18), (2.4.58), (2.4.76), and (2.4.77),

\[
((u_0^h \cdot \nabla)\tilde{\sigma}, \sigma - \tilde{\sigma}^h)_h = (((u_0^h - \tilde{u}) \cdot \nabla)\tilde{\sigma}, \sigma - \tilde{\sigma}^h)_h \\
\leq C \| u_0^h - \tilde{u} \|_{0,4} \| \nabla \tilde{\sigma} \|_{0,4,h} \| \sigma - \tilde{\sigma}^h \|_0 \\
\leq C(CH^{3/2} + Mh)(h^{-3/2} \| \tilde{\sigma} \|_0)(Ch^2 \| \sigma \|_2) \\
\leq C(CH^2 + M^2h^{3/2}) \| \tilde{\sigma} \|_0. \tag{2.4.78}
\]

Since \( \nabla \cdot u = 0 \), the second term in (2.4.75) can be written as \( \frac{1}{2}((\nabla \cdot (u_0^h - u))\tilde{\sigma}, \sigma - \tilde{\sigma}^h) \) and, using (2.3.17) and (2.4.58), we have

\[
\frac{1}{2}((\nabla \cdot (u_0^h - u))\tilde{\sigma}, \sigma - \tilde{\sigma}^h) \leq C \| \nabla (u_0^h - u) \|_0 \| \tilde{\sigma} \|_\infty \| \sigma - \tilde{\sigma}^h \|_h \\
\leq C(CH^{3/2}(Ch^{-1} \| \tilde{\sigma} \|_0)(Ch^2 \| \sigma \|_2) \\
\leq CCH^5/2 \| \tilde{\sigma} \|_0. \tag{2.4.79}
\]

Using (2.4.59) the third term in (2.4.75) becomes

\[
\langle (\sigma - \tilde{\sigma}^h)^-, \tilde{\sigma}^- - \tilde{\sigma}^+ \rangle_{h,u_0^h} \leq \langle (\sigma - \tilde{\sigma}^h)^- \rangle_{h,u_0^h} \langle (\tilde{\sigma}^- - \tilde{\sigma}^+) \rangle_{h,u_0^h} \\
\leq C \| u_0^h \|_\infty^{1/2} \| \sigma - \tilde{\sigma}^h \|_{0,G^k} \langle (\tilde{\sigma}^- - \tilde{\sigma}^+) \rangle_{h,u_0^h} \tag{2.4.80}
\]

\[
\leq C(CuM)^{1/2} Mh^{3/2} \langle (\tilde{\sigma}^- - \tilde{\sigma}^+) \rangle_{h,u_0^h}. \tag{2.4.81}
\]
Thus, by (2.4.78), (2.4.79), and (2.4.81), we obtain

\[ B^h(u_0^h, \sigma - \hat{\sigma}^h, \hat{\sigma}) \leq C \left( (CMh^2 + M^2h^{3/2} + CMh^{5/2}) \left\| \hat{\sigma} \right\|_0 + (Mh)^{3/2}\left\langle \left( \sigma^- - \hat{\sigma}^+ \right) \right\rangle_{h, u_0^h} \right). \]  

(2.4.82)

For the second term in (2.4.74), we have that

\[ B^h(u - u_0^h, \sigma, \hat{\sigma}) = (\left\langle (u - u_0^h) \cdot \nabla \right\rangle + \frac{1}{2} (\nabla \cdot (u - u_0^h)\sigma, \hat{\sigma}) + 0. \]

Using the embedding of \( H^1 \) in \( L^4 \), \( H^2 \) in \( L^\infty \), \( W^{2,2} \) in \( W^{1,4} \) and the Poincaré-Friedrichs inequality ([12]) we have

\[ B^h(u - u_0^h, \sigma, \hat{\sigma}) \leq C \left( \left\| u - u_0^h \right\|_{0,4} \left\| \nabla \sigma \right\|_{0,4} \left\| \hat{\sigma} \right\|_0 + \left\| \nabla (u - u_0^h) \right\|_0 \left\| \sigma \right\|_{\infty} \left\| \hat{\sigma} \right\|_0 \right) \]

\[ \leq C \left\| \nabla (u - u_0^h) \right\|_0 \left\| \sigma \right\|_2 \left\| \hat{\sigma} \right\|_0 \]

\[ \leq CCM^3h^{3/2} \left\| \hat{\sigma} \right\|_0. \]  

(2.4.83)

Thus, combining (2.4.74), (2.4.82), and (2.4.83) we have that

\[ \bar{\lambda} \left( B^h(u_0^h, \hat{\sigma}^h, \hat{\sigma}) - B^h(u, \sigma, \hat{\sigma}) \right) \leq C_3 \bar{\lambda} \left( \left( (CM + M^2)h^{3/2} + CMh^2 + CMh^{5/2} \right) \left\| \sigma \right\|_0 \right.

\[ + C_1^{1/2} (Mh)^{3/2} \left\langle \left( \sigma^- - \hat{\sigma}^+ \right) \right\rangle_{h, u_0^h} \right) \]  

(2.4.84)

To bound the third term on the RHS of (2.4.72) we have

\[ \bar{\lambda}(g_a(\sigma, \nabla u^h_0, \sigma)) - (g_a(\sigma, \nabla u, \sigma)) \]

\[ \leq \bar{\lambda} \left( \left\| g_a(\sigma - \hat{\sigma}^h, \nabla (u - u_0^h)) \right\| \right.

\[ + \left\| g_a(\sigma, \nabla u, \hat{\sigma}) \right\| + \left\| g_a(\sigma, \nabla (u - u_0^h), \hat{\sigma}) \right\| \right) \]

\[ \leq \bar{\lambda} \left( C \left\| \sigma - \hat{\sigma}^h \right\|_0 \left\| \nabla (u - u_0^h) \right\|_0 \left\| \hat{\sigma} \right\|_\infty \right.

\[ + C \left\| \sigma - \hat{\sigma}^h \right\|_0 \left\| \nabla u_0^h \right\|_\infty \left\| \hat{\sigma} \right\|_0 + C \left\| \sigma \right\|_\infty \left\| \nabla (u - u_0^h) \right\|_0 \left\| \hat{\sigma} \right\|_0 \right) \]

\[ \leq \bar{\lambda} \left( C(C'Mh^2)(Ch^{3/2})(Ch^{-1} \left\| \sigma \right\|_0) + C(C'Mh^2)(M) \left\| \hat{\sigma} \right\|_0 + C(M)(Ch^{3/2}) \left\| \hat{\sigma} \right\|_0 \right) \]

\[ \leq \bar{\lambda} C_4 \left( CMh^{3/2} + C'M^2h^2 + CM'h^{5/2} \right) \left\| \hat{\sigma} \right\|_0. \]  

(2.4.85)
For the fourth term on the RHS of (2.4.72) we can write

\[(\lambda - \bar{\lambda}) \left( B^h(u_0^h, \sigma_0^h, \hat{\sigma}) - B^h(u, \sigma, \hat{\sigma}) \right) \leq (\lambda - \bar{\lambda}) \left( |B^h(u_0^h, \sigma - \sigma_0^h, \hat{\sigma})| + |B^h(u - u_0^h, \sigma, \hat{\sigma})| \right). \tag{2.4.86}\]

The first term in (2.4.86) is done similarly to (2.4.78):

\[B^h(u_0^h, \sigma - \sigma_0^h, \hat{\sigma}) = -((u_0^h \cdot \nabla)\hat{\sigma}, \sigma - \sigma_0^h) - \frac{1}{2}((\nabla \cdot u_0^h)\hat{\sigma}, \sigma - \sigma_0^h) + \langle (\sigma - \sigma_0^h)^-, \hat{\sigma}^- - \hat{\sigma}^+ \rangle_{h, u_0^h}. \tag{2.4.87}\]

The first term requires a bound on \(\|u - u_0^h\|_\infty\). We note that, from the proof of Lemma 2.4.2 that

\[\|u - u_0\|_\infty \leq \|u - \tilde{u}\|_\infty + \|\tilde{u} - u_0\|_\infty \leq C \left( Mh + Mh^{3/2} + \bar{C}h \right) \leq C \bar{C}h. \tag{2.4.88}\]

With

\[((u_0^h \cdot \nabla)\hat{\sigma}, \sigma - \sigma_0^h)_{h} = (((u_0^h - u) \cdot \nabla)\hat{\sigma}, \sigma - \sigma_0^h)_{h} + ((u \cdot \nabla)\hat{\sigma}, \sigma - \sigma_0^h)_{h} \leq C \left((\bar{C}h) \|\sigma - \sigma_0^h\|_0 \|\nabla \hat{\sigma}\|_0 + \|u\|_\infty \|\nabla \hat{\sigma}\|_0 \|\sigma - \sigma_0^h\|_0\right) \leq C \left((\bar{C}h) (\bar{C}h^{3/2})(h^{-1}) + h^{-1}(M)(\bar{C}h^{3/2}) \right) \|\hat{\sigma}\|_0 \leq C \left(\bar{C}Mh^{1/2} + \bar{C}^2h^{3/2} \right) \|\hat{\sigma}\|_0\]

and, using \(\nabla \cdot u = 0\),

\[((\nabla \cdot u_0^h)\hat{\sigma}, \sigma - \sigma_0^h) \leq Ch^{-1}\|\nabla(u - u_0^h)\|_0 \|\sigma - \sigma_0^h\|_0 \|\hat{\sigma}\|_0 \leq C \bar{C}^2h^2 \|\hat{\sigma}\|_0\]
and

\[
((\sigma - \sigma_0^h)^-, \hat{\sigma}^- - \hat{\sigma}^+)_{h,u_0^h} \leq C \|u_0^h\|_{\infty}^{1/2} \|\sigma - \sigma_0^h\|_0 \langle \langle \hat{\sigma}^- - \hat{\sigma}^+ \rangle \rangle_{h,u_0^h}
\]

\[
\leq C (C_u M)^{1/2} \left( h^{-1/2} \|\sigma - \sigma_0^h\|_0 \langle \langle \hat{\sigma}^- - \hat{\sigma}^+ \rangle \rangle_{h,u_0^h} \right)
\]

\[
\leq C \overline{C} (C_u M)^{1/2} h \langle \langle \hat{\sigma}^- - \hat{\sigma}^+ \rangle \rangle_{h,u_0^h}
\]

we are able to bound (2.4.87) as

\[
B^h(u_0^h, \sigma - \sigma_0^h, \hat{\sigma})
\]

\[
\leq C \left( \overline{C} Mh^{1/2} + \overline{C}^2 h^{3/2} + \overline{C}^2 h^2 \right) \|\hat{\sigma}\|_0 + \overline{C} h (C_u M)^{1/2} \langle \langle \hat{\sigma}^- - \hat{\sigma}^+ \rangle \rangle_{h,u_0^h} \quad \text{(2.4.89)}
\]

Note that the second term in (2.4.86) is the same as the second term in (2.4.74), and is bounded in (2.4.83). Thus we have

\[
(\lambda - \overline{\lambda}) \left( B^h(u_0^h, \sigma_0^h, \hat{\sigma}) - B^h(u, \sigma, \hat{\sigma}) \right)
\]

\[
\leq C_5 (\lambda - \overline{\lambda}) \left( \overline{C} \left( Mh^{1/2} + (M + \overline{C})h^{3/2} + \overline{C} h^2 \right) \|\hat{\sigma}\|_0 + \overline{C} h (C_u M)^{1/2} \langle \langle \hat{\sigma}^- - \hat{\sigma}^+ \rangle \rangle_{h,u_0^h} \right) \quad \text{(2.4.90)}
\]

To bound the fifth term on the RHS of (2.4.72) we have

\[
(\lambda - \overline{\lambda}) \left( g_a(\sigma_0^h, \nabla u_0^h), \hat{\sigma} \right) - \left( g_a(\sigma, \nabla u), \hat{\sigma} \right)
\]

\[
= (\lambda - \overline{\lambda}) \left( \left( g_a(\sigma - \sigma_0^h, \nabla (u - u_0^h)), \hat{\sigma} \right)
\]

\[
- \left( g_a(\sigma - \sigma_0^h, \nabla u), \hat{\sigma} \right) - \left( g_a(\sigma, \nabla (u - u_0^h)), \hat{\sigma} \right) \right)
\]

\[
\leq (\lambda - \overline{\lambda}) \left( C \|\sigma - \sigma_0^h\|_0 \|\nabla (u - u_0^h)\|_0 \|\hat{\sigma}\|_\infty + C \|\sigma - \sigma_0^h\|_0 \|\nabla u\|_\infty \|\hat{\sigma}\|_0 
\]

\[
+ C \|\sigma\|_\infty \|\nabla (u - u_0^h)\|_0 \|\hat{\sigma}\|_0 \right)
\]

\[
\leq (\lambda - \overline{\lambda}) \left( C(\overline{C} h^{3/2})(\overline{C} h^{3/2})(Mh^{-1}) \|\hat{\sigma}\|_0 
\]

\[
+ C(\overline{C} h^{3/2})(M) \|\hat{\sigma}\|_0 + C(M)(\overline{C} h^{3/2}) \|\hat{\sigma}\|_0 \right)
\]

\[
\leq C_6(\lambda - \overline{\lambda}) \left( \overline{C} Mh^{3/2} + \overline{C}^2 h^2 \right) \|\hat{\sigma}\|_0. \quad \text{(2.4.91)}
\]
For the pressure term in (2.4.72), from [38] we have
\[
\| (p - q^h, \nabla \cdot \hat{u} ) \| \leq C \| p - q^h \|_0 \| \nabla \hat{u} \|_0
\]
\[
\leq C h^2 \| p \|_2 \| \nabla \hat{u} \|_0
\]
\[
\leq C_7 M h^2 \| \nabla \hat{u} \|_0.
\] (2.4.92)

Combining the estimates (2.4.73), (2.4.84), (2.4.85), (2.4.90), (2.4.91), and (2.4.92), we have
\[
\tilde{A}(u_h^0, (\hat{\sigma}, \hat{u}), (\hat{\sigma}, \hat{u})) + \overline{A}B^h(u_h^0, \hat{\sigma}, \hat{\sigma})
\]
\[
\leq C_1 M h^2 \| \hat{\sigma} \|_0 + C_2 M h^2 \| \nabla \hat{u} \|_0
\]
\[
+ C_3 \bar{\lambda} \left( \left( \overline{C} M + M^2 \right) h^{3/2} + \overline{C} M h^2 + \overline{C} M h^{5/2} \right) \| \hat{\sigma} \|_0
\]
\[
+ C_4 \bar{\lambda} \left( \overline{C} M h^{3/2} + C'M^2 h^2 + \overline{C} C'M h^{5/2} \right) \| \hat{\sigma} \|_0
\]
\[
+ C_5(\lambda - \bar{\lambda}) \left( \overline{C} \left( M h^{1/2} + (M + \overline{C}) h^{3/2} + \overline{C} h^2 \right) \right) \| \hat{\sigma} \|_0
\]
\[
+ \overline{C} h \left( C_u M \right)^{1/2} \langle \langle \hat{\sigma}^- - \hat{\sigma}^+ \rangle \rangle_{h,u_h^0}
\]
\[
+ C_6(\lambda - \bar{\lambda}) \left( \overline{C} M h^{3/2} + \overline{C}^2 h^2 \right) \| \hat{\sigma} \|_0 + C_7 M h^2 \| \nabla \hat{u} \|_0
\]
\[
\leq \left( C_1 M h^2 + C_3 \bar{\lambda} \left( \overline{C} M + M^2 \right) h^{3/2} + \overline{C} M h^2 + \overline{C} M h^{5/2} \right)
\]
\[
+ C_4 \bar{\lambda} \left( \overline{C} M h^{3/2} + C'M^2 h^2 + \overline{C} C'M h^{5/2} \right)
\]
\[
+ C_5(\lambda - \bar{\lambda}) \left( \overline{C} \left( M h^{1/2} + (M + \overline{C}) h^{3/2} + \overline{C} h^2 \right) \right)
\]
\[
+ C_6(\lambda - \bar{\lambda}) \left( \overline{C} M h^{3/2} + \overline{C}^2 h^2 \right) \| \hat{\sigma} \|_0
\]
\[
+ (C_2 + C_7) M h^2 \| \nabla \hat{u} \|_0
\]
\[
+ \left( C_3 \bar{\lambda} C_u^{1/2}(M h)^{3/2} + C_5(\lambda - \bar{\lambda}) \overline{C} h \left( C_u M \right)^{1/2} \right) \langle \langle \hat{\sigma}^- - \hat{\sigma}^+ \rangle \rangle_{h,u_h^0}. \] (2.4.93)
Recalling that $\lambda - \bar{\lambda} = K_1 h^{3/2}$ and $\lambda - \bar{\lambda} = K_2 h^{3/2}$ we have

\[
\tilde{A}(u_0^h, (\bar{\sigma}, \tilde{u}), (\bar{\sigma}, \tilde{u})) + \lambda B^h(u_0^h, \bar{\sigma}, \bar{\sigma}) \\
\leq \left( C_1 Mh^2 + C_3 \lambda \left( (\overline{CM} + M^2)h^{3/2} + \overline{CM}h^2 + \overline{CM}h^{5/2} \right) \\
+ C_4 \lambda \left( \overline{CM}h^{3/2} + C'M^2h^2 + \overline{CM}'Mh^{5/2} \right) \\
+ C_5 K_1 h^{3/2} \left( \overline{C} \left( Mh^{1/2} + (M + \overline{C})h^{3/2} + \overline{C}h^2 \right) \right) \\
+ C_6 K_2 h^{3/2} \left( \overline{CM}h^{3/2} + \overline{C}^2h^2 \right) \right) \|\bar{\sigma}\|_0 \\
+ (C_2 + C_7) Mh^2 \|\nabla \tilde{u}\|_0 \\
+ \left( C_3 \lambda C_1^{1/2}(Mh)^{3/2} + C_5 K_1 h^{3/2} \overline{C}h (C_uM)^{1/2} \right) \langle \langle \bar{\sigma}^- - \bar{\sigma}^+ \rangle \rangle_{h,u_0^h}. \quad (2.4.94)
\]

We note that the smallest power of $h$ on $\|\bar{\sigma}\|_0$ and $\langle \langle \bar{\sigma}^- - \bar{\sigma}^+ \rangle \rangle_{h,u_0^h}$ is $3/2$. Thus, neglecting the higher powers of $h$ we have, from (2.4.72) and (2.4.95),

\[
(1 - 4\lambda \overline{M})\|\bar{\sigma}\|_0^2 + 4\alpha(1 - \alpha)\|\nabla \tilde{u}\|_0^2 + \frac{\lambda}{2} \langle \langle \bar{\sigma}^- - \bar{\sigma}^+ \rangle \rangle_{h,u_0^h}^2 \\
\leq \tilde{A}(u_0^h, (\bar{\sigma}, \tilde{u}), (\bar{\sigma}, \tilde{u})) + \lambda B^h(u_0^h, \bar{\sigma}, \bar{\sigma}) \\
\leq \left( C_3 \lambda (\overline{CM} + M^2) + C_4 \lambda \overline{C}M \right) h^{3/2} \|\bar{\sigma}\|_0 \\
+ (C_2 + C_7) Mh^2 \|\nabla \tilde{u}\|_0 \\
+ C_3 \lambda C_1^{1/2} M^{3/2} h^{3/2} \langle \langle \bar{\sigma}^- - \bar{\sigma}^+ \rangle \rangle_{h,u_0^h}. \quad (2.4.95)
\]

Thus (2.4.95), (2.4.34), Young’s inequality, and $M \leq \overline{C}$ implies

\[
\frac{1}{2} (1 - 4\lambda \overline{M})\|\bar{\sigma}\|_0^2 + 2\alpha(1 - \alpha)\|\nabla \tilde{u}\|_0^2 + \frac{\lambda}{4} \langle \langle \bar{\sigma}^+ - \bar{\sigma}^- \rangle \rangle_{h,u_0^h}^2 \\
\leq \left( \frac{2 \left( C_3 \lambda M + C_4 \lambda M \right)^2}{(1 - 4\lambda \overline{M})} + \frac{h (C_2 + C_7)^2}{2\alpha(1 - \alpha)} + \frac{\lambda}{4} \left( C_3 C_1^{1/2} M^{1/2} \right)^2 \right) \frac{\overline{C}^2}{4} h^3. \quad (2.4.96)
\]

Let $\delta$ be given by

\[
2\delta^2 = \min \left\{ \frac{1}{2} (1 - 4\lambda \overline{M}), 2\alpha(1 - \alpha) \right\}.
\]
Then, (2.4.96) implies

\[
(\sqrt{2}\delta \|\hat{\sigma}\|_0)^2 + (\sqrt{2}\delta \|\nabla \hat{u}\|_0)^2 \\
\leq \left(\frac{2(2C_3\lambda_M + C_4\tilde{\lambda}M)^2}{(1-4\lambda M)} + \frac{h(C_2 + C_7)^2}{2\alpha(1-\alpha)} + \tilde{\lambda}(C_3C_u^{1/2}M^{1/2})^2\right) \frac{C^2}{4} h^3,
\]
i.e.,

\[
\delta \|\hat{\sigma}\|_0 + \delta \|\nabla \hat{u}\|_0 \\
\leq \left(\frac{2(2C_3\lambda M + C_4\tilde{\lambda}M)^2}{(1-4\lambda M)} + \frac{h(C_2 + C_7)^2}{2\alpha(1-\alpha)} + \tilde{\lambda}(C_3C_u^{1/2}M^{1/2})^2\right)^{1/2} \frac{C}{2} h^{3/2}.
\]

Hence, for \(M\) and \(h\) sufficiently small such that

\[
\frac{C}{\delta} \left(\frac{2(2\lambda M + \tilde{\lambda}M)^2}{(1-4\lambda M)} + \frac{h}{2\alpha(1-\alpha)} + \tilde{\lambda}M\right)^{1/2} \leq 1,
\]
we have

\[
\|\hat{\sigma}\|_0 + \|\nabla \hat{u}\|_0 \leq \frac{C}{2} h^{3/2}.
\]

Thus (2.4.60) implies

\[
\|\sigma - \sigma^h_1\|_0 + \|\nabla (u - u^h_1)\|_0 \\
\leq \|\hat{\sigma}\|_0 + \|\nabla \hat{u}\|_0 + \|\sigma - \hat{\sigma}\|_0 + \|\nabla (u - \hat{u})\|_0 \leq \frac{C}{2} h^{3/2} + \frac{C}{2} h^{3/2} = Ch^{3/2}.
\]

Hence \(\Phi(B_h) \subset B_h\).

Thus we have the following result.

**Theorem 2.4.1** Assume that \((\sigma^h_0, u^h_0)\) is the solution obtained after Step 1 of Algorithm 2.4.1. Then for \(M\) and \(h\) sufficiently small, there exists a fixed point \((\sigma^h, u^h)\) of Step 2 of Algorithm 2.4.1 that satisfies (2.3.13)-(2.3.14) and each of the iterates \((\sigma^h_i, u^h_i), i = 1, 2, \ldots\), satisfy

\[
\|\sigma - \sigma^h_i\|_0 + \|\nabla (u - u^h_i)\|_0 \leq Ch^{3/2}
\]
for constant $C > 0$.

**Proof:** If $M$ satisfies the small data condition of Lemma 2.4.1, the solution $\left(\sigma_0^h, u_0^h\right)$ of the defect step satisfies

$$\|\sigma - \sigma_0^h\|_0 + \|\nabla(u - u_0^h)\|_0 \leq C_0 h^{3/2}.$$ 

Then $\left(\sigma_0^h, u_0^h\right) \in \mathcal{B}_h$ where

$$\mathcal{B}_h = \left\{(\tau^h, v^h) \in \Sigma^h \times V^h : \|\sigma - \tau^h\|_0, \|\nabla(u - v^h)\|_0 \leq C_0 h^{3/2}\right\}.$$ 

Then by Lemmas 2.4.3-2.4.5, the mapping $\Phi$ defined by (2.4.26)-(2.4.27) satisfies $\Phi(\mathcal{B}_h) \subset \mathcal{B}_h$, and Schauder’s Fixed-Point Theorem guarantees the existence of a fixed point $\left(\sigma^h, u^h\right) \in \mathcal{B}_h$ of $\Phi$ solving (2.3.13)-(2.3.14).

2.4.4 Implementation of the Method

In computing approximations to (2.2.1)-(2.2.3) for a particular domain and set of problem parameters, a critical Weissenberg number $\lambda^*$ is encountered, and nonlinear solvers fail to converge for values of $\lambda > \lambda^*$. The defect-correction method presented in Section 2.4 seeks to extend the range of Weissenberg number for which solutions can be computed. A common approach to computing solutions for increasing $\lambda$ is to use a converged solution computed at a smaller value, say $\lambda_0$, as the initial iterate for the nonlinear solver ([16, 17, 20, 60, 66, 79, 80] are some examples). Once the solution is computed for $\lambda$, then it can be used as an initial iterate for an incrementally larger Weissenberg number. However, one will still encounter the critical value $\lambda^*$, and the traditional nonlinear solution algorithm will fail. Intuitively, one can then use Algorithm 2.4.1 to compute solutions at a Weissenberg number larger than $\lambda^*$, provided the defect parameters $\overline{\lambda}, \overline{\lambda} \leq \lambda^*$. In the following sections we discuss some of the considerations that arise when implementing this defect-correction method including an alternative correction iteration, and present a brief investigation into the influence of the parameters $\overline{\lambda}$ and $\overline{\lambda}$. 45
2.4.5 Newton Corrector Iteration

The correction step (Step 2) in Algorithm 2.4.1 uses a linearized iteration in order to solve the original nonlinear problem. This linearization will be referred to as the Picard corrector, as some terms that are present in the “full” linearization are not present (precisely those which contain updated velocity and lagged stress). This full linearization is obtained from computing the Fréchet derivative of the original problem at a known value \((\sigma_i^h, u_i^h, p_i^h)\), which we called the Newton linearization. Within the context of the defect correction method, the Newton corrector iteration has the form:

\[
A((\sigma_{i+1}^h, u_{i+1}^h), (\tau^h, v^h)) + \bar{\lambda}B^h(u_i^h, \sigma_{i+1}^h, \tau^h) + \bar{\lambda}B^h(u_{i+1}^h, \sigma_i^h, \tau^h) + \bar{\lambda}(g_a(\sigma_{i+1}^h, \nabla u_i^h), \tau^h) + \bar{\lambda}(g_a(\sigma_i^h, \nabla u_{i+1}^h), \tau^h) = 2\alpha(f, v^h) - (\lambda - 2\bar{\lambda})B^h(u_i^h, \sigma_i^h, \tau^h) - (\lambda - 2\bar{\lambda})(g_a(\sigma_i^h, \nabla u_i^h), \tau^h) \quad \forall(\tau^h, v^h) \in \Sigma^h \times V^h. \tag{2.4.97}
\]

The Newton corrector is expected to converge faster than the Picard corrector. However, convergence of the Newton corrector has not been proven. In addition, the computational implementation of (2.4.97) is nontrivial; the term \(B^h(u_i^h, \sigma_i^h, \tau^h)\) requires the calculation of the jump of the lagged stress \(\sigma_i^h\) in the direction of an unknown velocity \(u_i^h\). This difficulty is avoided by modifying (2.4.97) to use \(u_i^h\) for the jump term in all occurrences of \(B^h\).

2.4.6 Choosing the defected parameters

The defect-correction algorithm differs from the standard nonlinear solution approach by replacing \(\lambda\) with the reduced parameters \(\bar{\lambda}\) and \(\bar{\lambda}\). In the proof of the convergence of the method, it was necessary to set \(\bar{\lambda} = \lambda - K_1 h^{3/2}\) and \(\bar{\lambda} = \lambda - K_2 h^{3/2}\), \(K_1, K_2 \geq 0\). To compute approximations for a Weissenberg number larger than the critical value \(\lambda^*\), it may be necessary to set at least one of the defect parameters \(\bar{\lambda}\) and \(\bar{\lambda}\) to a value less than \(\lambda^*\). For example, if \(\bar{\lambda} = \lambda - K_2 h^{3/2} < \lambda^*\) is required, then we have the bound

\[
\lambda < \lambda^* + K_2 h^{3/2},
\]

46
which implies that as we refine the computational mesh, the bound on $\lambda$ will decrease. This undesired result is not entirely surprising, as it is commonly encountered that convergence of viscoelastic fluid flow solution methods for a fixed Weissenberg number degrades with mesh refinement [5].

Theoretically, the choice of the defect parameters also depends on the true solution and the various conditions (2.4.61), (2.4.34), and (2.4.62) that arise in the analysis of the method. As these conditions depend on unknown constants, it is not clear what values of $\bar{\lambda}$ and $\tilde{\lambda}$ are optimal for a given problem. At present, there is no good algorithm for choosing these defected parameters. The future development of an a posteriori error estimate for the defect correction method may lead to heuristics for choosing the parameters.

2.5 Numerical Experiments

To investigate the accuracy and effectiveness of the defect-correction method, computational experiments are designed to gauge these attributes. The method has been implemented using the finite element software package FreeFem++ [45] in 2-d. Linear systems are solved using the UMFPACK solver. As described in the theoretical analysis, continuous piecewise quadratic elements are used for velocity, continuous piecewise linears are used for pressure, and discontinuous piecewise linears are used for stress. The solution approach used for undefected nonlinear problems, as well as the nonlinear defect step of the defect-correction algorithm, is a standard Newton iteration scheme. Results obtained on two different problems are presented:

1. A nonphysical problem on a square domain that has a known analytic solution. Computations show that the method satisfies the theoretical spatial convergence rate and converges to the same solution as the standard undefected solution approach.

2. A commonly cited benchmark physical problem of a four-to-one planar contraction flow. Computations show that solutions computed under mesh refinement converge to a solution computed using the standard undefected solution approach on a very fine spatial mesh. Computations also show that the method is successful in computing solutions beyond the critical Weissenberg value.

The three different methods for computing solutions are labeled as:

DCP: A defect-correction method with a standard Newton nonlinear iteration in the defect step and a Picard-like linearization in the correction step.

2.5.1 Example 1

As in [29] and [58], \( \Omega = [0, 1] \times [0, 1] \subset \mathbb{R}^2 \), and chosen functions are added to the right-hand sides of (2.2.1)-(2.2.3) so that the true solution to the problem is given by

\[
\begin{pmatrix}
\sin(\pi x)y(y - 1) \\
\sin(x)(x - 1)y\cos(\frac{\pi y}{2})
\end{pmatrix}, \quad p = \cos(2\pi x)y(y - 1), \quad \sigma = 2\alpha D(u).
\]

The parameters \( \alpha, a \) in the equations are chosen as 0.5 and 0, respectively. A stopping criterion of

\[
\|u^h_i - u^h_{i-1}\|_{\infty}, \|\sigma^h_i - \sigma^h_{i-1}\|_{\infty} \leq 10^{-8}
\]

is used for the iterative nonlinear solver in both the undefected nonlinear solution approach and the defect step of the defect-correction algorithm. The linear correction step in the defect-correction algorithm is iterated to the same tolerance. The solution computed for \( \lambda = 4.0 \) was used as the initial iterate in all cases.

Table 2.5.1 presents results for \( \lambda = 5.0 \). Specifically, the errors results and mesh convergence rates obtained by the STD method and the DCP method for several values of \( \lambda \) and \( \tilde{\lambda} \) are given. The resulting errors for both methods are seen to be the same. It should be noted that the convergence rates obtained for the cases below are greater than the theoretical spatial convergence rate of \( 3/2 \). As is expected, for larger defects \( \lambda - \tilde{\lambda} \) and \( \lambda - \lambda \), the nonlinear and linear corrector iteration counts increase.

Also of interest is how DCP performs in comparison to DCN. In all experiments where both correction steps converged, each method obtained the same error result. Very few cases were observed in which one corrector method converged while the other did not. When the methods did not converge, the Newton corrector was more likely to diverge while the Picard corrector would stagnate. In most cases, the Picard iteration required more correction steps for convergence than the Newton iteration, as anticipated. Table 2.5.2 presents iteration counts obtained by both methods on the mesh \( h = 1/8 \) for several combinations of \( \lambda \) and \( \tilde{\lambda} \). In this table, NNI represents the number of nonlinear iterations.
Table 2.5.1 Errors and convergence rates for $\lambda = 5.0$ and selected values of $\bar{\lambda}$, $\tilde{\lambda}$, defect correction method (Picard corrector).

<table>
<thead>
<tr>
<th>$H^1$ error of $u$</th>
<th>$l_2$ error of $\sigma$</th>
<th>Number of nonlinear iterations (NNI)</th>
<th>Number of correction steps (NCS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 5.0$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{\lambda} = 4.9$</td>
<td>$5.5428E-02$</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>$\lambda = 4.5$</td>
<td></td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>$\bar{\lambda} = 4.5$</td>
<td>$5.5429E-02$</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>$\lambda = 4.0$</td>
<td></td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>$\bar{\lambda} = 4.0$</td>
<td>$5.5429E-02$</td>
<td>8</td>
<td>22</td>
</tr>
</tbody>
</table>

required in the defect step, NCS represents the number of linear correction steps, and an asterisk denotes that the particular method did not converge for the given parameters.

Table 2.5.2 Iteration counts for Picard and Newton correctors, $h = 1/8$, $\lambda = 5.0$, selected values of $\bar{\lambda}$ and $\tilde{\lambda}$

<table>
<thead>
<tr>
<th>$\lambda = 5.0$</th>
<th>NCS (Picard)</th>
<th>NCS (Newton)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{\lambda}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tilde{\lambda}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.9 4.9</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>4.5 4.5</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>4 4</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>3 3</td>
<td>8</td>
<td>129</td>
</tr>
<tr>
<td>5 4.5</td>
<td>8</td>
<td>29</td>
</tr>
<tr>
<td>4.5 5</td>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td>5 4</td>
<td>10</td>
<td>46</td>
</tr>
<tr>
<td>4 5</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>5 3</td>
<td>11</td>
<td>612</td>
</tr>
<tr>
<td>3 5</td>
<td>12</td>
<td>81</td>
</tr>
<tr>
<td>4.5 4</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>4 4.5</td>
<td>8</td>
<td>26</td>
</tr>
</tbody>
</table>


Additionally, we investigate the performance of the Picard corrector for different choices for the defect parameters. To examine this behavior, for $\lambda = 5$ and a grid size of $h = 1/8$, we fix one of the defect parameters and then vary the other, noting how many corrector iterations were required for convergence of the corrector at each value of the varied parameter. We remark that for the cases considered here the defect step required between 7 and 12 iterations. Figure 2.5.1 presents iteration counts required for various values of $\tilde{\lambda}$ with $\lambda$ fixed at 4.9, 4.5 and 4.0.

![Figure 2.5.1 Iteration counts for fixed values of $\lambda$, $\lambda = 5.0$, $h = 1/8$](image)

From Figure 2.5.1 we note that when $\lambda$ was fixed, a choice of $4 \leq \tilde{\lambda} \leq 5$ did not significantly affect the iteration count; however choosing $\tilde{\lambda} < 4$ required many more corrector iterations for convergence. This suggests that large single step defects may not be computationally efficient.
Figure 2.5.2 Iteration counts for fixed values of $\lambda$, $\lambda = 5.0$, $h = 1/8$

Figure 2.5.2 shows the results with the roles of the parameters are reversed. Note that the scale of the vertical axis is smaller than that of Figure 2.5.1.

In Figure 2.5.2 we also see that when $\lambda$ is fixed, the lowest number of corrector iterations were required when $\lambda \approx \lambda$. It is also observed that far fewer iterations were required in the far left range of the plot as compared with Figure 2.5.1. The information shown in Figures 2.5.1 and 2.5.2, together with the data presented in Table 2.5.2, indicate that it is reasonable to choose similar values for the two defect parameters.

2.5.2 Example 2

Numerical simulations of viscoelastic flow through a planar or axisymmetric contraction have been widely studied (see [5] or Chapter 8 of [65]). Here the case of planar...
flow through a contraction geometry with a ratio of 4:1 with respect to upstream and downstream channel widths is considered. The contraction angle is a fixed $3\pi/2$ and the channel lengths are sufficiently long to impose a fully developed Poiseuille flow in the inflow and outflow channels. The geometry of the computational domain is illustrated in Figure 2.5.3. The lower left corner of the domain corresponds to $x = y = 0$.

![Figure 2.5.3 Geometry of 4:1 contraction domain](image)

The factor $L$ is set to $1/4$ for these computations. On this domain the velocity boundary conditions are

$$
\mathbf{u} = \begin{bmatrix}
\frac{1}{32}(1 - y^2) \\
0
\end{bmatrix} \quad \text{on } \Gamma_{\text{in}}, \quad \mathbf{u} = \begin{bmatrix}
2 \left(\frac{1}{16} - y^2\right) \\
0
\end{bmatrix} \quad \text{on } \Gamma_{\text{out}}.
$$

Boundary conditions for $\sigma$ must be specified on the inflow boundary. From the constitutive equation (3.1.1) and the velocity conditions (2.5.1), for $u_{1,y} = \partial u_1/\partial y$, we have

$$
\sigma_{11} = \frac{-\alpha \lambda (a + 1) u_{1,y}^2}{(a^2 - 1) \lambda^2 u_{1,y}^2 - 1}, \quad \sigma_{12} = \frac{-\alpha u_{1,y}}{(a^2 - 1) \lambda^2 u_{1,y}^2 - 1}, \\
\sigma_{22} = \frac{-\alpha \lambda (a - 1) u_{1,y}^2}{(a^2 - 1) \lambda^2 u_{1,y}^2 - 1}, \quad \text{on } \Gamma_{\text{in}}.
$$

Symmetry conditions are imposed on the bottom of the computational domain. The parameter $\alpha$ is set to $8/9$ and the initial iterate was given by $(\sigma, \mathbf{u}, p) = (0, 0, 0)$. Computations were performed on three different meshes, M1, M2, and M3. Table 2.5.3 lists the characteristics of the meshes.

52
To determine the accuracy of the defect-correction method, solutions computed on successive meshes were compared to a solution computed by the STD method on a very fine spatial mesh (114811 degrees of freedom). For the results shown here, \( a = 1 \) (Oldroyd-B model) and \( \lambda = 0.7 \). Table 2.5.4 gives values of solutions norms computed by the STD and DCP methods. As the mesh is refined, the norms of the computed approximations approach the values computed on the very fine mesh.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mesh</th>
<th>( |u^h|_0 )</th>
<th>( |u^h|_1 )</th>
<th>( |\sigma^h|_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>STD</td>
<td>Fine</td>
<td>0.104166</td>
<td>0.595209</td>
<td>0.932091</td>
</tr>
<tr>
<td>DCP</td>
<td>M1</td>
<td>0.104243</td>
<td>0.595417</td>
<td>0.935189</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.5 )</td>
<td>M2</td>
<td>0.104207</td>
<td>0.595287</td>
<td>0.933602</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.5 )</td>
<td>M3</td>
<td>0.104183</td>
<td>0.595243</td>
<td>0.932683</td>
</tr>
<tr>
<td>DCP</td>
<td>M1</td>
<td>0.104243</td>
<td>0.595417</td>
<td>0.935193</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.35 )</td>
<td>M2</td>
<td>0.104206</td>
<td>0.595287</td>
<td>0.933606</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.35 )</td>
<td>M3</td>
<td>0.104183</td>
<td>0.595243</td>
<td>0.932687</td>
</tr>
<tr>
<td>DCP</td>
<td>M1</td>
<td>0.104243</td>
<td>0.595417</td>
<td>0.935192</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.4 )</td>
<td>M2</td>
<td>0.104206</td>
<td>0.595287</td>
<td>0.933604</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.6 )</td>
<td>M3</td>
<td>0.104183</td>
<td>0.595243</td>
<td>0.932686</td>
</tr>
<tr>
<td>DCP</td>
<td>M1</td>
<td>0.104243</td>
<td>0.595417</td>
<td>0.935187</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.6 )</td>
<td>M2</td>
<td>0.104207</td>
<td>0.595287</td>
<td>0.933599</td>
</tr>
<tr>
<td>( \tilde{\lambda} = 0.4 )</td>
<td>M3</td>
<td>0.104184</td>
<td>0.595243</td>
<td>0.932681</td>
</tr>
</tbody>
</table>

Table 2.5.4  Solution norms, successive meshes, \( \lambda = 0.7 \), \( a = 1 \).

For each method, \( \lambda^* \) was the “maximum” value for which the method converged. The values of \( \lambda^* \) obtained by STD, DCN, and DCP on all meshes for different values of \( a \) are presented in Table 2.5.5. For the values of \( a \) considered here, the \( \lambda^*-\)DCP values were...
the same as the $\lambda^*$-STD values for the case $a \neq 0$. However, DCN significantly improves the $\lambda^*$ values obtained by STD.

<table>
<thead>
<tr>
<th>Mesh</th>
<th>$a$</th>
<th>$\lambda^*$-STD</th>
<th>$\lambda^*$-DCP</th>
<th>$\lambda^*$-DCN</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.0</td>
<td>2.430</td>
<td>2.430</td>
<td>4.993</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.286</td>
<td>2.722</td>
<td>2.719</td>
</tr>
<tr>
<td></td>
<td>-1.0</td>
<td>1.465</td>
<td>1.465</td>
<td>3.121</td>
</tr>
<tr>
<td>M2</td>
<td>1.0</td>
<td>1.821</td>
<td>1.821</td>
<td>3.682</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.934</td>
<td>1.990</td>
<td>1.993</td>
</tr>
<tr>
<td></td>
<td>-1.0</td>
<td>1.321</td>
<td>1.321</td>
<td>2.738</td>
</tr>
<tr>
<td>M3</td>
<td>1.0</td>
<td>1.412</td>
<td>1.412</td>
<td>2.810</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.823</td>
<td>1.666</td>
<td>1.663</td>
</tr>
<tr>
<td></td>
<td>-1.0</td>
<td>1.170</td>
<td>1.170</td>
<td>2.185</td>
</tr>
</tbody>
</table>

Table 2.5.5  Critical $\lambda$ values for $(\sigma, u, p) = (0, 0, 0)$.

The two examples illustrate that the defect-correction approach can be effective in extending the range of Weissenberg numbers for which a numerical approximation to the viscoelastic fluid flow equations can be obtained. The optimal choices of the defect parameters is still an open question.
CHAPTER 3
CONTINUATION METHODS FOR JOHNSON-SEGALMAN FLUIDS

The incremental procedure described in Section 2.4.4 can be thought of as a continuation process. Continuation methods are a natural extension of the defect-correction method. For nonlinear problems, continuation methods have been developed as a way to analyze the behavior of solutions as a problem parameter is varied [2, 48, 49, 71, 64], and numerous computational algorithms and software packages have been developed to implement these methods [72, 74]. Predictor-corrector continuation methods trace the solution manifold through the parameter space by first forming a good approximation to the solution at an incremented value of the parameter and then correcting through an iterative scheme. This approach may identify singular and bifurcation points along the manifold which are of particular interest.

Continuation methods have been studied within the context of structural mechanics and elasticity [63, 53, 40] as well as fluid dynamics. For issues in simulating the Navier-Stokes equations at high Reynolds number, a common approach is to perform continuation in Reynolds number [42]. Carey and Krishnan [14] described a continuation method applied to a penalty approximation for Navier-Stokes, gave conditions for convergence for the method, and presented numerical results for the driven cavity problem. The authors also formulated a method for continuation in arc length. Gunzburger and Peterson [43] investigated predictor and steplength selection for continuation in Reynolds number and concluded that for certain values of the Reynolds number, the parameter steplength can be chosen independently of the type of predictor step used. Recently, de Almeida and Derby [22] described natural and pseudo-arclength continuation algorithms adapted for large-scale simulations of driven-cavity flows and successfully computed approximations for large values of the Reynolds number.

Similarly, continuation for viscoelastic fluid flow can be performed for increasing values of the Weissenberg number. Various implementations of continuation in Weissenberg
number can be found in [17, 66, 80, 20]. In [60], Mendelson et al. described an algorithm for continuation in Weissenberg number for Maxwell and second-order constitutive equations using a Galerkin finite element method. The authors observed numerically that, as the critical value of the Weissenberg number is approached, the length of the step that can be taken became infeasibly small. In [79], Yeh, et al. used natural and arclength continuation methods to arrive at a bifurcation in the numerical solution of an upper convected Maxwell fluid, at which the authors conjected is a product of the mathematical model of the fluid. However, later work has shown that the multiple solutions were an artifact of the numerical approximation (see [65]).

The objective of this chapter is to describe the application and implementation of continuation algorithms for the discontinuous Galerkin finite element approximation of the steady-state Johnson-Segalman model for viscoelastic fluid flow, and to use these algorithms to study the behavior of the resulting nonlinear system of equations as the Weissenberg number is increased. Simple and natural continuation methods are investigated with incremental increases in the Weissenberg number. Pseudo-arclength continuation reparametrizes the equations with respect to an arclength-like parameter and incorporates an additional constraint to be solved with the underlying system. The remainder of this chapter is organized as follows: in Section 3.1, the original continuous problem and its discrete approximation are described. In Section 3.2 simple and natural continuation approaches are discussed and applied to the discrete problem. In Section 3.3 the pseudo-arclength continuation method and various choices for the pseudo-arclength constraint are discussed and applied to the discrete problem. Numerical experiments with continuation methods are performed on the benchmark 4:1 contraction problem in Section 3.4.
3.1 Problem Description

The description of the model problem will be similar to that in Section 2.2. However, the case of nonhomogeneous boundary conditions for (1.4.8)-(1.4.12) is now considered:

\[ \sigma + \lambda (u \cdot \nabla)\sigma + \lambda g_a(\sigma, \nabla u) - 2\alpha D(u) = 0 \quad \text{in } \Omega, \tag{3.1.1} \]
\[ -\nabla \cdot \sigma - 2(1 - \alpha) \nabla \cdot D(u) + \nabla p = f \quad \text{in } \Omega, \tag{3.1.2} \]
\[ \text{div } u = 0 \quad \text{in } \Omega, \tag{3.1.3} \]
\[ u = u_\Gamma \quad \text{on } \Gamma, \tag{3.1.4} \]
\[ \sigma = \sigma_{\Gamma_{\text{in}}} \quad \text{on } \Gamma_{\text{in}}. \tag{3.1.5} \]

3.1.1 Continuous Problem and Variational Formulation

Existence of a solution to the problem (3.1.1)-(3.1.5) was shown by Renardy [68] under a small data assumption. (See also [41] and [33].) Specifically, if \( \Omega \) has a \( C^\infty \)-smooth boundary, with \( f \) and \( u_\Gamma \) sufficiently regular and small, the problem (3.1.1)-(3.1.5) admits a unique bounded solution \( (\sigma, u, p) \in H^2(\Omega) \times H^3(\Omega) \times H^2(\Omega) \).

The variational formulation is as described in Section 2.2.2 with some modifications due to the change in boundary conditions. Let

\[ X_r := H^1_0(\Omega) = \{ v \in H^1(\Omega) : v = r \quad \text{on } \Gamma \}, \]
\[ S := L_0^2(\Omega) = \{ q \in L^2(\Omega) : \int_\Omega q \, d\Omega = 0 \}, \]
\[ \Sigma_s := (L^2(\Omega))^{d \times d} \cap \{ \tau = (\tau_{ij}) : \tau_{ij} = \tau_{ji}, u \cdot \nabla \tau \in (L^2(\Omega))^{d \times d}, \text{ and } \tau = s \quad \text{on } \Gamma_{\text{in}} \}. \]

The variational formulation is: Given \( f \in H^{-1}(\Omega) \), find \( (\sigma, u, p) \in \Sigma_{\sigma_{\Gamma_{\text{in}}} \times X_{\Gamma_{\text{in}}} \times S} \) such that

\[ (\sigma, \tau) + \lambda ((u \cdot \nabla)\sigma, \tau) + \lambda (g_a(\sigma, \nabla u), \tau) - 2\alpha(D(u), \tau) = 0, \quad \forall \tau \in \Sigma_0 \tag{3.1.6} \]
\[ (\sigma, D(v)) + 2(1 - \alpha)(D(u), D(v)) - (p, \nabla \cdot v) = (f, v), \quad \forall v \in X_0, \tag{3.1.7} \]
\[ (q, \nabla \cdot u) = 0, \quad \forall q \in S. \tag{3.1.8} \]
3.1.2 Finite Element Approximation

The triangulation $T_h$ and the polynomial spaces $P_k(K)$ are as defined in 2.3. Let $P_k(T_h) = \bigcup_{K \in T_h} P_k(K)$. Let $T_h^\Omega : H^1(\Omega) \to P_k(T_h)$ and $T_h^\sigma : (L^2(\Omega))^{d \times d} \to P_k(T_h)$ denote suitable interpolation operators \cite{19,12}. Define the following finite element spaces for the approximation of $(\sigma, u, p)$:

\[
\Sigma_h^\sigma := (L^2(\Omega))^{d \times d} \cap \{ \tau^h = (\tau^h_{ij}) : \tau^h_{ij} = \tau^h_{ji}, \, \tau^h|_K \in P_1(K)^{d \times d} \, \forall K \in T_h, \quad \tau^h = I_1^\sigma s \text{ on } \Gamma_{in} \},
\]

\[
X_r^h := \{ v^h \in H^1(\Omega) \cap (C^0(\Omega))^d : v^h|_K \in P_2(K)^d, \forall K \in T_h, \, v^h = I_2^r \tau \text{ on } \Gamma \},
\]

\[
S_h := \{ q^h \in S \cap C^0(\Omega) : q^h|_K \in P_1(K), \forall K \in T_h \}.
\]

The notation described in Section 2.3 for the discontinuous Galerkin method will be used here. Introduce the operator $B^h$ on $X_r^h \times \Sigma_h^\sigma \times \Sigma_h^\sigma$ defined by

\[
B^h(u^h, \sigma^h, \tau^h) := ((u^h \cdot \nabla)\sigma^h, \tau^h)_h + \frac{1}{2} (\nabla \cdot u^h \sigma^h, \tau^h) + (\sigma^h + \sigma^-^h, \tau^h)_h, \quad (3.1.9)
\]

Note that the second term vanishes when $\nabla \cdot u^h = 0$. On occasion it will be necessary to consider (3.1.9) without the associated jump term. Therefore, let

\[
\bar{B}^h(u^h, \sigma^h, \tau^h) := ((u^h \cdot \nabla)\sigma^h, \tau^h)_h + \frac{1}{2} (\nabla \cdot u^h \sigma^h, \tau^h).
\]

The discontinuous Galerkin finite element approximation of (3.1.6)–(3.1.8) is then as follows. Given $f \in H^{-1}(\Omega)$, find $(\sigma^h, u^h, p^h) \in \Sigma_h^\sigma \times X_r^h \times S_h$ such that

\[
(\sigma^h, \tau^h) + \lambda B^h(u^h, \sigma^h, \tau^h) + \lambda (g(u^h, \nabla u^h), \tau^h) - 2\alpha (D(u^h), \tau^h) = 0, \quad \forall \tau^h \in \Sigma_0^h,
\]

\[
(\sigma^h, D(v^h)) + 2(1-\alpha) (D(u^h), D(v^h)) - (p^h, \nabla \cdot v^h) = (f, v^h), \quad \forall v^h \in X_0^h,
\]

\[
(q^h, \nabla \cdot u^h) = 0, \quad \forall q^h \in S^h.
\]

Existence of a solution to the discrete problem (3.1.11)-(3.1.13) has been shown by Baranger and Sandri \cite{8} under the assumption that the continuous problem (3.1.1)-(3.1.5) has a
bounded solution $(\sigma, u, p) \in H^2_{\sigma, u_\Gamma} (\Omega) \times H^3_{u_\Gamma} (\Omega) \times H^2 (\Omega)$. The error estimates
\[
\|\sigma - \sigma^h\|_0 + \|\nabla (u - u^h)\|_0 \leq Ch^{3/2}, \quad \|p - p^h\|_0 \leq Ch^{3/2}
\]
for constant $C > 0$, are also proven in [8] under the same assumptions.

3.1.3 Newton Iteration

For ease of notation, let $\Pi^h = \Sigma^h_0 \times X^h_0 \times S^h$ and let $\Pi_{bc}^h = \Sigma_{\sigma, u_\Gamma}^h \times X_{u_\Gamma} \times S^h$, with $u = (\sigma^h, u^h, p^h) \in \Pi_{bc}^h$ and $v = (\tau^h, v^h, q^h) \in \Pi^h$. Also let $(\Pi^h)^*$ denote the dual space of $\Pi^h$. With $A$ as defined in (2.2.6), (3.1.11)-(3.1.13) can be written as
\[
A((\sigma^h, u^h), (\tau^h, v^h)) + \lambda B^h(u^h, \sigma^h, \tau^h)
+ \lambda (g_a(\sigma^h, \nabla u^h), \tau^h) - (p^h, \nabla \cdot v^h) + (g^h, \nabla \cdot u^h)
= 2\alpha(f, v^h) \quad \forall (\tau^h, v^h, q^h) \in \Pi^h. \quad (3.1.14)
\]

For the remainder of this chapter, the superscript $h$ will be dropped from the expression of $\sigma^h, u^h, p^h, \tau^h, v^h$, and it is assumed that these quantities are elements of the discrete approximation spaces. Define the function $G : \Pi_{bc}^h \to (\Pi^h)^*$ by
\[
\langle G(u), v \rangle := A((\sigma, u), (\tau, v)) + \lambda B(u, \sigma, \tau) + \lambda (g_a(\sigma, \nabla u), \tau)
- (p, \nabla \cdot v) + (q, \nabla \cdot u) - 2\alpha(f, v) \quad \forall v \in \Pi^h. \quad (3.1.15)
\]

One approach to solving $G(u) = 0$ is to solve the system of nonlinear equations (3.1.15) via Newton iteration. Let $G_u = \partial G/\partial u$ denote the Jacobian of $G$ with respect to the unknown $u$. Given an initial iterate $u^0 \in \Pi_{bc}^h$, for $i = 0, 1, 2, \ldots$, solve for $\delta u^i \in \Pi^h$ satisfying the linear system
\[
\langle G_u(u^i)(\delta u^i), v \rangle = \langle -G(u^i), v \rangle \quad \forall v \in \Pi^h, \quad (3.1.16)
\]
setting $u^{i+1} := u^i + \delta u^i$ until $\|u^{i+1} - u^i\| < tol$ for an appropriate choice of norm and tolerance $tol$. The Jacobian is formed using the Fréchet derivative of $G$ at $u^i$ in the direction of $w$, i.e.,
\[
\langle G_u(u^i)(w), v \rangle := \lim_{\varepsilon \to 0} \frac{\langle G(u^i + \varepsilon w), v \rangle - \langle G(u^i), v \rangle}{\varepsilon}.
\]
Combining (3.1.15) and (3.1.16), the Newton iteration consists of the following linear problem: For \( i = 0, 1, 2, \ldots \), solve for \( \mathbf{u}^{i+1} = (\sigma^{i+1}, u^{i+1}, p^{i+1}) \in \Pi_{be}^h \) satisfying

\[
A((\sigma^{i+1} - \sigma^i, u^{i+1} - u^i), (\tau, v)) + \lambda B^h(\sigma^i, u^{i+1} - u^i, \tau) + \lambda B^h(u^{i+1} - u^i, \sigma^i, \tau) \\
+ \lambda(g_a(\sigma^{i+1} - \sigma^i, \nabla u^i), \tau) + \lambda(g_a(\sigma^i, \nabla u^{i+1} - u^i), \tau) \\
- (p^{i+1} - p^i, \nabla \cdot v) + (q, \nabla \cdot u^{i+1} - u^i)
\]

\[
= -A((\sigma^i, u^i), (\tau, v)) - \lambda B^h(\sigma^i, u^i, \tau) - \lambda(g_a(\sigma^i, \nabla u^i), \tau) \\
+ (p^i, \nabla \cdot v) - (q, \nabla \cdot u^i) + 2\alpha(f, v) \quad \forall (\tau, v, q) = v \in \Pi^h. \quad (3.1.17)
\]

The term

\[
B^h(u^{i+1} - u^i, \sigma^i, \tau) = u^{i+1} - u^i \cdot \nabla \sigma^i, \tau) + \frac{1}{2}(\nabla \cdot (u^{i+1} - u^i)\sigma^i, \tau) + (\sigma^+ - \sigma^-, \tau^+)_{h,u^{i+1}-u^i}
\]

requires the calculation of the stress jump in the direction of an unknown velocity. However, the contribution of the jump term

\[
(\sigma^+ - \sigma^-, \tau^+)_{h,u^{i+1}-u^i} = \sum_{K \in \mathcal{T}_h} \int_{\partial K^-(u^{i+1} - u^i)} \left( (\sigma^+(u^{i+1} - u^i) - \sigma^-(u^{i+1} - u^i)) : \tau^+(u^{i+1} - u^i) \right) \\
\quad \cdot |\mathbf{n} \cdot (u^{i+1} - u^i)| \, ds,
\]

itself is assumed to be negligible, as \( u^{i+1} - u^i = \delta u^i \) is usually small. Thus the jump term is dropped from (3.1.18). With this adjustment we can write (3.1.17) as

\[
A((\sigma^{i+1}, u^{i+1}), (\tau, v)) + \lambda B^h(\sigma^i, u^{i+1}, \tau) + \lambda \tilde{B}^h(u^{i+1}, \sigma^i, \tau) \\
+ \lambda(g_a(\sigma^{i+1}, \nabla u^i), \tau^h) + \lambda(g_a(\sigma^i, \nabla u^{i+1})), \tau) - (p^{i+1}, \nabla \cdot v) + (q, \nabla \cdot u^{i+1}) \\
= \lambda \tilde{B}^h(u^i, \sigma^i, \tau) + \lambda(g_a(\sigma^i, \nabla u^i), \tau) + 2\alpha(f, v) \quad \forall (\tau, v, q) = v \in \Pi^h. \quad (3.1.19)
\]

This iteration is performed for fixed values of the parameters \( a, \alpha \), and the Weissenberg number \( \lambda \). In order for (3.1.19) to converge, the initial iterate \( \mathbf{u}^0 \) must be within the radius of convergence of the solution of the discrete problem \( G(\mathbf{u}) = 0 \).
At some point in the process of increasing \( \lambda \), a value \( \lambda^* \) will be encountered for which the nonlinear iteration fails for values of \( \lambda \) just beyond \( \lambda^* \). As described by Owens and Phillips [65], there are a number of possibilities for the behavior of solution curves of the continuous problem (3.1.1)-(3.1.3) for increasing Weissenberg number. Although the interest here is the behavior of the discrete system (3.1.11)-(3.1.13), the same possibilities apply. Some of these are shown in Figure 3.1.1. To visualize this, consider a simplified planar representation of the solution manifold of \( \mathbf{G} \) in the \((\mathbf{u}, \lambda)\) space (for sake of illustration consider \( \lambda \) on the horizontal axis). Figure 3.1(a) represents the situation in which no solutions exist for (3.1.1)-(3.1.3) beyond \( \lambda^* \). In Figure 3.1(b), there is a gap in the solution curve, i.e., a range of \( \lambda \) for which solutions do not exist. The presence of a turning point in Figure 3.1(c) may indicate multiple solutions for certain values of \( \lambda \), and a bifurcation point (shown in Figure 3.1(d)) is present when multiple branches of the solution originate from

![Figure 3.1.1](image-url)

Figure 3.1.1 Some possibilities for the behavior of solution curves for high Weissenberg number.
λ*. In addition, the possibility of a unique solution for all nonnegative λ exists, however the previously described Newton iteration still may encounter a critical value for various choices of the parameters α and α.

### 3.2 Continuation in Weissenberg Number

As discussed in Section 3, convergence of the nonlinear iteration becomes more problematic as the Weissenberg number is increased. The defect-correction method has been applied to steady-state viscoelastic flows [29, 27, 57] for high Weissenberg number. In their approach, the defect step consisted of a nonlinear iteration in which the Weissenberg number was replaced with an artificially reduced value, and the correction step sought to improve on the approximation found in the defect step.

Continuation methods [49, 2, 71] provide a means for stepping along solution manifolds for varying values of a problem parameter or group of parameters. In the context of viscoelastic flows, the behavior of computed solution manifolds in the Weissenberg parameter space is of great interest. This leads to considering the Weissenberg number λ as a prime candidate for continuation.

#### 3.2.1 Simple Continuation

In some manner, continuation methods try to ensure that the initial iterate $u^0$ is within the radius of convergence of the solution to the discrete problem for the current parameter value. The basic approach to the nonlinear iteration (3.1.19) is to use $u^0 = 0$ as an initial iterate, regardless of the value of λ that is being solved for. A slightly more advanced approach is to first compute a solution for $\lambda = 0$, i.e., solve the corresponding linear Stokes problem to find $u^0$. This method should provide a better initial approximation than $u^0 = 0$ for $\lambda > 0$. However, as $\lambda$ increases, nonlinear iterations will still fail to converge with a Stokes initial iterate.

Consider $G(u) = G(u, \lambda)$ to be the nonlinear system of equations arising from the discrete problem (3.1.15) for a particular value of λ. Let $u_0$ be the solution computed by the nonlinear iteration (3.1.19) for $\lambda = \lambda_0$. Then $u_0$ serves as a “good” choice of initial iterate for the problem $G(u, \lambda_1)$, where $\lambda_1 = \lambda_0 + \Delta \lambda_0$ for some $\Delta \lambda_0 > 0$. This process...
can be thought of as a simple continuation in $\lambda$. Computations proceed along the solution manifold in the $(\lambda, u)$ space by incrementing $\lambda$ after each convergent nonlinear iteration, using the $u$ computed at a point on the curve as the initial iterate for the next larger value of $\lambda$. This process is described in Algorithm 3.2.1.

**Algorithm 3.2.1 (Simple Continuation in $\lambda$)** Let $(u_0, \lambda_0) \in \Pi_{bc}^h \times \mathbb{R}$ solve $G(u, \lambda) = 0$.

For $j = 0, 1, \ldots$, do

1. Determine the step length $\Delta \lambda_j$.
2. Set $u_{j+1}^0 = u_j$ and $\lambda_{j+1} = \lambda_j + \Delta \lambda_j$.
3. Solve $G(u_{j+1}, \lambda_{j+1}) = 0$ by the iteration: For $i = 0, 1, \ldots$, solve for $(u_{j+1}^i, \lambda_{j+1}) \in \Pi_{bc}^h \times \mathbb{R}$ satisfying

   \begin{align*}
   &A((\sigma_{j+1}^{i+1}, u_{j+1}^{i+1}, \tau, v) + \lambda_{j+1} B^h(u_{j+1}^{i+1}, \sigma_{j+1}^{i+1}, \tau) + \lambda_{j+1} \tilde{B}^h(u_{j+1}^{i+1}, \sigma_{j+1}^{i}, \tau) \\
   + \lambda_{j+1}(g_a(\sigma_{j+1}^{i+1}, \nabla u_{j+1}^{i+1}, \tau) + \lambda_{j+1}(g_a(\sigma_{j+1}^{i}, \nabla u_{j+1}^{i+1}, \tau) - (p_{j+1}, \nabla \cdot v) + (q, \nabla \cdot u_{j+1}^{i+1}) \\
   = \lambda_{j+1} \tilde{B}^h(u_{j+1}^{i+1}, \sigma_{j+1}^{i+1}, \tau) + \lambda_{j+1}(g_a(\sigma_{j+1}^{i}, \nabla u_{j+1}^{i}, \tau)) + 2\alpha(f, v) - (\tau, v, q) \in \Pi^h.
   \end{align*}

4. Go to Step 1.

The simple continuation approach can be effective in providing the Newton iteration for $G(u_j, \lambda_j) = 0$ with a good initial approximation. Figure 3.2.1 gives a plot of the solution manifold and a solution $(u_j, \lambda_j)$ along the curve. Note that in order for $u_j$ to be a satisfactory initial iterate for $\lambda_{j+1}$, the increment $\Delta \lambda_j$ must be chosen in such a way to ensure that $u_j$ is within the radius of convergence of the nonlinear operator (2.3.10)-(2.3.12).

A basic method for choosing the steplengths is to start with a moderate value for $\Delta \lambda_j$ and attempt the nonlinear iteration. If the iteration succeeds, then either set $\Delta \lambda_{j+1} := \Delta \lambda_j$ or $\Delta \lambda_{j+1} := \gamma \Delta \lambda_j$ for some $\gamma > 1$. If the iteration fails, then set $\Delta \lambda_{j+1} := \gamma \Delta \lambda_j$ for some $\gamma < 1$ and reattempt the nonlinear iteration. Once the steplength falls below a specified tolerance, the continuation procedure will terminate.

### 3.2.2 Natural Continuation

The simple continuation procedure may encounter difficulty if the point $(u_j, \lambda_j)$ is not sufficiently close to the point $(u_{j+1}, \lambda_{j+1})$. This can occur when attempting a step
of too large in magnitude, or when the solution curve experiences large changes in \( u \) for moderate changes in \( \lambda \) (high slope). In addition, the simple continuation process can be inefficient if there is only moderate change in \( u \) for a significant range of \( \lambda \) (small slope).

Forming a predicted value based upon the slope of the solution curve at the point \((u_j, \lambda_j)\) can provide a better initial iterate for the subsequent nonlinear iteration. This slope can be found by computing the quantity \( \frac{\partial u}{\partial \lambda} \) at \( \lambda_j \). Assume \( G \) is continuously differentiable in \( u \) and \( \lambda \), and \( u \) is continuously differentiable in \( \lambda \). Then, for \( G(u, \lambda) = 0 \) at \((u_j, \lambda_j)\), we have from the chain rule

\[
\left( \frac{\partial G}{\partial u} \right)_{(u_j, \lambda_j)} \left( \frac{\partial u}{\partial \lambda} \right)_{\lambda_j} + \frac{\partial G}{\partial \lambda} \bigg|_{(u_j, \lambda_j)} = 0 \quad \text{in } \Omega,
\]

or

\[
\left( G_u(u_j, \lambda_j) \right) \left( \frac{\partial u}{\partial \lambda} \right)_{\lambda_j}, v \right) = -\left( G_\lambda(u_j, \lambda_j), v \right) \quad \forall v \in \Pi^h. \quad (3.2.1)
\]
where $G_\lambda$ denotes the Fréchet derivative of $G$ with respect to the parameter $\lambda$. It is easy to see that

$$
\langle G_\lambda(u_j, \lambda_j), v \rangle = B^h(u_j, \sigma_j, \tau) + (g_a(\sigma_j, \nabla u_j), \tau) \quad \forall (\tau, v, q) = v \in \Pi^h.
$$

Once the tangent slope $\partial u / \partial \lambda$ at $\lambda_j$ has been found (denote this by $\hat{u}_j = (\hat{\sigma}_j, \hat{u}_j, \hat{p}_j)$), then for some steplength $\Delta \lambda_j$, set

$$
u_{j+1}^0 := u_j + \Delta \lambda_j \left. \frac{\partial u}{\partial \lambda} \right|_{\lambda_j} = u_j + \Delta \lambda_j \hat{u}_j,
$$
as the initial iterate for $\lambda_{j+1} := \lambda_j + \Delta \lambda_j$. The standard nonlinear iteration (3.1.19) can then be performed to approximate $(u_{j+1}, \lambda_{j+1})$. This procedure is described in Algorithm 3.2.2, and one step of the method is illustrated in Figure 3.2.2.

**Algorithm 3.2.2 (Natural Continuation in $\lambda$)** Let $(u_0, \lambda_0) \in \Pi_{bc}^h \times \mathbb{R}$ solve $G(u, \lambda) = 0$. For $j = 0, 1, \ldots$, do

1. Solve the linear problem (3.2.1) for $\hat{u}_j \in \Pi^h$ by:

   $$
   A((\hat{\sigma}_j, \hat{u}_j), (\tau, v)) + \lambda_j\tilde{B}^h(\hat{u}_j, \sigma_j, \tau) + \lambda_jB^h(u_j, \hat{\sigma}_j, \tau) + \lambda_j(g_a(\sigma_j, \nabla \hat{u}_j), \tau) + \lambda_j(g_a(\hat{\sigma}_j, \nabla u_j), \tau) - (\hat{p}_j, \nabla \cdot v) + (q, \nabla \cdot \hat{u}_j)
   $$

2. Determine the step length $\Delta \lambda_j$.

3. Set $\nu_{j+1}^0 = u_j + \Delta \lambda_j \hat{u}_j$ and $\lambda_{j+1} = \lambda_j + \Delta \lambda_j$.

4. Solve $G(u_{j+1}, \lambda_{j+1}) = 0$ by the iteration: For $i = 0, 1, \ldots$, solve for $(u_{j+1}^{i+1}, \lambda_{j+1}) \in \Pi_{bc}^h \times \mathbb{R}$ satisfying

   $$
   A((\sigma_{j+1}^{i+1}, \nu_{j+1}^{i+1}), (\tau, v)) + \lambda_{j+1}B^h(u_{j+1}^{i+1}, \sigma_{j+1}^{i+1}, \tau) + \lambda_{j+1}(g_a(\sigma_{j+1}^{i+1}, \nabla u_{j+1}^{i+1}), \tau) - (\nu_{j+1}^{i+1}, \nabla \cdot v) + (q, \nabla \cdot u_{j+1}^{i+1})
   $$

5. Go to Step 1.

This natural continuation procedure is sometimes described as Euler-Newton continuation [14], as the process of determining the predictor $u_{j+1}^0$ is similar in nature to the forward Euler method for ODEs, and a Newton iteration is utilized to solve the nonlinear
Algorithm 3.2.2 differs from Algorithm 3.2.1 with the addition of solving a linear problem for $\hat{u}_j$. It is expected that the $u^0_{j+1}$ found by natural continuation is a better initial iterate than merely the solution $u_j$, as used in the simple continuation algorithm. As a result, natural continuation is expected to be able to compute solutions for a larger range of Weissenberg number than simple continuation. In addition, the natural continuation approach may allow for a more aggressive steplength strategy than simple continuation. Den Heijer and Rheinboldt [24] derive some sophisticated algorithms for steplength selection for generalized natural continuation methods. Alternate predictor strategies may be used as well. For example, one may use a second-order Taylor series approximation to form $u^0$. Gunzburger and Peterson [43] showed that in the Navier-Stokes equations, for some cases the stepsizes in Reynolds number may be chosen independently of the type of predictor used.
A detailed discussion of steplength algorithms for viscoelastic flows is a topic to be analyzed in future work.

3.3 Turning Points and Pseudo-arclength Continuation

The natural continuation process can suffer difficulty at points along the solution manifold where the slope $\partial u / \partial \lambda$ is undefined, i.e., at turning points or singular points along the solution curve. In this case, the Jacobian $G_u$ will be singular and the linear problem (3.2.1) will not have a unique solution.

3.3.1 Parametrization with respect to arc length

As the natural continuation algorithm will fail near turning points, a different continuation approach is needed. Keller [48, 49] proposed the reparametrization of the solution curve with an arclength (or arclength-like) parameter. The description of the method presented here is similar to those found in [11, 15, 22, 39, 62] and others. In a manner similar to natural continuation, first a tangent to the solution curve is found and a predicted value that lies on the tangent is computed. Then an iterative procedure attempts to reconcile the predicted value back to the solution curve. The arclength parametrization requires additional information in the form of an equation describing the arclength, but allows for a more robust iteration that can proceed beyond turning points.

Let $\vartheta$ be a parameter describing the arc length of the solution manifold in the $(u, \lambda)$-space. Then we have

$$\left\| \frac{\partial u}{\partial \vartheta} \right\|^2 + \left( \frac{\partial \lambda}{\partial \vartheta} \right)^2 = 1,$$

where the norm in (3.3.1) is appropriate for $u$. For $u \in \Pi^h$, this norm will be

$$\|u\|^2 = \|\sigma\|^2 + \|u\|^2_1 + \|p\|^2_0.$$

Let $s$ represent an arclength-like parameter and consider $u(s)$ and $\lambda(s)$ to be functions of $s$. Let $\dot{u} = \partial u / \partial s$ and $\dot{\lambda} = \partial \lambda / \partial s$, and assume $(u_j, \lambda_j) = (u(s_j), \lambda(s_j))$ is a solution of $G(u, \lambda) = 0$. A unit tangent vector $[\dot{u}_j, \dot{\lambda}_j]^T$ to the curve is computed by solving

$$G_u \dot{u}_j + G_\lambda \dot{\lambda}_j = 0,$$

(3.3.2)
together with (3.3.1) above. This can be accomplished by first solving the linear system
\[
\langle G_u(u_j, \lambda_j) \hat{u}_j, v \rangle = -\langle G_\lambda(u_j, \lambda_j), v \rangle \quad \forall v \in \Pi^h,
\]
for \( \hat{u}_j \). Then \( \hat{u}_j \) and \( \hat{\lambda}_j \) can be determined by
\[
\dot{\lambda}_j = \pm \frac{1}{\sqrt{1 + \|\hat{u}_j\|^2}} ,
\]
\[
\dot{u}_j = \hat{\lambda}_j \hat{u}_j .
\]
Note that (3.3.4) and (3.3.5) imply \( \|\dot{u}_j\|^2 + |\dot{\lambda}_j|^2 = 1 \). The sign in (3.3.4) is chosen such that the angle between successive tangents on the curve is no less than 0 and no more than \( \pi/2 \). This results in a construction of the solution manifold that moves “forward” with respect to the parameter \( s [11, 22] \). To determine the correct sign, given two solutions \((u_{j-1}, \lambda_{j-1})\) and \((u_j, \lambda_j)\), compute the quantity
\[
\langle \dot{u}_j, u_j - u_{j-1} \rangle + \dot{\lambda}_j (\lambda_j - \lambda_{j-1}) ,
\]
where \( \langle \cdot, \cdot \rangle \) in (3.3.6) is the inner product that induces the norm in (3.3.1). If (3.3.6) is positive, then the choice of \((\dot{u}_j, \dot{\lambda}_j)\) means that the computation will proceed in the same direction as it did from \((u_{j-1}, \lambda_{j-1})\) to \((u_j, \lambda_j)\). If (3.3.6) is negative, the opposite sign should be chosen in (3.3.4).

Once the appropriate tangent direction has been chosen and the arclength parameter increment \( \Delta s_j = s_{j+1} - s_j \) is set, Euler predictors of
\[
\begin{align*}
u_{j+1}^0 &= u_j + \Delta s_j \hat{u}_j \\
\lambda_{j+1}^0 &= \lambda_j + \Delta s_j \hat{\lambda}_j
\end{align*}
\]
are chosen as the initial iterates for the nonlinear iteration. Note that, as \( \lambda_{j+1} = \lambda(s_{j+1}^i) \) is a function of the arclength parameter, its value may vary during the nonlinear iteration. This is precisely the flexibility that will allow computation beyond singular points.

To complete the nonlinear system of equations, \( G(u(s), \lambda(s)) = 0 \) is augmented with a suitable arclength condition or constraint given by some \( N(u(s), \lambda(s), s) = 0 \). This constraint is based upon the arclength equation (3.3.1). In practice, an approximation
to (3.3.1), such as a linearization, is used. Thus \( s \) is referred to as a pseudo-arclength parameter instead of the actual arc length. The discussion of the details of the pseudo-arclength constraints has been postponed until Section 3.3.2, however it should be noted that \( N \) must be continuously differentiable in both \( \mathbf{u} \) and \( \lambda \). Thus, the nonlinear system of equations to be solved is

\[
\begin{bmatrix}
G(\mathbf{u}(s), \lambda(s)) \\
N(\mathbf{u}(s), \lambda(s), s)
\end{bmatrix} = 0,
\]

(3.3.9)

and this can be accomplished by a Newton iteration of the form

\[
\begin{bmatrix}
G_u(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) & G_\lambda(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) \\
N_u(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) & N_\lambda(\mathbf{u}^i_{j+1}, \lambda^i_{j+1})
\end{bmatrix}
\begin{bmatrix}
\delta \mathbf{u}^i \\
\delta \lambda^i
\end{bmatrix} = -
\begin{bmatrix}
G(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) \\
N(\mathbf{u}^i_{j+1}, \lambda^i_{j+1})
\end{bmatrix}.
\]

(3.3.10)

The linear system in (3.3.10) has been shown to be nonsingular at turning points of the solution manifold [48]. The above procedure is summarized in Algorithm 3.3.1.

**Algorithm 3.3.1 (Pseudo-arclength Continuation)** Let \((\mathbf{u}_{-1}, \lambda_{-1}) \in \mathbf{\Pi}_bc^h \times \mathbb{R}\) and \((\mathbf{u}_0, \lambda_0) \in \mathbf{\Pi}_bc^h \times \mathbb{R}\) solve \(G(\mathbf{u}, \lambda) = 0\). For \(j = 0, 1, \ldots\), do

1. Solve the linear problem (3.3.3) for \(\tilde{\mathbf{u}}_j \in \mathbf{\Pi}^h\) by:

\[
A((\hat{\sigma}_j, \hat{\mathbf{u}}_j), (\tau, \nu)) + \lambda_j \hat{B}^h(\hat{\mathbf{u}}_j, \hat{\sigma}_j, \tau) + \lambda_j B^h(\mathbf{u}_j, \hat{\sigma}_j, \tau) + \lambda_j (g_a(\hat{\sigma}_j, \nabla \hat{\mathbf{u}}_j), \tau) + (\hat{p}_j, \nabla \hat{\mathbf{u}}_j) - (\hat{p}_j, \nabla \cdot \hat{\mathbf{u}}_j)
\]

\[
= -B^h(\mathbf{u}_j, \sigma_j, \tau) - (g_a(\sigma_j, \nabla \mathbf{u}_j), \tau) \forall (\tau, \nu, q) \in \mathbf{\Pi}^h.
\]

2. Set \(\lambda_j = (1 + ||\tilde{\mathbf{u}}_j||^2)^{-1/2}\) and \(\mathbf{u}_j = \lambda_j \tilde{\mathbf{u}}_j\).

3. Compute \(\omega := \langle \tilde{\mathbf{u}}_j, \mathbf{u}_j - \mathbf{u}_{j-1} \rangle + \hat{\lambda}_j (\lambda_j - \lambda_{j-1})\). If \(\omega < 0\), set \(\hat{\lambda}_j = -(1 + ||\tilde{\mathbf{u}}_j||^2)^{-1/2}\) and \(\tilde{\mathbf{u}}_j = \hat{\lambda}_j \tilde{\mathbf{u}}_j\).

4. Determine the pseudo-arclength stepsize \(\Delta s_j\).

5. Set \(\mathbf{u}^0_{j+1} = \mathbf{u}_j + \Delta s_j \tilde{\mathbf{u}}_j\) and \(\lambda_{j+1} = \lambda_j + \Delta s_j \hat{\lambda}\).

6. Solve \([G(\mathbf{u}_{j+1}(s), \lambda_{j+1}(s)), N(\mathbf{u}_{j+1}(s), \lambda_{j+1}(s), s)]^T = 0\) by the iteration: For \(i = 0, 1, \ldots\), solve the linear system (3.3.10) for \((\delta \mathbf{u}^i, \delta \lambda^i) \in \mathbf{\Pi}^h \times \mathbb{R}\) satisfying

\[
\begin{bmatrix}
G_u(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) & G_\lambda(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) \\
N_u(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) & N_\lambda(\mathbf{u}^i_{j+1}, \lambda^i_{j+1})
\end{bmatrix}
\begin{bmatrix}
\delta \mathbf{u}^i \\
\delta \lambda^i
\end{bmatrix} = -
\begin{bmatrix}
G(\mathbf{u}^i_{j+1}, \lambda^i_{j+1}) \\
N(\mathbf{u}^i_{j+1}, \lambda^i_{j+1})
\end{bmatrix},
\]

with \(\mathbf{u}^{i+1}_{j+1} := \mathbf{u}^i_{j+1} + \delta \mathbf{u}^i\) and \(\lambda^{i+1}_{j+1} := \lambda^i_{j+1} + \delta \lambda^i\).

7. Go to Step 1.
The block Jacobians $G_u$ and $G_\lambda$ are as described in the natural continuation algorithm, while the vectors $N_u$ and $N_\lambda$ will be described in Section 3.3.2. Note that each pass of Algorithm 3.3.1 does not terminate with a predetermined value for $\lambda_{j+1}$.

3.3.2 Pseudo-arclength constraints

As stated in Section 3.3.1, the pseudo-arclength constraint $N(u(s), \lambda(s), s)$ must be continuously differentiable with respect to $u$ and $\lambda$. There are several choices for $N$ that serve as good defining functions for $s$. However, $N$ must contain some characteristic of the solution that is to be measured along the solution curve.

Orthogonal Constraint

Presented by Keller [48], the most commonly used pseudo-arclength constraint is derived from a linear approximation to (3.3.1). This condition is given by

$$N_1(u(s), \lambda(s), s) := \langle \dot{u}_j, (u(s) - u(s_j)) \rangle + \dot{\lambda}_j (\lambda(s) - \lambda(s_j)) - (s - s_j) = 0, \quad (3.3.11)$$

and it requires that successive solution iterates lie on the hyperplane orthogonal to the tangent vector $[\dot{u}_j, \dot{\lambda}_j]^T$ and at a distance of $\Delta s_j$ from the solution $(u(s_j), \lambda(s_j)) = (u_j, \lambda_j)$. Figure 3.3.1 gives a graphical representation of the tangent $[\dot{u}_j, \dot{\lambda}_j]^T$ and the iterates satisfying $N_1$. Because of the nature of the geometry of the iterates, $N_1$ will be referred to as an “orthogonal” constraint. The orthogonality can be seen as follows. Note that from the structure of the predicted values (3.3.7) and (3.3.8), we have

$$\dot{u}_j = \frac{1}{\Delta s_j} (u_{j+1}^0 - u_j) \quad \text{and} \quad \dot{\lambda}_j = \frac{1}{\Delta s_j} (\lambda_{j+1}^0 - \lambda_j). \quad (3.3.12)$$

Now $N_1(u_{j+1}^i, \lambda_{j+1}^i, s_{j+1})$ and (3.3.12) yield the relation

$$\langle (u_{j+1}^0 - u_j), (u_{j+1}^i - u_j) \rangle + (\lambda_{j+1}^0 - \lambda_j) (\lambda_{j+1}^i - \lambda_j) = (\Delta s_j)^2,$$

or,

$$\langle (u_{j+1}^0 - u_j), (u_{j+1}^i - u_{j+1}^0) \rangle + \langle (u_{j+1}^0 - u_j), (u_{j+1}^0 - u_j) \rangle + (\lambda_{j+1}^0 - \lambda_j) (\lambda_{j+1}^i - \lambda_{j+1}^0) + (\lambda_{j+1}^0 - \lambda_j) (\lambda_{j+1}^0 - \lambda_j) = (\Delta s_j)^2. \quad (3.3.13)$$
From Figure 3.3.1 it is easy to see that

\[(\Delta s_j)^2 = \|u^{0}_{j+1} - u^i_j\|^2 + |\lambda^{0}_{j+1} - \lambda^i_j|^2, \tag{3.3.14}\]

and thus, subtracting (3.3.14) from (3.3.13) we have

\[\langle (u^{0}_{j+1} - u^i_j), (u^{0}_{j+1} - u^0_{j+1}) \rangle + (\lambda^{0}_{j+1} - \lambda^i_j) \lambda^{0}_{j+1} - \lambda^i_j = 0, \tag{3.3.15}\]

which implies that the vectors

\[
\begin{pmatrix}
  u^{0}_{j+1} - u^i_j \\
  \lambda^{0}_{j+1} - \lambda^i_j
\end{pmatrix}
\]

and

\[
\begin{pmatrix}
  u^{0}_{j+1} - u^0_{j+1} \\
  \lambda^{0}_{j+1} - \lambda^0_{j+1}
\end{pmatrix}
\]

are orthogonal. For implementation in Algorithm 3.3.1, the derivatives \(N_{1,u}\) and \(N_{1,\lambda}\) evaluated at \((u^i_{j+1}, \lambda^i_{j+1}, s)\) are necessary. These are given as

\[N_{1,u}(u^i_{j+1}, \lambda^i_{j+1}, s) = \dot{u}_j \quad \text{and} \quad N_{1,\lambda}(u^i_{j+1}, \lambda^i_{j+1}, s) = \dot{\lambda}_j.\]
Spherical Constraint

Another constraint presented by Keller [48] enforces successive iterates to be a particular distance from the previous solution. This constraint is of the form

\[ N_2(u(s), \lambda(s), s) := \|u(s) - u(s_j)\|^2 + |\lambda(s) - \lambda(s_j)|^2 - (s - s_j)^2 = 0, \quad (3.3.16) \]

and it requires that successive iterates lie on the sphere centered at \((u_j, \lambda_j)\) of radius \(s - s_j\) and will thus be referred to as a "spherical" constraint. Figure 3.3.2 gives a graphical representation of the tangent \([\dot{u}_j, \dot{\lambda}_j]^T\) and the iterates satisfying \(N_2\). The derivatives of \(N_2\) are found to be

\[ N_{2,u}(u_{j+1}^i, \lambda_{j+1}^i, s) = 2 \left( u_{j+1}^i - u_j^i \right), \quad (3.3.17) \]
\[ N_{2,\lambda}(u_{j+1}^i, \lambda_{j+1}^i, s) = 2 |\lambda_{j+1}^i - \lambda_j^i|. \quad (3.3.18) \]

This constraint is less frequently used than \(N_1\) in pseudo-arclength continuation descriptions and implementations [11, 14, 15, 25, 74, 62], perhaps due to the required recomputation of

Figure 3.3.2 Illustration of the pseudo-arclength condition \(N_2\).
(3.3.17) and (3.3.18) for each linear iteration $i$. The orthogonal constraint requires no such recomputation as the derivatives $N_1 \mathbf{u}$ and $N_1 \lambda$ do not vary with respect to $i$.

**Constraint Weighting**

Both (3.3.11) and (3.3.16) can be modified by employing a weighting parameter for the $\mathbf{u}$ and $\lambda$ terms. This weighting modifies the geometry of successive iterates by either changing their angle with respect to the tangent or reshaping the ellipsoid on which the iterates lie. The weighting allows for a greater contribution from either the $\mathbf{u}$ term or the $\lambda$ term, and enhances the flexibility of the pseudo-arclength continuation method. Specifically, for $0 \leq \theta \leq 2$, weighted constraints can be written as

$$
\tilde{N}_1(\mathbf{u}(s), \lambda(s), s) = \theta \langle \dot{\mathbf{u}}_j, (\mathbf{u}(s) - \mathbf{u}(s_j)) \rangle + (2 - \theta) \dot{\lambda}_j (\lambda(s) - \lambda(s_j)) - (s - s_j), \quad (3.3.19)
$$

$$
\tilde{N}_2(\mathbf{u}(s), \lambda(s), s) = \theta \| \mathbf{u}(s) - \mathbf{u}(s_j) \|^2 + (2 - \theta) |\lambda(s) - \lambda(s_j)|^2 - (s - s_j)^2. \quad (3.3.20)
$$

**3.3.3 Combined Methods**

The natural and pseudo-arclength continuation methods described above can be combined to create a powerful and flexible solution approach for large-scale problems, as was done by de Almeida and Derby [22] for the Navier-Stokes equations. The less-expensive natural continuation can be used by default and the procedure can switch to pseudo-arclength continuation when a region of high curvature of the solution manifold is encountered. To avoid the high cost of computing the curvature, the implementation described in [22] uses the angle between successive tangents on the curve to determine if the curvature threshold has been met.

**3.4 Numerical experiments**

To investigate the performance of the various continuation algorithms discussed in Sections 3.2 and 3.3, the algorithms have been implemented using the finite element software package **FreeFem++** [45] in 2-d. Linear systems are solved using the **UMFPACK** solver [21]. As described in Section 3.1, continuous piecewise quadratic elements are used for velocity, continuous piecewise linears are used for pressure, and discontinuous piecewise linears are used for stress.
3.4.1 Four-to-one Contraction Flow

As in Section 2.5, the case of planar flow through a contraction geometry with a ratio of 4:1 with respect to upstream and downstream channel widths is considered. The contraction angle is a fixed \(3\pi/2\) and the channel lengths are sufficiently long to impose a fully developed Poiseuille flow in the inflow and outflow channels. The geometry of the computational domain is illustrated in Figure 2.5.3 of Section 2.5. On this domain the velocity boundary conditions are given by (2.5.1). Boundary conditions for \(\sigma\) are specified on the inflow boundary by (2.5.2). Symmetry conditions are imposed on the bottom of the computational domain. The parameter \(\alpha\) is set to 8/9. Computations were performed on two different meshes, M1 and M2. Table 2.5.3 of Section 2.5 lists the characteristics of the meshes.

3.4.2 High Weissenberg Number Results

Of particular interest is the behavior of the solution manifold for large values of \(\lambda\). To investigate this, Algorithms 3.2.1 and 3.2.2 are run from a starting value of \(\lambda = 0\) and \(\mathbf{u}^0 = 0\) on each computational mesh for selected values of the \(a\) parameter. Algorithm 3.3.1 (with the \(N_2\) constraint) requires two initial solutions, the solution above is used as well as the solution computed at \(\lambda = 1\). The initial step length in \(\lambda\) or \(s\) is set to 1, and upon failure of the nonlinear iteration, the steplength is reduced by half and the iteration reattempted. The continuation process terminates when the steplength falls below \(10^{-6}\). As a comparison, results are given from a “no continuation” approach of using \(\mathbf{u}^0 = 0\) as the initial iterate for all values of \(\lambda\). A stopping criterion of

\[
\|\mathbf{u}^i - \mathbf{u}^{i-1}\|_\infty \leq 10^{-8}
\]

is used for the nonlinear iteration, and the nonlinear iteration was terminated if the stopping criterion had not been satisfied after 200 iterations.

Table 3.4.1 gives the maximum \(\lambda\) values for each solution approach. All of the continuation methods vastly increase the range of \(\lambda\) for which solutions can be computed over using \(\mathbf{u}^0 = 0\) for all values of \(\lambda\). It is also observed that while simple and natural continuation produced similar high Weissenberg number limits, pseudo-arclength continuation was
Table 3.4.1  Maximum values of $\lambda$ obtained by the different solution methods, selected values of $a$.

<table>
<thead>
<tr>
<th>Mesh</th>
<th>$a$</th>
<th>No Continuation</th>
<th>Simple Continuation</th>
<th>Natural Continuation</th>
<th>Pseudo-arclength Continuation ($N_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.0</td>
<td>2.430</td>
<td>11.1279</td>
<td>11.1279</td>
<td>11.1291</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>1.286</td>
<td>18.1992</td>
<td>18.2187</td>
<td>18.3423</td>
</tr>
<tr>
<td></td>
<td>-1.0</td>
<td>1.465</td>
<td>8.9040</td>
<td>8.9399</td>
<td>8.9560</td>
</tr>
<tr>
<td>M2</td>
<td>1.0</td>
<td>1.821</td>
<td>9.3370</td>
<td>9.3460</td>
<td>9.3467</td>
</tr>
<tr>
<td></td>
<td>0.0</td>
<td>0.934</td>
<td>9.0184</td>
<td>9.0234</td>
<td>9.1348</td>
</tr>
<tr>
<td></td>
<td>-1.0</td>
<td>1.321</td>
<td>9.1563</td>
<td>9.3007</td>
<td>9.4050</td>
</tr>
</tbody>
</table>

able to exceed those limits in all cases. However, the amounts by which pseudo-arclength continuation surpassed the other methods could be considered insignificant. Plots of the solution norm $\|u_j\|$, where

$$\|u_j\|^2 = \|\sigma\|^2_0 + \|u\|^2_1 + \|p\|^2_0,$$

computed from the pseudo-arclength continuation method for both meshes are shown in Figure 3.4.1. In terms of the computational effort required by each algorithm, the simple and natural continuation methods require less work than the pseudo-arclength method. However, while using the same steplength selection and reduction strategy, the pseudo-arclength continuation method requires far fewer steps and steplength reductions to reach its maximum value of $\lambda$. The number of steps and steplength reductions required for each method for computations with $a = 1$ on mesh M1 are reported in Table 3.4.2.

The behavior of computed approximations for increasing Weissenberg number is of great interest, in particular for the Oldroyd-B constitutive model ($a = 1$). In Figure 3.4.2, the horizontal velocity of the fluid computed on mesh M2 along the line of symmetry (the center of the computational domain) is shown for several values of $\lambda$, including the limiting value.
Figure 3.4.1 Solution norm computed by pseudo-arclength continuation.

<table>
<thead>
<tr>
<th>Method</th>
<th>max $\lambda$</th>
<th>Total # of steps</th>
<th># of steplength reductions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>11.1279</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>Natural</td>
<td>11.1279</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>Pseudo-arclength</td>
<td>11.1291</td>
<td>12</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 3.4.2 Number of steps and steplength reductions required for $a = 1$ on mesh M1.

It is observed that, as the value of $\lambda$ increases, the magnitude of the velocity tends to increase just beyond the contraction of the domain, and fluctuates downstream. This may
suggest that for highly elastic fluids, as the fluid enters the contraction channel, the elastic features of the fluid cause a speed-up/slow-down behavior. These observations have been seen previously in numerical experiments by Keunings and Crochet [51] for the Phan Thien-Tanner model. This behavior could lead to a loss of stability of solutions at higher values of $\lambda$; one hypothesis is that no steady-state solutions (only temporally unsteady solutions) exist beyond some critical value of $\lambda$. The same observation was also made during numerical experiments with a defect-correction method for viscoelasticity [27]. Trebotich, et al. [77] also observed this situation and suggest the wave-like behavior is related to the elastic Mach number.

As discussed in [69], there is a stress singularity at the reentrant corner of contraction flows. The inability of the nonlinear solver to converge for values of $\lambda$ greater than 9.3466 ($a = 1$, mesh M2) may be due to the steep stress gradients that develop near the reentrant corner. Figure 3.4.3 is a plot of the $\sigma_{11}$ component of the stress along the horizontal line $y = 0.24$ (the contraction occurs at $y = 0.25$). The plot shows that for increasing $\lambda$, the

Figure 3.4.2 Horizontal velocity $u_1$ along the line of symmetry for increasing $\lambda$. 
maximum value of $\sigma_{11}$ near the corner increases considerably. Figure 3.4.4 is a plot of the velocity field for $\lambda = 9.3466$ ($a = 1$, mesh M2).

![Plot of stress component $\sigma_{xx}$ along line $y = 0.24$, mesh M2, $a = 1$](image)

Figure 3.4.3 Stress component $\sigma_{11}$ along a line near the reentrant corner for increasing $\lambda$.

![Velocity field and speed contour, $\lambda = 9.3466$, $a = 1$, mesh M2.](image)

Figure 3.4.4 Velocity field and speed contour, $\lambda = 9.3466$, $a = 1$, mesh M2.
3.4.3 Pseudo-Arclength Constraints

Also of interest is how the pseudo-arclength constraints \( N_1 \) and \( N_2 \), defined in (3.3.11) and (3.3.16) respectively, performed relative to each other. As mentioned in Section 3.3.2, \( N_2 \) is less often used in reported implementations, perhaps due to its need for recomputation at each iteration. One objective in studying \( N_2 \) is to learn if the geometry specified by the constraint has an impact on its convergence behavior and efficiency. As \( N_2 \) requires successive iterates to lie on a sphere of radius \( \Delta s_j \), centered at \((u_j, \lambda_j)\), successive iterates will sweep out an arc in the \((u, \lambda)\)-space. This structure may be more efficient than the orthogonality required by \( N_1 \) in regions of high curvature of the solution manifold, in particular by attaining convergence with a larger stepsize than is possible for \( N_1 \).

To compare the performance of each constraint in a region of low or moderate curvature, computations were performed using both approaches for mesh M1 and \( a = 1 \), starting with solutions computed at \( \lambda_{-1} = 0 \) and \( \lambda_0 = 1.0 \). The initial steplength was set to \( \Delta s_0 = 1.0 \) and computations were stopped once the continuation process had reached \( \lambda = 11.0 \). Table 3.4.3 lists performance statistics for both approaches during these computations. The performance of the two methods is very comparable from \( \lambda = 1.0 \) to \( \lambda = 11.0 \) - both

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Final ( \lambda )</th>
<th># of Steps</th>
<th># of Steplength Reductions</th>
<th>Average # of Iterations/Step</th>
<th>Average CPU Seconds/Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_1 )</td>
<td>11.04878019</td>
<td>14</td>
<td>3</td>
<td>26.7</td>
<td>113.7</td>
</tr>
<tr>
<td>( N_2 )</td>
<td>11.07273844</td>
<td>14</td>
<td>2</td>
<td>26.7</td>
<td>114.5</td>
</tr>
</tbody>
</table>

Table 3.4.3 Performance statistics for pseudo-arclength constraints to pass \( \lambda = 11.0 \), \( \Delta s_0 = 1.0 \).

methods required, the same number of steps and, on average the same number of iterations per step. A slightly larger average CPU time for \( N_2 \) may reflect the recomputation of \( N_{2,u} \) and \( N_{2,\lambda} \) required at each step.

To examine the behavior of the constraints in a region of high curvature, computations were performed for mesh M1 and \( a = 1 \), starting with solutions computed at
\( \lambda_{-1} = 11.1 \) and \( \lambda_0 = 11.124 \). The initial steplength was set to \( \Delta s_0 = 0.01 \). Table 3.4.4 lists performance statistics for both approaches. From Table 3.4.4, it is evident that the spherical constraint \( N_2 \) is much more efficient than the orthogonal constraint \( N_1 \), as \( N_2 \) required only one step of length 0.01 to exceed the target value, while \( N_1 \) required 25, with 8 additional convergence failures.

The same experiment was repeated with an initial steplength of \( \Delta s_0 = 0.0001 \). Table 3.4.5 lists performance statistics for each constraint. Again we see that far fewer steps, steplength reductions, and average iterations are required for \( N_2 \) to reach the same value of \( \lambda \) (in this case, \( \lambda = 11.12884 \)).

One disadvantage encountered using the \( N_2 \) constraint was that a tendency to “sweep around” and skip over the “forward” solution curve and find a solution lying “behind” the current location. This tendency was most pronounced for values of \( \lambda \) near the

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Final ( \lambda )</th>
<th># of Steps</th>
<th># of Steplength Reductions</th>
<th>Average # of Iterations/Step</th>
<th>Average CPU Seconds/Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_1 )</td>
<td>11.12838171</td>
<td>25</td>
<td>8</td>
<td>189.3</td>
<td>785.5</td>
</tr>
<tr>
<td>( N_2 )</td>
<td>11.12838919</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>419.4</td>
</tr>
</tbody>
</table>

Table 3.4.4 Performance statistics for pseudo-arclength constraints to pass \( \lambda = 11.12838 \), \( \Delta s_0 = 0.01 \).

<table>
<thead>
<tr>
<th>Constraint</th>
<th># of Steps</th>
<th># of Steplength Reductions</th>
<th>Average # of Iterations/Step</th>
<th>Average CPU Seconds/Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_1 )</td>
<td>270</td>
<td>7</td>
<td>177.3</td>
<td>1346.3</td>
</tr>
<tr>
<td>( N_2 )</td>
<td>122</td>
<td>0</td>
<td>56.7</td>
<td>265.9</td>
</tr>
</tbody>
</table>

Table 3.4.5 Performance statistics for pseudo-arclength constraints, \( \lambda = 11.12884 \), \( \Delta s_0 = 0.0001 \).
maximum that was obtained by $N_2$. To correct this behavior, a check was implemented: upon determination of $(u_{j+1}, \lambda_{j+1})$, the two vectors

$$
\begin{bmatrix}
  u_{j+1} - u_j \\
  \lambda_{j+1} - \lambda_j
\end{bmatrix}
$$

and

$$
\begin{bmatrix}
  u_j - u_{j-1} \\
  \lambda_j - \lambda_{j-1}
\end{bmatrix}
$$

were compared to see if they pointed in opposite direction. If they did, then the iteration had swung back around on the curve. In that case, the point $(u_{j+1}, \lambda_{j+1})$ was discarded, $\Delta s_j$ was reduced, and the iteration reattempted. The net result is that instances where the iteration wrapped back onto the curve were treated as convergence failures. With this correction in place, $N_2$ was still a more efficient approach in regions of high curvature of the solution manifold.

Another difference in the approximations obtained using $N_1$ and $N_2$ was that, for mesh M1 and $a = 1$, a turning point was computed with $N_2$, while $N_1$ terminated due to slow convergence before reaching the turning point. In Figure 3.4.5, the solution curves traced by both methods are shown for $\lambda$ close to the maximum. The plot on the right is a closer look at the turn than the plot on the left. The turning point produces values of

Figure 3.4.5  Turning point computed by $N_2$ for mesh M1, $a = 1$, $\Delta s_0 = 0.0001$.

$\lambda$ for which two approximations have been computed. In particular, for $\lambda = 11.1289745$, a solution was computed on both the lower branch and upper branch of the curve. The norm
of the difference of the two solutions was computed to be \( 8.6638 \times 10^{-4} \). Figure 3.4.6 shows a plot of the residual velocity field when the lower branch solution is subtracted from the upper branch solution. Although the magnitude of the difference in the solutions is small

![Velocity field, upper branch — lower branch, \( \lambda = 11.1289745 \), mesh M1, \( a = 1 \).](image)

relative to the full velocity field on either branch, the vortex observed in the difference between the solutions may indicate a temporal instability, as solutions on the upper branch beyond the turning point may be unsteady.

There are previous reports of the existence of a turning point in viscoelastic contraction flows [23, 34, 50, 79] and other problem geometries [9, 61, 78]. Some observations indicate that the solution behavior is similar to that of a loss of temporal stability. However, turning points observations were later dismissed as numerical artifact as they were not preserved under mesh refinement [65]. In the computations here, no other combination of mesh size and \( a \) parameter produced a turning point. Nevertheless, the existence of the turning point provides an important indicator of the instability of the discretized system of equations describing the flow.
4.1 Introduction

In this chapter we investigate the solution of a non-linear generalized Stokes problem using a dual-mixed formulation. The non-linear generalized Stokes problem arises in modeling flows of, for example, biological fluids, lubricators, paints, polymeric fluids, where the fluid's extra stress is assumed to be a non-linear function of the fluid's velocity gradient tensor. Several examples of such constitutive equations are given in Section 1.4.3.

The work in this chapter extends the investigations of [7, 31, 32, 59, 36]. In [7] Baranger, Najib, and Sandri provided an analysis for the existence and uniqueness of the modeling equations in appropriate Sobolev spaces and gave an error analysis of a finite element approximation method applied to the primitive variables $(\sigma, p, u)$. In [32], Farhloul and Manouzi extended the analysis in [31] of a mixed formulation for $r$-Laplacian problems to the appropriate Sobolev space setting. Manouzi and Farhloul in [59] reformulated the modeling equations into a saddle point problem and used a mixed formulation to study the existence and uniqueness of the solution, again in appropriate Sobolev spaces. An error analysis for the finite element approximation was also given. In each of [7, 31, 32, 59], the analysis used the assumption that the equation describing the extra stress $\sigma$ in terms of the velocity gradient tensor $\nabla u$ or the deformation tensor $D(u)$ was invertible to give $\nabla u$ or $D(u)$ as a function of $\sigma$. Gatica, González, and Meddahi in [36] reformulated the modeling equations, using the total stress tensor $\psi$ in place of the $\sigma$ $(\psi = \sigma - pI)$ and introducing an additional variable for $\nabla u$. Doing so, their formulation used the constitutive equation for $\sigma$ as a function of $\nabla u$ and reduced the regularity requirement for the velocity. Advantages for this approach include: (i) more flexibility in choosing the approximating finite element space for $u$, (ii) Dirichlet boundary conditions for $u$ become natural boundary conditions and are easily incorporated into the variational formulations, (iii) avoids the assumption of
expressing $\nabla u$ was a function of $\sigma$. A disadvantage in this formulation is that additional unknowns are introduced. The analysis provided by Gatica et al. in [36] was restricted to a Hilbert space setting.

In this chapter we recast the formulation described in [36] in appropriate Sobolev spaces. Because of the non-linearity in the constitutive equation this problem is more appropriately studied in Sobolev spaces which more accurately characterize the regularity of the solution. A description of the notation used in this paper, the mathematical problem, and the dual-mixed variational formulation is given in Section 4.2. Existence and uniqueness of the variational formulation is studied in Section 4.3. In Section 4.4 the finite element approximation is presented and analyzed. Numerical results are given in Section 4.5.

### 4.2 Mathematical Setting

The equations of a fluid obeying a shear-thinning type constitutive law discussed in Section 1.4.3 are given by

$$\sigma = g(\nabla u) \text{ in } \Omega, \quad (4.2.1)$$

$$-\nabla \cdot \sigma + \nabla p = f \text{ in } \Omega, \quad (4.2.2)$$

$$\text{div } u = 0 \text{ in } \Omega, \quad (4.2.3)$$

$$u = u_\Gamma \text{ on } \Gamma. \quad (4.2.4)$$

**Remark 4.2.1** From (4.2.3) it follows that $u_\Gamma$ must satisfy the compatibility condition

$$\int_{\Gamma} u_\Gamma \cdot n \, d\Gamma = 0,$$

where $n$ denotes the outward pointing unit normal vector to $\Omega$.

Used in the analysis below are the following function spaces and norms.

$$T := (L^r(\Omega))^{n \times n} = \{ \tau = (\tau_{ij}); \; \tau_{ij} \in L^r(\Omega); \; i, j = 1, \ldots, n \},$$

with norm $\| \tau \|_r := (\int_\Omega |\tau|^r \, d\Omega)^{1/r}$.

$$T' := (L^r(\Omega))^{n \times n} \quad \text{and} \quad T'_{\text{div}} := \{ \tau \in T'; \; \text{div } \tau \in (L^r(\Omega))^n \}.$$
with norm \( \|\tau\|_\text{div}^{r'} := \left( \int_\Omega (|\tau|^{r'} + |\text{div} \tau|^{r'}) \, d\Omega \right)^{1/r'} \). Let \( U := (L^r(\Omega))^n \), and \( P := L^{r'}(\Omega) \). For a Banach space \( X \), \( X^* \) denotes its dual space with associated norm \( \| \cdot \|_{X^*} \).

Note that \( T^* = T' \), and \( (T')^* = T \).

Motivated by (1.4.13)-(1.4.15), assume that \( \text{g} : T \to T^* \) is a bounded, continuous, strictly monotone operator \([70]\). (For the case of a power law fluid, monotonicity is shown in \([75, 18]\).) Assume that \( \text{g} \) satisfies the following assumptions. (Verification of these properties for a power law or Carreau law fluid is given in \([6]\)).

**Lemma 4.2.1** ([75], Proposition 3.1) There exist positive constants \( \hat{C}_1 \) and \( \hat{C}_2 \) such that, for \( s, t, w \in T \),

\[
\int_\Omega (\text{g}(s) - \text{g}(t)) : (s - t) \, d\Omega \\
\geq \hat{C}_1 \left( \int_\Omega |\text{g}(s) - \text{g}(t)| |s - t| \, d\Omega + \frac{\|s - t\|_T^2}{\|s\|_T^{2-r} + \|t\|_T^{2-r}} \right), \quad (4.2.5)
\]

\[
\int_\Omega (\text{g}(s) - \text{g}(t)) : w \, d\Omega \\
\leq \hat{C}_2 \left\| \frac{|s - t|}{|s| + |t|} \right\|_\infty^{2-r} \left( \int_\Omega |\text{g}(s) - \text{g}(t)| |s - t| \, d\Omega \right)^{\frac{r-1}{r}} \|w\|_T, \quad (4.2.6)
\]

with the convention that \( \text{g}(s) = 0 \) if \( s = 0 \) and \( |s(x) - t(x)|/(|s(x)| + |t(x)|) = 0 \) if \( s(x) = t(x) = 0 \).

In order to obtain the dual-mixed formulation, introduce two new variables, \( \phi \) and \( \psi \).

\[
\phi := \nabla u, \quad (4.2.7)
\]

\[
\psi := \sigma - pI, \quad \text{the total stress tensor}, \quad (4.2.8)
\]

\[
= \text{g}(\phi) - pI, \quad \text{using (4.2.1)}. \quad (4.2.9)
\]

With the definition of \( \psi \) a variational form for (4.2.2) can written as

\[
- \int_\Omega \mathbf{v} \cdot \text{div} \psi \, d\Omega = \int_\Omega \mathbf{v} \cdot \mathbf{f} \, d\Omega, \quad \text{for} \ \mathbf{v} \in T. \quad (4.2.10)
\]
Note that from the definition of $\phi$ we have that, for sufficiently smooth functions,

$$0 = -\int_{\Omega} \phi : \tau \, d\Omega + \int_{\Omega} \nabla u : \tau \, d\Omega$$

$$= -\int_{\Omega} \phi : \tau \, d\Omega + \int_{\Gamma} (\tau \cdot n) \cdot u_{\Gamma} \, d\Gamma - \int_{\Omega} u \cdot \text{div} \tau \, d\Omega$$

(4.2.11)

and the condition (4.2.3) is equivalent to

$$\text{tr}(\phi) = 0.$$  

(4.2.12)

Combining (4.2.1), (4.2.11), and (4.2.10) a variational formulation of (4.2.1)-(4.2.4) is: Given $f \in \left(L^{r'}(\Omega)\right)^n$, $u_{\Gamma} \in \left(W^{1-1/r, r}(\Gamma)\right)^n$, determine $(\phi, \psi, p, u) \in T \times T'_{\text{div}} \times P \times U$ such that

$$\int_{\Omega} g(\phi) : \varsigma \, d\Omega - \int_{\Omega} \psi : \varsigma \, d\Omega - \int_{\Omega} p \text{tr}(\varsigma) \, d\Omega = 0, \quad \forall \varsigma \in T,$$

(4.2.13)

$$-\int_{\Omega} \tau : \phi \, d\Omega - \int_{\Omega} q \text{tr}(\phi) \, d\Omega - \int_{\Omega} u \cdot \text{div} \tau \, d\Omega = -\int_{\Gamma} (\tau \cdot n) \cdot u_{\Gamma} \, d\Gamma,$$

$$\forall (\tau, q) \in T'_{\text{div}} \times P,$$  

(4.2.14)

$$-\int_{\Omega} v \cdot \text{div} \psi \, d\Omega = \int_{\Omega} v \cdot f \, d\Omega, \forall v \in U.$$  

(4.2.15)

Note that equations (4.2.13)-(4.2.15) do not uniquely define a solution; as adding $(0, c \mathbf{I}, -c, \mathbf{0})$ to a solution $(\phi, \psi, p, u)$, also satisfies (4.2.13)-(4.2.15) for any $c \in \mathbb{R}$. In order to guarantee uniqueness we proceed as in [3, 13, 36] and impose, via a Lagrange multiplier, the constraint $\int_{\Omega} \text{tr}(\psi) \, d\Omega = 0$. The variational formulation may then be restated as: Given $f \in \left(L^{r'}(\Omega)\right)^n$, $u_{\Gamma} \in \left(W^{1-1/r, r}(\Gamma)\right)^n$, determine $(\phi, \psi, p, u, \lambda) \in T \times T'_{\text{div}} \times P \times U \times \mathbb{R}$ such that

$$\int_{\Omega} g(\phi) : \varsigma \, d\Omega - \int_{\Omega} \psi : \varsigma \, d\Omega - \int_{\Omega} p \text{tr}(\varsigma) \, d\Omega = 0, \quad \forall \varsigma \in T,$$

(4.2.16)

$$-\int_{\Omega} \tau : \phi \, d\Omega - \int_{\Omega} q \text{tr}(\phi) \, d\Omega - \int_{\Omega} u \cdot \text{div} \tau \, d\Omega + \lambda \int_{\Omega} \text{tr}(\tau) \, d\Omega$$

$$= -\int_{\Gamma} (\tau \cdot n) \cdot u_{\Gamma} \, d\Gamma, \forall (\tau, q) \in T'_{\text{div}} \times P,$$  

(4.2.17)

$$-\int_{\Omega} v \cdot \text{div} \psi \, d\Omega + \eta \int_{\Omega} \text{tr}(\psi) \, d\Omega = \int_{\Omega} v \cdot f \, d\Omega, \forall (v, \eta) \in U \times \mathbb{R}.$$  

(4.2.18)
Remark 4.2.2 As commented in [36], the value of the Lagrange multiplier $\lambda$ is 0, as can be seen from the choice of $\tau = I$ and $q = -1$. However, it is included in the variational formulation so that the formulation has a twofold saddle point structure.

To formally rewrite (4.2.16)-(4.2.18) as a twofold saddle point problem define the following operators:

$A : T \rightarrow T'$, $B : T \rightarrow (T'_{\text{div}} \times P)^*$, $C : T'_{\text{div}} \times P \rightarrow (U \times \mathbb{R})^*$.

\[
[A(\phi), \varsigma] := \int_\Omega g(\phi) : \varsigma \, d\Omega, \quad (4.2.19)
\]
\[
[B(\phi), (\tau, q)] := -\int_\Omega \tau : \phi \, d\Omega - \int_\Omega q \text{tr}(\phi) \, d\Omega, \quad (4.2.20)
\]
\[
[C(\psi, p), (v, \eta)] := -\int_\Omega v \cdot \text{div} \psi \, d\Omega + \eta \int_\Omega \text{tr}(\psi) \, d\Omega. \quad (4.2.21)
\]

The modeling equations can then be written in the form

\[
[A(\phi), \varsigma] + [B(\varsigma), (\psi, p)] = 0, \forall \varsigma \in T, \quad (4.2.22)
\]
\[
[B(\phi), (\tau, q)] + [C(\tau, q), (u, \lambda)] = -\int_\Gamma (\tau \cdot n) \cdot u_r \, d\Gamma, \forall (\tau, q) \in T'_{\text{div}} \times P, \quad (4.2.23)
\]
\[
[C(\psi, p), (v, \eta)] = \int_\Omega v \cdot f \, d\Omega, \forall (v, \eta) \in U \times \mathbb{R}, \quad (4.2.24)
\]

or equivalent, in the form of a twofold saddle point equation,

\[
[A(\phi), \varsigma] + [\varsigma, B^*(\psi, p)] = 0, \forall \varsigma \in T, \quad (4.2.25)
\]
\[
[B(\phi), (\tau, q)] + [(\tau, q), C^*(u, \lambda)] = -\int_\Gamma (\tau \cdot n) \cdot u_r \, d\Gamma, \forall (\tau, q) \in T'_{\text{div}} \times P, \quad (4.2.26)
\]
\[
[C(\psi, p), (v, \eta)] = \int_\Omega v \cdot f \, d\Omega, \forall (v, \eta) \in U \times \mathbb{R}, \quad (4.2.27)
\]

where $B^*$ and $C^*$ denote the respective adjoint operators of $B$ and $C$, respectively.

4.3 Existence and Uniqueness

Intuitively, the solution of (4.2.25)-(4.2.27) can be found using the following steps.

1. Find a particular solution to (4.2.27), i.e. $(\psi_0, p_0)$ such that

\[
[C(\psi_0, p_0), (v, \eta)] = \int_\Omega v \cdot f \, d\Omega, \forall (v, \eta) \in U \times \mathbb{R}. \quad (4.3.1)
\]
2. Let \( \psi = \tilde{\psi} + \psi_0 \) and \( p = \tilde{p} + p_0 \). Rewriting (4.2.25)-(4.2.27) we have

\[
[A(\phi), \varsigma] + [\varsigma, B^*(\tilde{\psi}, \tilde{p})] = -[\varsigma, B^*(\psi_0, p_0)], \forall \varsigma \in T, \quad (4.3.2)
\]

\[
[B(\phi), (\tau, q)] + [(\tau, q), C^*(u, \lambda)] = -\int_{\Gamma} (\tau \cdot n) \cdot u_{\Gamma} \, d\Gamma, \forall (\tau, q) \in T'_{div} \times P, (4.3.3)
\]

\[
[C(\tilde{\psi}, \tilde{p}), (v, \eta)] = 0, \forall (v, \eta) \in U \times \mathbb{R}, \quad (4.3.4)
\]

3. Introduce a subspace of \( T'_{div} \times P \) defined by

\[
Z_1 := \left\{ (\tau, q) \in T'_{div} \times P : [(\tau, q), C^*(v, \eta)] = 0, \forall (v, \eta) \in U \times \mathbb{R} \right\}.
\]

Note that \( Z_1 \) can equivalently be defined as

\[
Z_1 := \left\{ (\tau, q) \in T'_{div} \times P : [C(\tau, q), (v, \eta)] = 0, \forall (v, \eta) \in U \times \mathbb{R} \right\}.
\]

Equations (4.3.2)-(4.3.4) are then replaced by

\[
[A(\phi), \varsigma] + [\varsigma, B^*(\tilde{\psi}, \tilde{p})] = -[\varsigma, B^*(\psi_0, p_0)], \forall \varsigma \in T, \quad (4.3.5)
\]

\[
[B(\phi), (\tau, q)] = -\int_{\Gamma} (\tau \cdot n) \cdot u_{\Gamma} \, d\Gamma, \forall (\tau, q) \in Z_1. \quad (4.3.6)
\]

4. Find a particular solution to (4.3.6), \( \phi_0 \), i.e.

\[
[B(\phi_0), (\tau, q)] = -\int_{\Gamma} (\tau \cdot n) \cdot u_{\Gamma} \, d\Gamma, \forall (\tau, q) \in Z_1. \quad (4.3.7)
\]

5. Let \( \phi = \tilde{\phi} + \phi_0 \). Rewriting (4.3.5)-(4.3.6) we have

\[
[A(\tilde{\phi} + \phi_0), \varsigma] + [\varsigma, B^*(\tilde{\psi}, \tilde{p})] = -[\varsigma, B^*(\psi_0, p_0)], \forall \varsigma \in T, \quad (4.3.8)
\]

\[
[B(\tilde{\phi}), (\tau, q)] = 0, \forall (\tau, q) \in Z_1. \quad (4.3.9)
\]

6. Introduce a subspace of \( T \) defined by

\[
Z_2 := \left\{ \varsigma \in T : [\varsigma, B^*(\tau, q)] = 0, \forall (\tau, q) \in Z_1 \right\}.
\]

Note that \( Z_2 \) can equivalently be defined as

\[
Z_2 := \left\{ \varsigma \in T : [B(\varsigma), (\tau, q)] = 0, \forall (\tau, q) \in Z_1 \right\}.
\]

Equations (4.3.8)-(4.3.9) are then replaced by

\[
[A(\tilde{\phi} + \phi_0), \varsigma] = -[\varsigma, B^*(\psi_0, p_0)], \forall \varsigma \in Z_2. \quad (4.3.10)
\]

7. Solve (4.3.10) for \( \tilde{\phi} \), from which we then get our solution for \( \phi, \phi = \tilde{\phi} + \phi_0 \).

8. From (4.3.8) we solve for \( (\tilde{\psi}, \tilde{p}) \) satisfying

\[
[\varsigma, B^*(\tilde{\psi}, \tilde{p})] = -[\varsigma, B^*(\psi_0, p_0)] - [A(\tilde{\phi} + \phi_0), \varsigma], \forall \varsigma \in T, \quad (4.3.11)
\]

which we use to determine \( \psi = \tilde{\psi} + \psi_0 \) and \( p = \tilde{p} + p_0 \).
9. Finally, we solve for \((\mathbf{u}, \lambda)\) using (4.3.3)

\[
[(\tau, q), \mathbf{C}^*(\mathbf{u}, \lambda)] = -\int_{\Gamma} (\tau \cdot \mathbf{n}) \cdot \mathbf{u} \, d\Gamma - [\mathbf{B}(\phi), (\tau, q)], \forall (\tau, q) \in \mathcal{T}_d \times P.
\]

(4.3.12)

Given the above, there are five issues to address.

(i) The solvability of \(\tilde{\phi}\) in (4.3.10), Step 7.

(ii) The solvability of \((\tilde{\psi}, \tilde{p})\) in (4.3.11), Step 8.

(iii) The solvability of \((\mathbf{u}, \lambda)\) in (4.3.12), Step 9.

(iv) The existence of a particular solution \((\psi_0, p_0)\) to (4.3.1), Step 1.

(v) The existence of a particular solution \(\phi_0\) to (4.3.7), Step 4.

**Issue (i):**

Given the stated assumptions on \(g, A\) is a bounded, continuous, strictly monotone operator on a reflexive Banach space. The solvability of \(\tilde{\phi}\) follows from monotone operator theory [70].

**Issue (ii):**

For the solvability of \((\tilde{\psi}, \tilde{p})\) we need to show that \(\mathbf{B}^*\) satisfies the following inf-sup condition:

\[
\inf_{(\tau, q) \in Z_1} \sup_{c \in \mathcal{T}} \frac{[c, \mathbf{B}^*(\tau, q)]}{\|c\|_\mathcal{T} \| (\tau, q) \|_{\mathcal{T}_d \times P}} \geq c_1,
\]

which is equivalent to

\[
\inf_{(\tau, q) \in Z_1} \sup_{c \in \mathcal{T}} \frac{\|B(c), (\tau, q)\|}{\|c\|_\mathcal{T} \| (\tau, q) \|_{\mathcal{T}_d \times P}} \geq c_1.
\]

(4.3.13)

(4.3.14)

**Issue (iii):**

For the solvability of \((\mathbf{u}, \lambda)\) we need to show that \(\mathbf{C}^*\) satisfies the following inf-sup condition:

\[
\inf_{(\mathbf{u}, \lambda) \in \mathcal{U} \times \mathbb{R}} \sup_{(\tau, q) \in \mathcal{T}_d \times P} \frac{[(\tau, q), \mathbf{C}^*(\mathbf{u}, \lambda)]}{\| (\tau, q) \|_{\mathcal{T}_d \times P} \| (\mathbf{u}, \lambda) \|_{\mathcal{U} \times \mathbb{R}}} \geq c_2,
\]

which is equivalent to

\[
\inf_{(\mathbf{u}, \lambda) \in \mathcal{U} \times \mathbb{R}} \sup_{(\tau, q) \in \mathcal{T}_d \times P} \frac{\| C(\tau, q), (\mathbf{u}, \lambda) \|}{\| (\tau, q) \|_{\mathcal{T}_d \times P} \| (\mathbf{u}, \lambda) \|_{\mathcal{U} \times \mathbb{R}}} \geq c_2.
\]

(4.3.15)

(4.3.16)

For **issues (iv) and (v)** note the following:
Lemma 4.3.1 ([38], Remark 4.2, pg. 61) Let \((X, \| \cdot \|_X)\) and \((M, \| \cdot \|_M)\) be two reflexive Banach spaces. Let \((X^*, \| \cdot \|_{X^*})\) and \((M^*, \| \cdot \|_{M^*})\) be their corresponding dual spaces. Let \(B : X \to M^*\) be a linear continuous operator and \(B^* : M^{**} \to X\) the dual operator of \(B\). Let \(V = \ker(B)\) be the kernel of \(B\); we denote by \(V^0 \subset X^*\) the polar set of \(V : V^0 = \{x^* \in X^*, [x^*, v] = 0, \forall v \in V\}\) and \(\hat{B} : (X/V) \to M^*\) the quotient operator associated with \(B\).

The following three properties are equivalent:

(i) \(\exists \beta > 0\), such that

\[
\inf_{q \in M} \sup_{v \in X} \frac{[Bv, q]}{\|q\|_M \|v\|_X} \geq \beta,
\]

(ii) \(B^*\) is an isomorphism from \(M^{**}\) onto \(V^0\) and

\[
\|B^*q\| \geq C_B \|q\|_{M^{**}} \quad \forall q \in M^{**},
\]

(iii) \(\hat{B}\) is an isomorphism from \((X/V)\) onto \(M^*\) and

\[
\|\hat{B}\hat{v}\| \geq C_B \|\hat{v}\|_{(X/V)} \quad \forall \hat{v} \in (X/V).
\]

Part (iii) of Lemma 4.3.1 will guarantee the existence of particular solutions, i.e. \(\psi_0, p_0, \phi_0\), and part (i) will guarantee existence and uniqueness of \(\hat{\psi}, \hat{p}, u,\) and \(\lambda\). Thus, we have two inf-sup conditions which we need to establish.

4.3.1 Inf-sup Condition for \(B\)

Define the null space for the operator \(C\), \(Z_1\), as

\[
Z_1 := \left\{ (\tau, q) \in T'_{\text{div}} \times P : [C(\tau, q), (v, \eta)] = 0, \forall (v, \eta) \in U \times \mathbb{R} \right\},
\]

\[
= \left\{ (\tau, q) \in T'_{\text{div}} \times P : \text{div} \tau = 0 \text{ in } \Omega, \text{ and } \int_{\Omega} tr(\tau) d\Omega = 0 \right\}. \quad (4.3.17)
\]

Note that for \((\tau, q) \in Z_1\), \(\|\tau\|_{T'_{\text{div}}} = \|\tau\|_{T'}\). Helpful in establishing the inf-sup condition for \(B\) is the following lemma.

Lemma 4.3.2 (See Lemma 3.1 in [3] for Hilbert space setting.)

For \(\tau \in T'_{\text{div}}\) satisfying \(\int_{\Omega} tr(\tau) d\Omega = 0\), let \(\tau^0 = \tau - \frac{1}{n} tr(\tau)I\). Then, there exists \(C,\)
depending only \( \Omega \), such that

\[
\|\boldsymbol{\tau}\|_{L^{r'}} \leq C \left( \|\boldsymbol{\tau}^0\|_{L^{r'}} + \|\text{div} \, \boldsymbol{\tau}\|_{W^{-1,r'}} \right). \tag{4.3.18}
\]

**Proof:** Now, there exists a non-zero function \( \varphi \in L^r(\Omega) \) such that

\[
\|\text{tr} (\boldsymbol{\tau})\|_{L^{r'}(\Omega)} \|\varphi\|_{L^r(\Omega)} = \int_{\Omega} \text{tr} (\boldsymbol{\tau}) \varphi \, d\Omega. \tag{4.3.19}
\]

Since \( \int_{\Omega} \text{tr} (\boldsymbol{\tau}) \, d\Omega = 0 \), we can assume \( \int_{\Omega} \varphi \, d\Omega = 0 \) (shift \( \varphi \) by its average). From [35], pg. 116 (see also [12] pg. 220), given \( \varphi \in L^r(\Omega) \), \( 1 < r < \infty \) with \( \int_{\Omega} \varphi \, d\Omega = 0 \), then there exists \( \boldsymbol{v} \in W^{1,r}_0(\Omega) \) and a constant \( C \) such that

\[
\text{div} \, \boldsymbol{v} = \varphi \quad \text{in} \quad \Omega \quad \text{and} \quad \|\boldsymbol{v}\|_{W^{1,r}(\Omega)} \leq C \|\varphi\|_{L^r(\Omega)}. \tag{4.3.20}
\]

From (4.3.19) and (4.3.20),

\[
\frac{1}{nC} \|\text{tr} (\boldsymbol{\tau})\|_{L^{r'}(\Omega)} \|\boldsymbol{v}\|_{W^{1,r}(\Omega)} \leq \frac{1}{n} \int_{\Omega} \text{tr} (\boldsymbol{\tau}) \text{div} \, \boldsymbol{v} \, d\Omega
\]

\[
= \frac{1}{n} \int_{\Omega} \text{tr} (\boldsymbol{\tau}) \nabla d \boldsymbol{v} \, d\Omega
\]

\[
= \int_{\Omega} (\boldsymbol{\tau} - \boldsymbol{\tau}^0) : \nabla d \boldsymbol{v} \, d\Omega \quad \text{(using the defn. of} \, \boldsymbol{\tau}^0) \]

\[
= -\int_{\Omega} (\boldsymbol{\tau}^0 : \nabla d \boldsymbol{v} + \text{div} \, \boldsymbol{\tau} \cdot \boldsymbol{v}) \, d\Omega
\]

\[
\leq \left( \|\boldsymbol{\tau}^0\|_{L^{r'}(\Omega)} + \|\text{div} \, \boldsymbol{\tau}\|_{W^{-1,r'}(\Omega)} \right) \|\boldsymbol{v}\|_{W^{1,r}(\Omega)}.
\]

**Lemma 4.3.3** There exists a constant \( c_1 > 0 \) such that

\[
\inf_{(\boldsymbol{\tau},q) \in \mathcal{Z}_1} \sup_{\phi \in \mathcal{T}} \frac{[\mathbf{B} (\phi) : (\boldsymbol{\tau},q)]}{\|\phi\|_T \| (\tau,q) \|_{T_{\text{div},P}}} \geq c_1.
\]

**Proof:** We establish the inf-sup condition by considering two cases.

**Case 1:** \( \|q\|_P \leq \|\boldsymbol{\tau}\|_{T_{\text{div}}} \).

Let

\[
\boldsymbol{\tau}^0 = \boldsymbol{\tau} - \frac{1}{n} \text{tr} (\boldsymbol{\tau}) \mathbf{I}, \quad \text{and} \quad \phi = -|\tau^0|^{r'/r-1} \tau^0 / \|\tau^0\|_{T_{\text{div}}}^{r'-1}. \tag{4.3.21}
\]
Note that $\phi \in T$, and $\|\phi\|_T = 1$. Then,

$$\frac{[B(\phi), (\tau, q)]}{\|\phi\|_T} = \int_\Omega \frac{|\tau^0|^{r'/r-1}}{\|\tau^0\|^{r'/r-1}_T} \tau : \tau^0 \, d\Omega + \int_\Omega q \frac{|\tau^0|^{r'/r-1}}{\|\tau^0\|^{r'/r-1}_T} \text{tr} (\tau^0) \, d\Omega$$

$$= \frac{1}{\|\tau^0\|^{r'-1}_T} \int_\Omega |\tau^0|^{r'/r-1} \tau^0 : \tau^0 \, d\Omega,$$

$$\text{as } \text{tr} (\tau^0) = 0, \text{ and } \tau : \tau^0 = \tau^0 : \tau^0$$

$$= \|\tau^0\|_T r' \geq \frac{1}{C} \|\tau\|_T \quad \text{as } (\tau, q) \in Z_1, \text{ see (4.3.18))}$$

$$\geq \frac{1}{2C} \left( \|\tau\|_{T'_{\text{div}}} + \|q\|_P \right) = \frac{1}{2C} \| (\tau, q) \|_{T'_{\text{div}} \times P}. \quad (4.3.22)$$

**Case 2:** $\|q\|_P \geq \|\tau\|_{T'_{\text{div}}}$. Let

$$\phi = \frac{-qI + \tau |^{r'/r-1}}{\|qI + \tau\|^{r'-1}_T}. \quad (4.3.23)$$

Again, $\phi \in T$, and $\|\phi\|_T = 1$. For this choice of $\phi$,

$$\frac{[B(\phi), (\tau, q)]}{\|\phi\|_T} = \int_\Omega \frac{|qI + \tau|^{r'/r-1}}{\|qI + \tau\|^{r'-1}_T} (\tau : (qI + \tau) + q \text{tr} (qI + \tau)) \, d\Omega$$

$$= \int_\Omega \frac{|qI + \tau|^{r'/r-1}}{\|qI + \tau\|^{r'-1}_T} (qI + \tau) : (qI + \tau) \, d\Omega$$

$$= \|qI + \tau\|_{T'} \geq \|qI\|_{T'} - \|\tau\|_{T'}$$

$$= n^{1/r'} \|q\|_P - \|\tau\|_{T'}$$

$$\geq (n^{1/r'} - 1) \|q\|_P$$

$$\geq (n^{1/r'} - 1)/2 (\|q\|_P + \|\tau\|_{T'_{\text{div}}})$$

$$= C \| (\tau, q) \|_{T'_{\text{div}} \times P}. \quad (4.3.24)$$

4.3.2 Inf-sup Condition for $C$

The following two lemmas are helpful in establishing the inf-sup condition for $C$.  

92
Lemma 4.3.4 Let \( 0 T'_{\text{div}} := \{ \tau \in T'_{\text{div}} : \int_\Omega \text{tr}(\tau) \, d\Omega = 0 \} \). Then, there exists \( C > 0 \) such that for any \( u \in U \)

\[
\sup_{\tau \in 0 T'_{\text{div}}} \frac{\int_\Omega u \cdot \text{div} \tau \, d\Omega}{\| \tau \|_{T'_{\text{div}}}} \geq C \sup_{\tau \in T'_{\text{div}}, \tau \neq 0} \frac{\int_\Omega u \cdot \text{div} \tau \, d\Omega}{\| \tau \|_{T'_{\text{div}}}}. \tag{4.3.25}
\]

Proof: For \( \tau \in T'_{\text{div}} \), let \( \tau_0 = \tau - \frac{1}{n|\Omega|} (\int_\Omega \text{tr}(\tau) \, d\Omega) \mathbf{I} \). Then, \( \tau_0 \in 0 T'_{\text{div}} \), and \( \text{div} \tau = \text{div} \tau_0 \). Let

\[
\varsigma := |\tau_0|^{r'/r - 1} \tau_0 + \frac{\text{sgn} \left( (\int_\Omega \text{tr}(\tau) \, d\Omega) (\int_\Omega |\tau_0|^{r'/r - 1} \tau_0 \, d\Omega) \right)}{n |\Omega|} \left( \int_\Omega |\tau_0|^{r'/r - 1} \tau_0 \, d\Omega \right) \mathbf{I}.
\]

Note that as \( \| |\tau_0|^{r'/r - 1} \tau_0 \|_{L^r} = \left( \int_\Omega |\tau_0|^{r'} \, d\Omega \right)^{1/r} = \| \tau_0 \|_{L^{r'}}^{r'/r} \),

and

\[
\left| \int_\Omega |\tau_0|^{r'/r - 1} \text{tr}(\tau_0) \, d\Omega \right| \leq \sqrt{n} \int_\Omega |\tau_0|^{r'/r} \, d\Omega \leq C \| \tau_0 \|_{L^{r'}}^{r'/r} \left( \int_\Omega 1^{r'} \, d\Omega \right)^{1/r'} \leq C \| \tau_0 \|_{L^{r'}}^{r'/r}.
\]

Thus

\[
\| \varsigma \|_{L^r} \leq C \| \tau_0 \|_{L^{r'}}^{r'/r}. \tag{4.3.26}
\]

We have that

\[
\| \tau \|_{L^{r'}} = \sup_{\sigma \in L^r} \frac{(\tau, \sigma)}{\| \sigma \|_{L^r}}. \tag{4.3.27}
\]

Now, using \( \tau_0 \in 0 T'_{\text{div}} \),

\[
(\tau, \varsigma) = \int_\Omega |\tau_0|^{r'} \, d\Omega + \frac{1}{n|\Omega|} \left( \int_\Omega \text{tr}(\tau) \, d\Omega \right) \left( \int_\Omega |\tau_0|^{r'/r - 1} \text{tr}(\tau_0) \, d\Omega \right)
+ \frac{1}{n|\Omega|} \left( \int_\Omega \text{tr}(\tau) \, d\Omega \right) \left( \int_\Omega |\tau_0|^{r'/r - 1} \text{tr}(\tau_0) \, d\Omega \right)
\geq \| \tau_0 \|_{L^{r'}}^{r'/r}. \tag{4.3.28}
\]
Therefore, from (4.3.26), (4.3.27), and (4.3.28) we have that $\|\boldsymbol{\tau}\|_{L^r'} \geq C \|\boldsymbol{\tau}_0\|_{L^r'}$. Combining the above we obtain

$$\int_{\Omega} \mathbf{u} \cdot \text{div} \mathbf{\tau} d\Omega = \int_{\Omega} \mathbf{u} \cdot \text{div} \mathbf{\tau}_0 d\Omega \leq C \int_{\Omega} \mathbf{u} \cdot \text{div} \mathbf{\tau}_0 d\Omega,$$

from which (4.3.25) then follows.

\[\Box\]

**Lemma 4.3.5** Given $\mathbf{w} \in (L^r'(\Omega))^2$, there exists $\mathbf{\tau} \in T_{\text{div}}'$ such that

$$\text{div} (\mathbf{\tau}) = \mathbf{w} \text{ in } \Omega, \text{ and } \|\mathbf{\tau}\|_{T_{\text{div}}'} \leq C \|\mathbf{w}\|_{L^r'(\Omega)}. \quad (4.3.29)$$

**Proof:** We show the construction of a suitable $\mathbf{\tau}$ for the case $n = 2$. The method extends in a straightforward manner for $n = 3$. For $\mathbf{w} := \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \in (L^r'(\Omega))^2$, let $w_1^0 := w_1 - \frac{1}{|\Omega|} \int_{\Omega} w_1 d\Omega$. Note that $w_1^0 \in L^r(\Omega)$ and $\int_{\Omega} w_1^0 d\Omega = 0$. Then, from [35], pg. 116, there exist $\mathbf{v} := \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ and a constant $C$ such that

$$\text{div} \mathbf{v} = w_1^0 \text{ in } \Omega, \text{ and } \|\mathbf{v}\|_{W^{1,r'}(\Omega)} \leq C \|w_1^0\|_{L^r'(\Omega)}.$$

Next consider the problem: Find $\mathbf{z} \in \{ \mathbf{f} : \mathbf{f} \in L^r'(\Omega) \text{ and } f_y \in L^r'(\Omega) \}$ such that

$$\text{div} \begin{pmatrix} v_1 \\ z \end{pmatrix} = w_2 \text{ in } \Omega.$$

This is equivalent to finding $z_y \in L^r'(\Omega)$ such that

$$z_y = w_2 - v_{2x} \text{ in } \Omega.$$

As $\Omega$ is bounded, without loss of generality, assume $\Omega \subset \{(x,y) : 0 < x < a, 0 < y < b\}$. For $g \in L^r'(\Omega)$, let $\tilde{g}$ denote the extension by zero of $g$ to $\mathbb{R}^2 \setminus \Omega$. For $(x, y) \in \Omega$, let

$$z(x, y) = \int_0^y \left( w_2(x, s) - v_{2x}(x, s) \right) ds.$$
Now
\[ \| \nabla \|_{L^r}(\Omega) = \int_0^b \int_0^a \left( \int_0^y \left( w_2 - v_{2x} \right) d\gamma \right)^{r'} dx dy \]
\[ \leq \int_0^b \int_0^a \left( \int_0^b \left( \int_0^y \left( w_2 - v_{2x} \right) d\gamma \right)^{r'/r} \right) dx dy \]
\[ \leq b^r \| w_2 - v_{2x} \|_{L^r}(\Omega) \]
\[ \leq C \left( \| w_2 \|_{L^r}(\Omega) + \| v_{2x} \|_{L^r}(\Omega) \right) \leq C \left( \| w_2 \|_{L^r}(\Omega) + \| w_0 \|_{L^r}(\Omega) \right). \]

Also, note that
\[ \| w_0 \|_{L^r}(\Omega) \leq \| w_1 \|_{L^r}(\Omega) + \left\| \frac{1}{|\Omega|} \int_{\Omega} w_1 d\Omega \right\|_{L^r}(\Omega), \]
and
\[ \left\| \frac{1}{|\Omega|} \int_{\Omega} w_1 d\Omega \right\|_{L^r}(\Omega) \leq 2 \| w_1 \|_{L^r}(\Omega). \]

Let \( \tau(x, y) := \begin{bmatrix} v_1 + x \frac{1}{|\Omega|} \int_{\Omega} w_1 d\Omega & v_2 \\ v_2 & z \end{bmatrix} \). Then
\[ \text{div} \tau = \text{div} \begin{bmatrix} v_1 & v_2 \\ v_2 & z \end{bmatrix} + \text{div} \begin{bmatrix} x \frac{1}{|\Omega|} \int_{\Omega} w_1 d\Omega & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} w_0^0 \\ w_2 \end{bmatrix} + \begin{bmatrix} 1 \frac{1}{|\Omega|} \int_{\Omega} w_1 d\Omega \\ 0 \end{bmatrix} = \mathbf{w}. \]

Moreover,
\[ \| \tau \|_{T_{\text{div}}'} \leq C \| \mathbf{w} \|_{L^r}(\Omega). \]
Lemma 4.3.6 There exists a constant $c_2 > 0$ such that
\[
\inf_{(u,\lambda) \in U \times R} \sup_{(\tau,q) \in T'_{div} \times P} \frac{[C(\tau,q), (u,\lambda)]}{\| (\tau,q) \|_{T'_div \times P} \| (u,\lambda) \|_{U \times R}} \geq c_2.
\]

Proof: We establish the inf-sup condition by considering two cases.

Case 1. $|\lambda| \geq \|u\|_U$.

For this case we have
\[
\sup_{(\tau,q) \in T'_{div} \times P} \frac{[C(\tau,q), (u,\lambda)]}{\| (\tau,q) \|_{T'_div \times P} \| (u,\lambda) \|_{U \times R}} \geq \frac{C(\lambda I,0), (u,\lambda)}{\| \lambda I \|_{T'_div}} = \frac{n \lambda^2 |\Omega|}{|\lambda| n^{r'/2} |\Omega|^{1/r'}} \geq C\| (u,\lambda) \|_{U \times R}.
\]

(4.3.30)

Case 2. $|\lambda| \leq \|u\|_U$.

Using Lemma 4.3.4,
\[
\sup_{(\tau,q) \in T'_{div} \times P} \frac{[C(\tau,q), (u,\lambda)]}{\| (\tau,q) \|_{T'_div \times P} \| (u,\lambda) \|_{U \times R}} \geq \sup_{\tau_0 \in 0T'_{div}} \frac{[C(\tau_0,0), (u,\lambda)]}{\| \tau_0 \|_{T'_div}} = \frac{-\int_{\Omega} u \cdot \text{div} \tau_0 \, d\Omega}{\| \tau \|_{T'_div}} \geq C\|u\|_U \geq C\| (u,\lambda) \|_{U \times R}.
\]

(4.3.31)

Choose $w \in (L^{r'}(\Omega))^n$ such that $\|u\|_{L'} \|w\|_{L^{r'}} = \int_{\Omega} u \cdot w \, d\Omega$, and let $\tau$ be as given in (4.3.29).

Then,
\[
\sup_{(\tau,q) \in T'_{div} \times P} \frac{[C(\tau,q), (u,\lambda)]}{\| (\tau,q) \|_{T'_div \times P} \| (u,\lambda) \|_{U \times R}} \geq C \frac{-\int_{\Omega} u \cdot \text{div}(\tau) \, d\Omega}{\| \tau \|_{T'_div}} \geq C \frac{\int_{\Omega} u \cdot w \, d\Omega}{\|w\|_{L^{r'}}} \geq C\| (u,\lambda) \|_{U \times R}.
\]

(4.3.32)

We now summarize the above results in the following theorem.

Theorem 4.3.1 There exists a unique solution $(\phi,\psi,p,u,\lambda) \in T \times T'_{div} \times P \times U \times R$ satisfying (4.2.22)-(4.2.24). In addition, we have that
\[
\|\phi\|_T \leq C \left( \|u\|_{1-1/r,r,T} + \|f\|_{r'/r,0,r',\Omega} \right).
\]

(4.3.33)
Proof: Existence and uniqueness of the solution \((\phi, \psi, p, u, \lambda) \in T \times T'_{\text{div}} \times P \times U \times \mathbb{R}\) follows directly from the continuity and monotonicity of \(g\) and the inf-sup conditions (4.3.14) and (4.3.16).

To show (4.3.33), we begin with some preliminary bounds. Let \(\phi = \tilde{\phi} + \phi_0\), \(\psi = \tilde{\psi} + \psi_0\), and \(p = \tilde{p} + p_0\) be as given in (4.3.1)-(4.3.12). Note that from (4.3.1) and the inf-sup condition (4.3.16), we have that there exists \((v, \eta) \in U \times \mathbb{R}\) such that

\[
\frac{c_2}{2} \leq \frac{\|C(\psi_0, p_0, (v, \eta))\|_{U \times \mathbb{R}} \|\psi_0, p_0\|_{T'_{\text{div}} \times P}}{\|v\|_{U \times \mathbb{R}} \|\psi_0, p_0\|_{T'_{\text{div}} \times P}} = \frac{\int_{\Omega} v \cdot f \, d\Omega}{\|v\|_{U \times \mathbb{R}} \|\psi_0, p_0\|_{T'_{\text{div}} \times P}} \leq \frac{\|v\|_U \|f\|_{L'(\Omega)}^n}{\|\psi_0\|_{T'} + \|p_0\|_P} = \frac{\|f\|_{0, r', \Omega}}{\|\psi_0\|_{T'} + \|p_0\|_P},
\]

(4.3.34)

or

\[
\|\psi_0\|_{T'} + \|p_0\|_P \leq \frac{2}{c_2} \|f\|_{0, r', \Omega}.
\]

(4.3.35)

From Lemma (4.3.1) (i) and (iii) with the associations \(B = B, M = Z_1 \subset T'_{\text{div}} \times P, X = T,\) and \(V = \ker(B)\), we have that there exists a \(\dot{\phi} \in T/V\) such that

\[
[B(\dot{\phi}), (\tau, q)] = -\int_{\Gamma} (\tau \cdot n) \cdot u_{\Gamma} \, d\Gamma, \quad \forall (\tau, q) \in Z_1,
\]

with \(\|\dot{\phi}\|_{T/V} \leq \frac{1}{C_B} \|u_{\Gamma}\|_{1 - 1/r, \Gamma} \). Note that \(\|\dot{\phi}\|_{T/V} := \inf_{\zeta \in \dot{\phi}} \|\zeta\|_T\). As the cosets in \(T/V\) are closed, we can choose \(\phi_0 \in \dot{\phi}\) such that

\[
\|\phi_0\|_T = \|\dot{\phi}\|_{T/V} \leq \frac{1}{C_B} \|u_{\Gamma}\|_{1 - 1/r, \Gamma}.
\]

(4.3.36)

From (4.2.5) we have that

\[
C_1 \left(\|\tilde{\phi} + \phi_0\|_T^2 + \int_{\Omega} |g(\tilde{\phi} + \phi_0)| \tilde{\phi} + \phi_0 | \, d\Omega\right) \leq \int_{\Omega} g\left(\tilde{\phi} + \phi_0\right) : (\tilde{\phi} + \phi_0) \, d\Omega.
\]

(4.3.37)
Using (4.2.19) with \( \zeta = \tilde{\phi} \) we have

\[
[A(\tilde{\phi} + \phi_0), \tilde{\phi}] = \int_{\Omega} g(\tilde{\phi} + \phi_0) : \tilde{\phi} \, d\Omega
\]

\[
= \int_{\Omega} g(\tilde{\phi} + \phi_0) \cdot (\tilde{\phi} + \phi_0) 
+ \int_{\Omega} g(\tilde{\phi} + \phi_0) : \phi_0 \, d\Omega
\]

\[
\geq \hat{C}_1 \left( \|\tilde{\phi} + \phi_0\|_T^r + \int_{\Omega} |g(\tilde{\phi} + \phi_0)| ||\tilde{\phi} + \phi_0||_T \right)
\]

\[
- \hat{C}_2 \left( \int_{\Omega} |g(\tilde{\phi} + \phi_0)||\tilde{\phi} + \phi_0||_T \right)^{1/r'} ||\phi_0||_T
\]

\[
\geq \hat{C}_1 \|\tilde{\phi} + \phi_0\|_T + \left( \hat{C}_1 - \frac{\epsilon_1 \hat{C}_2}{r'} \right) \int_{\Omega} |g(\tilde{\phi} + \phi_0)||\tilde{\phi} + \phi_0||_T
\]

\[
- \frac{\hat{C}_2}{r \epsilon_1} ||\phi_0||_T^r. \quad (4.3.38)
\]

Now we also have from (4.3.10), using Young’s inequality and the triangle inequality,

\[
[A(\tilde{\phi} + \phi_0), \tilde{\phi}] = - [B(\tilde{\phi}), (\psi_0, p_0)]
\]

\[
= \int_{\Omega} \psi_0 : \tilde{\phi} \, d\Omega + \int_{\Omega} p_0 \, tr(\tilde{\phi}) \, d\Omega
\]

\[
\leq \|\psi_0\|_{T'} \|\tilde{\phi}\|_T + \sqrt{n} \|p_0\|_P \|\tilde{\phi}\|_T
\]

\[
\leq \frac{2\epsilon_2}{r} \|\tilde{\phi}\|_T^r + \frac{1}{r' \epsilon_2} \left( \|\psi_0\|_{T'}^r + \sqrt{n} \|p_0\|_P^r \right)
\]

\[
\leq \frac{2\epsilon_2}{r} \|\tilde{\phi} + \phi_0\|_T^r + \frac{2\epsilon_2}{r} \|\phi_0\|_T^r
\]

\[
+ \frac{1}{r' \epsilon_2} \left( \|\psi_0\|_{T'}^r + \sqrt{n} \|p_0\|_P^r \right). \quad (4.3.39)
\]

Combining (4.3.38), (4.3.39), and \( \phi = \tilde{\phi} + \phi_0 \), we have

\[
\left( \hat{C}_1 - \frac{2\epsilon_2}{r} \right) \|\phi\|_T^r + \left( \hat{C}_1 - \frac{\epsilon_1 \hat{C}_2}{r'} \right) \int_{\Omega} |g(\phi)| ||\phi||_T \, d\Omega
\]

\[
\leq \left( \frac{\hat{C}_2}{r \epsilon_1} + \frac{2\epsilon_2}{r} \right) \|\phi_0\|_T + \frac{1}{r' \epsilon_2} \left( \|\psi_0\|_{T'}^r + \sqrt{n} \|p_0\|_P^r \right). \quad (4.3.40)
\]

Together with (4.3.35), (4.3.36) and choices for \( \epsilon_1, \epsilon_2 \) that ensure

\[
\hat{C}_1 - \frac{2\epsilon_2}{r} > 0, \quad \text{and} \quad \hat{C}_1 - \frac{\epsilon_1 \hat{C}_2}{r'} > 0,
\]

the result (4.3.33) is shown.
4.4 Finite Element Approximation

Let $\Omega \subset \mathbb{R}^n$ be a polygonal domain and let $T_h$ be a triangulation of $\Omega$ into triangles $(n = 2)$ or tetrahedrals $(n = 3)$. Thus

$$\Omega = \bigcup K, \quad K \in T_h,$$

and assume that there exist constants $\gamma_1, \gamma_2$ such that

$$\gamma_1 h \leq h_K \leq \gamma_2 \rho_K$$

(4.4.1)

where $h_K$ is the diameter of triangle (tetrahedral) $K$, $\rho_K$ is the diameter of the greatest ball (sphere) included in $K$, and $h = \max_{K \in T_h} h_K$. Define the finite-dimensional subspaces $T_h \subseteq T$, $T_{\text{div}, h} \subseteq T_{\text{div}}$, $P_h \subseteq P$, and $U_h \subseteq U$. Then the discrete formulation of (4.2.16)-(4.2.18) is defined as: Given $f \in \left(L^r(\Omega)\right)^n$, $\mathbf{u}_T \in \left(W^{1-1/r,r}(\Gamma)\right)^n$, determine $(\phi_h, \psi_h, p_h, \mathbf{u}_h, \lambda_h) \in T_h \times T_{\text{div}, h} \times P_h \times U_h \times \mathbb{R}$ such that

$$\int_{\Omega} g(\phi_h) : \mathbf{s}_h \, d\Omega - \int_{\Omega} \psi : \mathbf{s}_h \, d\Omega - \int_{\Omega} p_h \, \text{tr} \mathbf{s}_h \, d\Omega = 0, \forall \mathbf{s}_h \in T_h, \quad (4.4.2)$$

$$- \int_{\Omega} \mathbf{t}_h : \phi_h \, d\Omega - \int_{\Omega} q_h \, \text{tr} \phi_h \, d\Omega - \int_{\Omega} \mathbf{u}_h \cdot \text{div} \mathbf{t}_h \, d\Omega + \lambda_h \int_{\Omega} \text{tr} \mathbf{t}_h \, d\Omega$$

$$= - \int_{\Gamma} (\mathbf{t}_h \cdot \mathbf{n}) \cdot \mathbf{u}_\Gamma \, d\Gamma, \forall (\mathbf{t}_h, q_h) \in T_{\text{div}, h} \times P_h, \quad (4.4.3)$$

$$- \int_{\Omega} \mathbf{v}_h \cdot \text{div} \psi_h \, d\Omega + \eta_h \int_{\Omega} \text{tr} \psi_h \, d\Omega = \int_{\Omega} \mathbf{v}_h \cdot f \, d\Omega, \forall (\mathbf{v}_h, \eta_h) \in U_h \times \mathbb{R}. \quad (4.4.4)$$

Restricting the domain of the operators $A$, $B$, and $C$ to the appropriate finite-dimensional subspaces we can write (4.4.2)-(4.4.4) as a twofold saddle point system:

$$[A(\phi_h), \mathbf{s}_h] + [\mathbf{s}_h, B^*(\psi_h, p_h)] = 0, \forall \mathbf{s}_h \in T_h, \quad (4.4.5)$$

$$[B(\phi_h), (\mathbf{t}_h, q_h)] + [(\mathbf{t}_h, q_h), C^*(\mathbf{u}_h, \lambda_h)] = - \int_{\Gamma} (\mathbf{t}_h \cdot \mathbf{n}) \cdot \mathbf{u}_\Gamma \, d\Gamma,$$

$$\forall (\mathbf{t}_h, q_h) \in T_{\text{div}, h} \times P_h, \quad (4.4.6)$$

$$[C(\psi_h, p_h), (\mathbf{v}_h, \eta_h)] = \int_{\Omega} \mathbf{v}_h \cdot f \, d\Omega, \forall (\mathbf{v}_h, \eta_h) \in U_h \times \mathbb{R}. \quad (4.4.7)$$

The corresponding discrete kernels of $B$ and $C$ are defined similarly. We have

$$Z_{1h} := \left\{ (\mathbf{t}_h, q_h) \in T_{\text{div}, h} \times P_h : [C(\mathbf{t}_h, q_h), (\mathbf{v}_h, \eta_h)] = 0, \forall (\mathbf{v}_h, \eta_h) \in U_h \times \mathbb{R} \right\}.$$
and
\[ Z_{2h} := \{ \varsigma_h \in T_h : [B(\varsigma_h), (\tau_h, q_h)] = 0, \forall (\tau_h, q_h) \in Z_{1h} \} \, . \]

### 4.4.1 A Priori Estimates

**Theorem 4.4.1** Let \( 1 < r < 2 \) and \( g \) satisfy (4.2.5) and (4.2.6). Let \( (\phi, \psi, p, u, \lambda) \in T \times T'_{\text{div}} \times P \times U \times \mathbb{R} \) solve (4.2.16)-(4.2.18). Assume that

1. The nonlinear operator \( A : T_h \rightarrow T_{h}^\prime \) is a bounded, continuous, strictly monotone operator
2. There exists a positive constant \( c_1 \) such that
   \[
   \inf_{(\tau_h, q_h) \in \mathcal{Z}_{1h}} \sup_{\varsigma_h \in T_h} \frac{[B(\varsigma_h), (\tau_h, q_h)]}{\|\varsigma_h\|_T \|(\tau_h, q_h)\|_{T'_{\text{div}} \times P}} \geq c_1 . \tag{4.4.8}
   \]
3. There exists a positive constant \( c_2 \) such that
   \[
   \inf_{(u_h, \lambda_h) \in U_h \times \mathbb{R}} \sup_{(\tau_h, q_h) \in T'_{\text{div}, h}} \frac{[C(\tau_h, q_h), (u_h, \lambda_h)]}{\|(\tau_h, q_h)\|_{T'_{\text{div}} \times P} \|(u_h, \lambda_h)\|_{U_h \times \mathbb{R}}} \geq c_2 . \tag{4.4.9}
   \]

Then, for \( f \in \left( L^{r'}(\Omega) \right)^n \) and \( u_\Gamma \in \left( W^{1,1-r; r} (\Gamma) \right)^n \), there exists a unique solution
\( (\phi_h, \psi_h, p_h, u_h, \lambda_h) \in T_h \times T'_{\text{div}, h} \times P_h \times U_h \times \mathbb{R} \) to the problem (4.4.5)-(4.4.7). In addition, we have
\[
\|\phi_h\|_T \leq C \left( \|u_\Gamma\|_{1-1/r, r; \Gamma} + \|f\|_{0, r'; \Omega} \right) , \tag{4.4.10}
\]
for some constant \( C > 0 \).

**Proof:** With the assumptions as stated above, existence and uniqueness of
\( (\phi_h, \psi_h, p_h, u_h, \lambda_h) \in T_h \times T'_{\text{div}, h} \times P_h \times U_h \times \mathbb{R} \) solving (4.4.5)-(4.4.7) follows directly from the continuous solution approach outlined in Section 4.3 and summarized in Theorem 4.3.1. The proof of (4.4.10) mirrors the approach for showing (4.3.33).

**Theorem 4.4.2** Assume the hypotheses of Theorem 4.4.1 are satisfied. Also assume that for \( h \) sufficiently small, there is a constant \( c_3 > 0 \) such that
\[
\inf_{(\tau_h, q_h) \in T'_{\text{div}, h} \times P_h} \sup_{(\varsigma_h, \psi_h, \eta_h) \in T_h \times U_h \times \mathbb{R}} \frac{[B(\varsigma_h), (\tau_h, q_h)] + [C(\tau_h, q_h), (\psi_h, \eta_h)]}{\|\varsigma_h, \psi_h, \eta_h\|_{T \times U \times \mathbb{R}}} \geq c_3 . \tag{4.4.11}
\]
where \( \| (\varsigma_h, \mathbf{v}_h, \eta_h) \|_{T \times U \times \mathbb{R}} = \| \varsigma_h \|_T + \| \mathbf{v}_h \|_U + \| \lambda_h \|_\mathbb{R} \). Then

\[
\| \phi - \phi_h \|_T + \| \mathbf{u} - \mathbf{u}_h \|_U + | \lambda - \lambda_h |
\leq C \left\{ \inf_{\varsigma_h \in T_h} \| \phi - \varsigma_h \|_T^{r/2} + \inf_{\mathbf{v}_h \in U_h} \| \mathbf{u} - \mathbf{v}_h \|_U^{r/2} + \inf_{\tau_h \in T_{\text{div}, h}} \| \psi - \tau_h \|_{T'} + \inf_{q_h \in P_h} \| p - q_h \|_P \right\}, \tag{4.4.12}
\]

\[
\| \psi - \psi_h \|_{T_{\text{div}}'} + \| p - p_h \|_P \leq C \left\{ \inf_{\varsigma_h \in T_h} \| \phi - \varsigma_h \|_T^{r'/r'} + \inf_{\mathbf{v}_h \in U_h} \| \mathbf{u} - \mathbf{v}_h \|_U^{r'/r'} + \inf_{\tau_h \in T_{\text{div}, h}} \| \psi - \tau_h \|_{T'}^{2/r'} + \inf_{q_h \in P_h} \| p - q_h \|_P^{2/r'} \right\}. \tag{4.4.13}
\]

for some constant \( C > 0 \).

**Proof:** From the discrete formulation (4.4.2)-(4.4.4) we have that the approximation \((\phi_h, \psi_h, p_h, \mathbf{u}_h, \lambda_h)\) satisfies

\[
[\mathbf{A}(\phi_h), \varsigma_h] + [\mathbf{B}(\varsigma_h), (\psi_h, p_h)] = 0, \forall \varsigma_h \in T_h,
\]

\[
[\mathbf{B}(\phi_h), (\tau_h, q_h)] + [\mathbf{C}(\tau_h, q_h), (\mathbf{u}_h, \lambda_h)] = - \int_{\Gamma} (\tau_h \cdot \mathbf{n}) \cdot \mathbf{u}_r \, d\Gamma, \forall (\tau_h, q_h) \in T_{\text{div}, h} \times P_h,
\]

\[
[\mathbf{C}(\psi_h, p_h), (\mathbf{v}_h, \eta_h)] = \int_{\Omega} \mathbf{v}_h \cdot \mathbf{f} \, d\Omega, \forall (\mathbf{v}_h, \eta_h) \in U_h \times \mathbb{R}.
\]

Define the following subspaces:

\[
\tilde{Z}_{1h} := \left\{ (\tau_h, q_h) \in T_{\text{div}, h} \times P_h : [\mathbf{C}(\tau_h, q_h), (\mathbf{v}_h, \eta_h)] = \int_{\Omega} \mathbf{v}_h \cdot \mathbf{f} \, d\Omega, \right\}, \tag{4.4.14}
\]

and

\[
\tilde{Z}_{2h} := \left\{ \varsigma_h \in T_h : [\mathbf{B}(\varsigma_h), (\tau_h, q_h)] + [(\tau_h, q_h), \mathbf{C}^*(\mathbf{u}_h, \lambda_h)] = - \int_{\Gamma} (\tau_h \cdot \mathbf{n}) \cdot \mathbf{u}_r \, d\Gamma, \right\}. \tag{4.4.15}
\]

101
From (4.2.5) and the definition of $A$ (4.2.19), we have,

$$
\hat{C}_1 \frac{\|\phi - \phi_h\|_T^2}{\|\phi\|_T^{2-r} + \|\phi_h\|_T^{2-r}} + \hat{C}_1 \int_\Omega |g(\phi) - g(\phi_h)| |\phi - \phi_h| \, d\Omega
$$

$$
\leq \int_\Omega (g(\phi) - g(\phi_h)) : (\phi - \phi_h) \, d\Omega, \quad (4.4.16)
$$

and

$$
\int_\Omega (g(\phi) - g(\phi_h)) : (\phi - \phi_h) \, d\Omega = [A(\phi) - A(\phi_h), \phi - \phi_h]
$$

$$
= [A(\phi) - A(\phi_h), \phi - \zeta_h]
$$

$$
- [A(\phi) - A(\phi_h), \zeta_h - \phi_h]. \quad (4.4.17)
$$

We examine the first term on the RHS of (4.4.17). From (4.2.6) and Young’s inequality, we have

$$
[A(\phi) - A(\phi_h), \phi - \zeta_h] = \int_\Omega (g(\phi) - g(\phi_h)) : (\phi - \zeta_h) \, d\Omega
$$

$$
\leq \hat{C}_2 \left\| \frac{\phi - \phi_h}{|\phi| + |\phi_h|} \right\|_{0,\infty}^{(2-r)/r} \left( \int_\Omega |g(\phi) - g(\phi_h)| |\phi - \phi_h| \, d\Omega \right)^{(r-1)/r} \|\phi - \zeta_h\|_T
$$

$$
\leq \frac{\hat{C}_2' \epsilon_1}{r} \int_\Omega |g(\phi) - g(\phi_h)| |\phi - \phi_h| \, d\Omega + \frac{1}{r \epsilon_1} \|\phi - \zeta_h\|_T^r. \quad (4.4.18)
$$
For the second term on the RHS of (4.4.17), if \( \varsigma_h \in \tilde{Z}_{2h} \), we have

\[
[A(\phi) - A(\phi_h), \varsigma_h - \phi_h] = [A(\phi), \varsigma_h - \phi_h] - [A(\phi_h), \varsigma_h - \phi_h]
\]

\[
= -[B(\varsigma_h - \phi_h), (\psi, p)] + [B(\varsigma_h - \phi_h), (\psi_h, p_h)]
\]

\[
= [B(\phi_h - \varsigma_h), (\psi, p)] \quad \text{(as \( \varsigma_h, \phi_h \in \tilde{Z}_{2h} \))}
\]

\[
= [B(\phi_h - \varsigma_h), (\psi, p)] - [B(\phi_h - \varsigma_h), (\tau_h, q_h)] \quad \text{(for \( (\tau_h, q_h) \in \tilde{Z}_{1h} \))}
\]

\[
= [B(\phi_h - \varsigma_h), (\psi - \tau_h, p - q_h)]
\]

\[
= -\int_{\Omega} (\phi_h - \phi) : (\psi - \tau_h) d\Omega - \int_{\Omega} (p - q_h) tr(\phi_h - \phi) d\Omega
\]

\[
- \int_{\Omega} (\phi - \varsigma_h) : (\psi - \tau_h) d\Omega - \int_{\Omega} (p - q_h) tr(\phi - \varsigma_h) d\Omega
\]

\[
\leq \|\phi - \phi_h\|_{T'} \|\psi - \tau_h\|_{T'} + \sqrt{n} \|p - q_h\|_{P} \|\phi - \phi_h\|_{T}
\]

\[
+ \|\phi - \varsigma_h\|_{T} \|\psi - \tau_h\|_{T'} + \sqrt{n} \|p - q_h\|_{P} \|\phi - \varsigma_h\|_{T}
\]

\[
\leq \frac{\epsilon_2 + \epsilon_3}{2} \|\phi - \phi_h\|^2_{T} + \frac{1}{2\epsilon_2} \|\psi - \tau_h\|^2_{T'} + \frac{\sqrt{n}}{2\epsilon_3} \|p - q_h\|^2_{P}
\]

\[
+ \frac{\epsilon_4 + \epsilon_5}{r} \|\phi - \varsigma_h\|^r_{T} + \frac{1}{\epsilon_4 r'} \|\psi - \tau_h\|^r_{T'} + \frac{\sqrt{n}}{\epsilon_5 r'} \|p - q_h\|^r_{P}. \quad (4.4.19)
\]

Combining (4.4.16)-(4.4.19) with \( \epsilon_4 = \epsilon_5 = 1 \) we have

\[
\left( \frac{\hat{C}_1}{\|\phi\|^2_{T} + \|\phi_h\|^2_{T}} - \frac{\epsilon_2 + \epsilon_3}{2} \right) \|\phi - \phi_h\|^2_{T}
\]

\[
+ \left( \frac{\hat{C}_1}{\|\phi\|^2_{T} + \|\phi_h\|^2_{T}} - \frac{\epsilon_2 + \epsilon_3}{2} \right) \int_{\Omega} |g(\phi) - g(\phi_h)| |\phi - \phi_h| \, d\Omega
\]

\[
\leq \left( \frac{1}{r\epsilon_1} + \frac{2}{r} \right) \|\phi - \varsigma_h\|^r_{T} + \frac{1}{2\epsilon_2} \|\psi - \tau_h\|^2_{T'} + \frac{1}{r'} \|\psi - \tau_h\|^r_{T'}
\]

\[
+ \frac{\sqrt{n}}{2\epsilon_3} \|p - q_h\|^2_{P} + \frac{\sqrt{n}}{r'} \|p - q_h\|^r_{P}. \quad (4.4.20)
\]

Choosing \( \epsilon_1, \epsilon_2, \epsilon_3 \) small enough to ensure

\[
\left( \frac{\hat{C}_1}{\|\phi\|^2_{T} + \|\phi_h\|^2_{T}} - \frac{\epsilon_2 + \epsilon_3}{2} \right) > 0,
\]

and

\[
\left( \frac{\hat{C}_1}{\|\phi\|^2_{T} + \|\phi_h\|^2_{T}} \right) > 0,
\]
we have
\[
\| \phi - \phi_h \|_T^2 \leq C \left\{ \inf_{\varsigma_h \in \tilde{Z}_{2h}} \| \phi - \varsigma_h \|_T \right. \\
\left. + \inf_{(\tau_h, q_h) \in \tilde{Z}_{1h}} \left( \| \psi - \tau_h \|_{T'}^2 + \| \psi - \tau_h \|_{T'}^r + \| p - q_h \|_P^2 + \| p - q_h \|_{P'}^r \right) \right\}. \tag{4.4.21}
\]

Note that this also implies
\[
\int_\Omega |g(\phi) - g(\phi_h)| \, |\phi - \phi_h| \, d\Omega \leq C \left\{ \inf_{\varsigma_h \in \tilde{Z}_{2h}} \| \phi - \varsigma_h \|_T \right. \\
\left. + \inf_{(\tau_h, q_h) \in \tilde{Z}_{1h}} \left( \| \psi - \tau_h \|_{T'}^2 + \| \psi - \tau_h \|_{T'}^r + \| p - q_h \|_P^2 + \| p - q_h \|_{P'}^r \right) \right\}. \tag{4.4.22}
\]

The estimate (4.4.21) holds for \((\varsigma_h, \tau_h, q_h) \in \tilde{Z}_{2h} \times \tilde{Z}_{1h} \subseteq T_h \times T'_{\text{div}, h} \times P_h.\) In order to show that this estimate holds in all of \(T_h \times T'_{\text{div}, h} \times P_h,\) we employ a lifting argument similar to that in [30]. Define the subspace
\[
\tilde{W}_h := \left\{ \varsigma_h \in T_h : [B(\varsigma_h), (\tau_h, q_h)] + [(\tau_h, q_h), C^*(u_h, \lambda_h)] \\
= - \int_\Gamma (\tau_h \cdot \mathbf{n}) \cdot u_{\Gamma} \, d\Gamma \quad \forall (\tau_h, q_h) \in T'_{\text{div}, h} \times P_h \right\}.
\]

We first show that (4.4.21) holds for all \(\varsigma_h \in T_h.\) Then we show that (4.4.21) holds for all \((\tau_h, q_h) \in T'_{\text{div}, h} \times P_h.\)

Note that \(\varsigma_h \in \tilde{W}_h \Rightarrow \varsigma_h \in \tilde{Z}_{2h}.\) Thus, for \(v_h \in U_h,\)
\[
\inf_{\varsigma_h \in \tilde{Z}_{2h}} \| \phi - \varsigma_h \|_T \leq \inf_{\varsigma_h \in \tilde{W}_h} \| (\phi, u) - (\varsigma_h, v_h) \|_{T \times U}. \tag{4.4.23}
\]

From the inf-sup condition (4.4.11), there exist operators \(\Pi_T : T \rightarrow T_h\) and \(\Pi_U : U \rightarrow U_h\) such that
\[
[B(\varsigma - \Pi_T \varsigma), (\tau_h, q_h)] + [C(\tau_h, q_h), (v - \Pi_U v, \lambda_h)] = 0, \quad \forall (\tau_h, q_h) \in T'_{\text{div}, h} \times P_h, \tag{4.4.24}
\]
and
\[
\| (\Pi_T \varsigma, \Pi_U v) \|_{T \times U} \leq \tilde{C} \| (\varsigma, v) \|_{T \times U}, \quad \forall (\varsigma, v) \in T \times U. \tag{4.4.25}
\]
Now, let \((s_h, v_h) \in T_h \times U_h\) and set \(\tilde{\phi} := s_h - \Pi_T(s_h - \phi)\) and \(\tilde{u} := v_h - \Pi_U(v_h - u)\). Note that \((\tilde{\phi}, \tilde{u}) \in T_h \times U_h\). Then for all \((\tau_h, q_h) \in T'_{\text{div}, h} \times P_h\),

\[
[B(\tilde{\phi}), (\tau_h, q_h)] + [C(\tau_h, q_h), (\tilde{u}, \lambda_h)] = [B(\phi), (\tau_h, q_h)] + [C(\tau_h, q_h), (u, \lambda_h)]
\]

\[
= - \int_{\Gamma} (\tau_h \cdot n) \cdot u_h \, d\Gamma . \tag{4.4.26}
\]

Thus \(\tilde{\phi} \in \tilde{W}_h\). Now, using (4.4.25), we have

\[
\|(\tilde{\phi}, \tilde{u}) - (s_h, v_h)\|_{T \times U} = \|(\Pi_T(\phi - s_h), \Pi_U(u - v_h))\|_{T \times U}
\]

\[
\leq \tilde{C} \|(\phi - s_h, u - v_h)\|_{T \times U} . \tag{4.4.27}
\]

Thus we have

\[
\inf_{s_h \in Z_{2h}} \|\phi - s_h\|_T \leq \inf_{(s_h, v_h) \in W_{h} \times U_h} \|(\phi, u) - (s_h, v_h)\|_{T \times U}
\]

\[
\leq \inf_{(s_h, v_h) \in T_h \times U_h} \|(\phi, u) - (\tilde{\phi}, \tilde{u})\|_{T \times U}
\]

\[
\leq \inf_{(s_h, v_h) \in T_h \times U_h} \left( \|(\phi, u) - (s_h, v_h)\|_{T \times U} + \|(\tilde{\phi}, \tilde{u}) - (s_h, v_h)\|_{T \times U} \right)
\]

\[
\leq (1 + \tilde{C}) \inf_{(s_h, v_h) \in T_h \times U_h} \|(\phi, u) - (s_h, v_h)\|_{T \times U} , \tag{4.4.28}
\]

which lifts the best approximation of \(\phi\) from \(\tilde{Z}_{2h}\) to \(T_h\).

Now, we must also show

\[
\inf_{(\tau_h, q_h) \in \tilde{Z}_{1h}} \|(\psi, p) - (\tau_h, q_h)\|_{T'_{\text{div}, h} \times P} \leq C \inf_{(\tau_h, q_h) \in T'_{\text{div}, h} \times P_h} \|(\psi, p) - (\tau_h, q_h)\|_{T'_{\text{div}, h} \times P} . \tag{4.4.29}
\]

From (4.4.9), we have the existence of operators \(\Pi_{T'} : T'_{\text{div}, h} \to T'_{\text{div}, h}\) and \(\Pi_P : P \to P_h\) such that

\[
[C(\tau - \Pi_{T'}, \tau, q - \Pi_P q), (v_h, \eta_h)] = 0 , \quad \forall (v_h, \eta_h) \in U_h \times \mathbb{R} , \tag{4.4.30}
\]

and

\[
\|(\Pi_{T'} \tau, \Pi_P q)\|_{T'_{\text{div}, h} \times P} \leq \tilde{C} \|(\tau, q)\|_{T'_{\text{div}, h} \times P} . \tag{4.4.31}
\]

Now for \((\tau_h, q_h) \in T'_{\text{div}, h} \times P_h\), let \(\tilde{\psi} := \tau_h - \Pi_{T'}(\tau_h - \psi)\) and \(\tilde{p} := q_h - \Pi_P(q_h - p)\). Note that \((\tilde{\psi}, \tilde{p}) \in T'_{\text{div}, h} \times P_h\). Then for all \((v_h, \eta_h) \in U_h \times \mathbb{R}\) we have

\[
[C(\tilde{\psi}, \tilde{p}), (v_h, \eta_h)] = [C(\psi, p), (v_h, \eta_h)] = \int_{\Omega} v_h \cdot f \, d\Omega . \tag{4.4.32}
\]

105
So \((\tilde{\psi}, \tilde{p}) \in \tilde{Z}_{1h}\). Now, using (4.4.31) we have

\[
\| (\tilde{\psi}, \tilde{p}) - (\tau_h, q_h) \|_{T'_{\text{div}} \times P} = \| (\Pi_{T'} (\psi - \tau_h), \Pi_P (p - q_h)) \|_{T'_{\text{div}} \times P} \\
\leq \hat{C} \| (\psi - \tau_h, p - q_h) \|_{T'_{\text{div}} \times P}. \tag{4.4.33}
\]

Thus

\[
\inf_{(\tau_h, q_h) \in \tilde{Z}_{1h}} \| (\psi, p) - (\tau_h, q_h) \|_{T'_{\text{div}} \times P} \leq \inf_{(\tau_h, q_h) \in T'_{\text{div}} \times P_h} \| (\psi, p) - (\tilde{\psi}, \tilde{p}) \|_{T'_{\text{div}} \times P} \\
\leq (1 + \hat{C}) \inf_{(\tau_h, q_h) \in T'_{\text{div}} \times P_h} \| (\psi, p) - (\tau_h, q_h) \|_{T'_{\text{div}} \times P}. \tag{4.4.34}
\]

This lifts the best approximation of \((\psi, p)\) from \(\tilde{Z}_{1h}\) to \(T'_{\text{div}} \times P\).

Thus, from (4.4.21), (4.4.28), and (4.4.34) we have

\[
\| \phi - \phi_h \|^2_T \leq C \left\{ \inf_{\zeta_h \in T_h} \| \phi - \zeta_h \|_T^2 + \inf_{\mathbf{u}_h \in U_h} \| \mathbf{u} - \mathbf{v}_h \|_U^2 \\
+ \inf_{\tau_h \in T'_{\text{div}, h}} \left( \| \psi - \tau_h \|^2_{T'} + \| \tilde{\psi} - \tau_h \|^2_{T'} \right) \\
+ \inf_{q_h \in P_h} \left( \| p - q_h \|^2_{P} + \| p - q_h \|^2_{P} \right) \right\}. \tag{4.4.35}
\]
We have, from the discrete inf-sup condition for $B$,

$$
c_1 \left( \| \psi_h - \tau_h \|_{T'_{dV}} + \| p_h - q_h \|_P \right) \leq \sup_{\mathbf{s}_h \in T_h} \frac{\{ B(s_h), (\psi_h - \tau_h, p_h - q_h) \}}{\| s_h \|_T}
$$

\[
= \sup_{\mathbf{s}_h \in T_h} \left( \frac{-\int_{\Omega} \mathbf{s}_h : (\psi_h - \tau_h) \, d\Omega - \int_{\Omega} (p_h - q_h) \, tr(s_h) \, d\Omega}{\| \mathbf{s}_h \|_T} \right)
\]

\[
\leq \sup_{\mathbf{s}_h \in T_h} \left( \frac{-\int_{\Omega} \mathbf{s}_h : (\psi_h - \psi) \, d\Omega - \int_{\Omega} (p_h - p) \, tr(s_h) \, d\Omega}{\| \mathbf{s}_h \|_T} \right)
\]

\[
\leq \sup_{\mathbf{s}_h \in T_h} \left( \frac{\| \mathbf{B}(s_h), (\psi_h - \psi, p_h - p) \|}{\| \mathbf{s}_h \|_T} + \| \psi - \tau_h \|_{T'} + \sqrt{n} \| p - q_h \|_P \right)
\]

\[
= \sup_{\mathbf{s}_h \in T_h} \left( \frac{\| \mathbf{A}(\phi_h), s_h - [\mathbf{A}(\phi), s_h] \|}{\| \mathbf{s}_h \|_T} + \| \psi - \tau_h \|_{T'} + \sqrt{n} \| p - q_h \|_P \right)
\]

\[
= \sup_{\mathbf{s}_h \in T_h} \left( \frac{\int_{\Omega} (g(\phi_h) - g(\phi)) : s_h \, d\Omega}{\| \mathbf{s}_h \|_T} + \| \psi - \tau_h \|_{T'} + \sqrt{n} \| p - q_h \|_P \right).
\]  

The first term on the RHS of (4.4.36) can be handled using (4.2.6), (4.4.22), and the a priori bounds on $\phi$ and $\phi_h$:

\[
\sup_{\mathbf{s}_h \in T_h} \frac{\int_{\Omega} (g(\phi_h) - g(\phi)) : s_h \, d\Omega}{\| \mathbf{s}_h \|_T} \leq \sup_{\mathbf{s}_h \in T_h} \hat{C}_2 \left( \frac{\| \phi - \phi_h \|_{\infty}^{(2-r)/r} \left( \int_{\Omega} |g(\phi) - g(\phi_h)| \, |\phi - \phi_h| \, d\Omega \right)^{(r-1)/r} \| \mathbf{s}_h \|_T}{\| \mathbf{s}_h \|_T} \right)
\]

\[
\leq \hat{C}_2 \left( \int_{\Omega} |g(\phi) - g(\phi_h)| \, |\phi - \phi_h| \, d\Omega \right)^{1/r'} (as \ |\phi - \phi_h|/(|\phi| + |\phi_h|) \leq 1)
\]

\[
\leq C \left\{ \inf_{s_h \in Z_{2h}} \| \phi - s_h \|_T^{r'/r'} + \inf_{(\tau_h, q_h) \in Z_{1h}} \left( \| \psi - \tau_h \|_{T'}^{2/r'} + \| p - q_h \|_P^{2/r'} + \| p - q_h \|_P \right) \right\}. \]  

(4.4.37)
Combining (4.4.36), (4.4.37), and an application of the triangle inequality imply

\[
\|\psi - \psi_h\|_{T''_{\text{div}}} + \|p - p^h\|_P \\
\leq C \left\{ \inf_{\zeta_h \in \tilde{Z}_{2h}} \|\phi - \zeta_h\|_{T'}^{r'/r} + \inf_{(\tau_h, q_h) \in \tilde{Z}_{1h}} \left( \|\psi - \tau_h\|_{T'}^{2/r'} + \|\psi - \tau_h\|_{T'} \\
+ \|p - q_h\|_{P}^{2/r'} + \|p - q_h\|_P \right) \right\}. \quad (4.4.38)
\]

Now the previously described argument to lift the best approximations of \((\zeta_h, \tau_h, q_h)\) from \(\tilde{Z}_{1h} \times \tilde{Z}_{1h}\) to \(T_h \times T'_{\text{div},h} \times P_h\) can be applied here. Thus we have, from (4.4.38)

\[
\|\psi - \psi_h\|_{T''_{\text{div}}} + \|p - p^h\|_P \\
\leq C \left\{ \inf_{\zeta_h \in T_h} \|\phi - \zeta_h\|_{T'}^{r'/r} + \inf_{\varphi_h \in U_h} \|\mathbf{u} - \varphi_h\|_{U'}^{r'/r} + \inf_{\tau_h \in T'_{\text{div},h}} \left( \|\psi - \tau_h\|_{T'}^{2/r'} + \|\psi - \tau_h\|_{T'} \\
+ \inf_{q_h \in P_h} \left( \|p - q_h\|_{P}^{2/r'} + \|p - q_h\|_P \right) \right) \right\}. \quad (4.4.39)
\]

From the discrete inf-sup condition for \(C\) we have,

\[
c_2(\|\mathbf{u}_h - \mathbf{v}_h\|_U + |\lambda_h - \lambda|) \leq \sup_{(\tau_h, q_h) \in T'_{\text{div},h} \times P_h} \frac{[C(\tau_h, q_h), (\mathbf{u}_h - \mathbf{v}_h, \lambda_h - \lambda)]}{\|\tau_h\|_{T'_{\text{div},h} \times P_h}}
\leq \sup_{(\tau_h, q_h) \in T'_{\text{div},h} \times P_h} \frac{-\int_{\Omega} (\mathbf{u}_h - \mathbf{u}) \cdot \text{div}(\tau_h) \, d\Omega - (\lambda_h - \lambda) \int_{\Omega} \text{tr}(\tau_h) \, d\Omega}{\|\tau_h\|_{T'_{\text{div},h} \times P_h}} + \frac{\|\mathbf{u} - \mathbf{v}_h\|_U}{\|\tau_h\|_{T'_{\text{div}}}^2} + \|q_h\|_P. \quad (4.4.40)
\]

Now the first term on the RHS of (4.4.40) gives

\[
\sup_{(\tau_h, q_h) \in T'_{\text{div},h} \times P_h} \frac{-\int_{\Omega} (\mathbf{u}_h - \mathbf{u}) \cdot \text{div}(\tau_h) \, d\Omega - (\lambda_h - \lambda) \int_{\Omega} \text{tr}(\tau_h) \, d\Omega}{\|\tau_h\|_{T'_{\text{div},h} \times P_h}} = \sup_{(\tau_h, q_h) \in T'_{\text{div},h} \times P_h} \frac{[C(\tau_h, q_h), (\mathbf{u}_h - \mathbf{u}, \lambda_h - \lambda)]}{\|\tau_h\|_{T'_{\text{div},h} \times P_h}}
= \sup_{(\tau_h, q_h) \in T'_{\text{div},h} \times P_h} \frac{[\mathbf{B}(\varphi_h - \varphi), (\tau_h, q_h)]}{\|\tau_h\|_{T'_{\text{div},h} \times P_h}} \leq \frac{\|\varphi - \varphi_h\|_{T'} \|\tau_h\|_{T'} + \sqrt{n} \|q_h\|_P \|\varphi - \varphi_h\|_{T}}{\|\tau_h\|_{T'_{\text{div}}} + \|q_h\|_P}
\leq \|\varphi - \varphi_h\|_T + \sqrt{n} \|\varphi - \varphi_h\|_T \leq C \|\varphi - \varphi_h\|_T. \quad (4.4.41)
\]
Thus, from (4.4.35), (4.4.40), (4.4.41), and the triangle inequality we have

\[
\|u - uh\|_U + |\lambda - \lambda_h| \leq C \left\{ \inf_{\varsigma_h \in T_h} \|\phi - \varsigma_h\|^{r/2}_{T} + \inf_{v_h \in U_h} \left( \|u - v_h\|^{r/2}_{U} + \|u - v_h\|_U \right) + \inf_{\tau_h \in T_{\text{div, } h}} \left( \|\psi - \tau_h\|_{T'} + \|\psi - \tau_h\|^{r'/2}_{T'} \right) + \inf_{q_h \in P_h} \left( \|p - q_h\|_{P} + \|p - q_h\|^{r'/2}_{P} \right) \right\}.
\]

(4.4.42)

Thus the estimates (4.4.12) and (4.4.13) are proven.

### 4.4.2 Discrete Approximation Spaces

Let \( n = 2 \). Let \( K \in T_h \) and let \( P_k(K) \) be the set of all polynomials in the variables \( x_1, x_2 \) of degree less than or equal to \( k \) defined on the triangle \( K \). Let \( RT_k(K) \) be the 2-vector of Raviart-Thomas elements \([67, 73]\) on \( K \) defined by

\[
RT_k(K) = (P_k(K))^2 + \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mathcal{P}_k(K).
\]

Let the following discrete spaces be defined as:

\[
T_h := \{ \phi \in T : \phi|_K \in \mathcal{P}_0(K)^{2 \times 2}, \ \forall K \in T_h \},
\]

\[
T'_{\text{div, } h} := \left\{ \psi \in T'_{\text{div}} : \psi = (\psi_1 \ psi_2)^T|_K \in (RT_0(K))^2, \right. \]

\[
(\psi_1 \ psi_2)^T|_K \in RT_0(K), \ \forall i \in \{1, 2\}, \ \forall K \in T_h \right\},
\]

\[
P_h := \{ p \in P : p|_K \in \mathcal{P}_0(K), \ \forall K \in T_h \},
\]

\[
U_h := \{ u \in U : u|_K \in \mathcal{P}_0(K)^2, \ \forall K \in T_h \}.
\]

Let \( s > 1 \) and let \( \mathcal{T}_0^h : (W^{1,s} (\Omega))^{2 \times 2} \rightarrow T'_{\text{div, } h} \) be the lowest-order Raviart-Thomas interpolation operator \([67, 13, 26]\) defined by, for row \( j = 1, 2 \) of \( \tau \in T'_{\text{div}}, \)

\[
\int_{e_i} (\tau_j - \mathcal{T}_0^h \tau_j) \cdot n_{e_i} \, ds = 0, \ \forall e_i \in \partial K, \ \forall K \in T_h,
\]

(4.4.43)

where \( n_{e_i} \) denotes the outer unit normal vector to edge \( e_i \) of \( K \). For every \( K \in T_h \), \( \mathbf{\tau}|_K \in \left(W^{1,s}(K)\right)^{2 \times 2} \), and thus \( \mathbf{\tau}|_{\partial K} \in \left(W^{1-1/s,s}(\partial K)\right)^{2 \times 2} \subset \left(L^s(\partial K)\right)^{2 \times 2} \subset \left(L^1(\partial K)\right)^{2 \times 2} \).
Thus we have, if \( \text{div } \tau \in \left( W^{m,r'}(\Omega) \right)^2 \) with \( 0 \leq m \leq 1 \),
\[
\| \tau - T_0^0 \tau \|_{0,r',\Omega} \leq C h^m |\tau|_{m,r',\Omega}, 
\] (4.4.44)
\[
\| \text{div} (\tau - T_0^0 \tau) \|_{0,r',\Omega} \leq C h^m |\text{div } \tau|_{m,r',\Omega},
\] (4.4.45)
and, for \( v \in U \),
\[
\int_{\Omega} v \cdot \text{div} (\tau - T_0^0 \tau) \, d\Omega = 0, \quad \forall \tau \in T'_\text{div}.
\] (4.4.46)

**Lemma 4.4.1** For the choices of \( T_h, T'_\text{div}, P_h \), and \( U_h \) above, there exists a positive constant \( c_1 \) such that
\[
\inf_{(\tau_h,q_h) \in Z_h} \sup_{\phi_h \in T_h} \frac{[B(\phi_h), (\tau_h,q_h)]}{\| \phi_h \|_T \| (\tau_h,q_h) \|_{T'_\text{div} \times P}} \geq c_1 .
\] (4.4.47)

**Proof:** Note that for \( \tau_h \in T'_\text{div}, \text{div } \tau_h = 0 \) implies \( \tau_h \in (P_0(K))^{2 \times 2} \) for all \( K \in T_h \). We consider two cases.

**Case 1:** \( \|q_h\|_P \leq \|\tau_h\|_{T'_\text{div}} \).
Let \( \tau_h^0 = \tau_h - \frac{1}{h} \text{tr}(\tau_h) I \), and \( \phi_h = -|\tau_h^0|^{r'/r-1} \tau_h^0 / \|\tau_h^0\|_{T'}^{r'-1} \). Note that \( \phi_h \in T_h \) as \((\tau_h,q_h) \in Z_h \) implies \( \tau_h^0 \in (P_0(K))^{2 \times 2} \). Proceeding as in Case 1 of the proof of Lemma 4.3.3, we obtain (4.4.47).

**Case 2:** \( \|q_h\|_P \geq \|\tau_h\|_{T'_\text{div}} \).
Let
\[
\phi_h = \frac{-q_h I + \tau_h |q_h|^{r'/r-1}(q_h I + \tau_h)}{\|q_h I + \tau_h\|_{T'}^{r'-1}} .
\]
Again \( \phi_h \in T_h \) as \( q_h \in P_0(K) \) and \( \tau_h \in (P_0(K))^{2 \times 2} \) for all \( K \in T_h \). Proceeding as in Case 2 of the proof of Lemma 4.3.3, we obtain (4.4.47).

**Lemma 4.4.2** For the choices of \( T_h, T'_\text{div}, P_h \), and \( U_h \) above, there exists a positive constant \( c_2 \) such that
\[
\inf_{(u_h,\lambda_h) \in U_h \times \mathbb{R}} \sup_{(\tau_h,q_h) \in T'_\text{div} \times P_h} \frac{[C(\tau_h,q_h), (u_h,\lambda_h)]}{\| (\tau_h,q_h) \|_{T'_\text{div} \times P} \| (u_h,\lambda_h) \|_{U \times \mathbb{R}}} \geq c_2 .
\] (4.4.48)

**Proof:** As in the approach to the proof of Lemma 4.3.6, we consider two cases:

**Case 1:** \( |\lambda_h| \geq \|u_h\|_U \).

The choice \( (\tau_h,q_h) = (\lambda_h I, 0) \in T'_\text{div} \times P_h \) shows the result as in Case 1 of the proof of
Lemma 4.3.6.

Case 2: $|\lambda_h| \leq \|u_h\|_U$.

Note that Lemma 4.3.4 applies to the subspace $T_{\text{div},h}^\prime \subset T_{\text{div}}^\prime$, thus we have

$$
sup_{(\tau_h,q_h) \in T_{\text{div},h}^\prime \times P_h} \frac{[C(\tau_h, q_h), (u_h, \lambda_h)]}{\| (\tau_h, q_h) \|_{T_{\text{div}}^\prime \times P}} \geq sup_{\tau_0 \in \partial \tau_{\text{div},h}^\prime} \frac{[C(\tau_0, 0), (u_h, \lambda_h)]}{\| \tau_0 \|_{T_{\text{div}}^\prime}} = sup_{\tau \in T_{\text{div},h}^\prime} \frac{-\int_{\Omega} u_h \cdot \text{div} \tau_0 d\Omega}{\| \tau_0 \|_{T_{\text{div},h}^\prime}} \geq C sup_{\tau \in T_{\text{div},h}^\prime} \frac{-\int_{\Omega} u_h \cdot \text{div} \tau_h d\Omega}{\| \tau_h \|_{T_{\text{div}}^\prime}}. \quad (4.4.49)
$$

Now we proceed in a manner similar to that of Proposition 5 of [59] (as well as Proposition 3.1 of [32]). Let $w$ be the solution of the Laplacian problem

$$
-\Delta w = |u_h|^{r-2}u_h, \quad \text{in } \Omega,
$$

$$
w = 0, \quad \text{on } \Gamma.
$$

Note that $|u_h|^{r-2}u_h \in (W^{0,r} (\Omega))^2$. Hence, from [38], this problem has a unique solution $w \in (W_0^{2,r} (\Omega))^2$, and there exists a constant $C > 0$ such that

$$
\|w\|_{2,r',\Omega} \leq C \|u_h|^{r-2}u_h\|_{0,r',\Omega}
$$

$$
= C \left( \int_{\Omega} |u_h|^{r-2}u_h|^{r'} d\Omega \right)^{1/r'}
$$

$$
= C \left( \int_{\Omega} |u_h|^{r'(r-1)} d\Omega \right)^{(r-1)/r}
$$

$$
= C \left( \int_{\Omega} |u_h|^r d\Omega \right)^{(1/r)(r-1)} = C \|u_h\|_{0,r,\Omega}^{r-1}. \quad (4.4.50)
$$

Now let $\tau^* = \nabla w$. Thus from (4.4.50) we have

$$
\|\tau^*\|_{1,r',\Omega} = \|w\|_{2,r',\Omega} \leq C \|u_h\|_{0,r,\Omega}^{r-1}, \quad (4.4.51)
$$

and $\nabla \cdot \tau^* = \Delta w = -|u_h|^{r-2}u_h$. Thus $\tau^* \in T_{\text{div}}^\prime$ and we have

$$
\|\tau^*\|_{T_{\text{div}}^\prime} \leq C \|u_h\|_{0,r,\Omega}^{r-1}. \quad (4.4.52)
$$

111
Let $\tau_h = T_{0}^{0} \tau^*$.

Then we have, from (4.4.46), that

$$- \int_{\Omega} \mathbf{u}_h \cdot \text{div} \, \tau_h \, d\Omega = - \int_{\Omega} \mathbf{u}_h \cdot \text{div} \, \tau^* \, d\Omega$$

$$= \int_{\Omega} |\mathbf{u}_h|^{r-2} \mathbf{u}_h \cdot \mathbf{u}_h \, d\Omega$$

$$= \|\mathbf{u}_h\|_U^r. \quad (4.4.53)$$

The estimate (4.4.45) and the reverse triangle inequality gives, for $m = 0$,

$$\|\nabla \cdot \tau_h\|_{0,r',\Omega} \leq C \|\nabla \cdot \tau^*\|_{0,r',\Omega}. \quad (4.4.54)$$

Then, from (4.4.44), (4.4.52), (4.4.54), and the triangle inequality, we have

$$\|\tau_h\|_{T'_{\text{div}}} \leq \|\tau^*\|_{0,r',\Omega} + \|\nabla \cdot \tau_h\|_{0,r',\Omega}$$

$$\leq C \left( \|\tau^*\|_{0,r',\Omega} + \|\tau^* - \tau_h\|_{0,r',\Omega} + \|\nabla \cdot \tau^*\|_{0,r',\Omega} \right)$$

$$\leq C \left( \|\tau^*\|_{T'_{\text{div}}} + \|\tau^*\|_{1,r',\Omega} + \|\mathbf{u}\|_{0,r',\Omega}^{r-1} \right)$$

$$\leq C \left( \|\mathbf{u}_h\|_{0,r',\Omega}^{r-1} + h \|\mathbf{u}_h\|_{0,r',\Omega}^{r-1} + \|\mathbf{u}_h\|_{0,r',\Omega}^{r-1} \right)$$

$$\leq C \|\mathbf{u}_h\|_U^{r-1}. \quad (4.4.55)$$

Combining (4.4.49), (4.4.53), and (4.4.55), we have that

$$\sup_{(\tau_h,q_h) \in T'_{\text{div},h} \times P_h} \left[ \frac{|C(\tau_h,q_h), (\mathbf{u}_h,\lambda_h)|}{\|\tau_h\|_{T'_{\text{div}}} \times P_h} \right] \geq - \int_{\Omega} \mathbf{u}_h \cdot \text{div} \, \tau_h \, d\Omega \geq \frac{\|\mathbf{u}_h\|_U}{C \|\mathbf{u}_h\|_U^{r-1}}$$

$$\geq C \|\mathbf{u}_h\|_U \geq c_2 \|(\mathbf{u}_h,\lambda_h)\|_{U \times \mathbb{R}},$$

and thus we obtain (4.4.48).

In order to apply the abstract a priori estimate from Theorem 4.4.2, we must first show that the chosen approximation spaces satisfy the condition (4.4.11).

**Lemma 4.4.3** For $h$ sufficiently small, there is a constant $c_3 > 0$ such that

$$\inf_{(\tau_h,q_h) \in T'_{\text{div},h} \times P_h} \sup_{(\mathbf{v}_h,\eta_h) \in T \times U \times \mathbb{R}} \left[ |B(\mathbf{s}_h, (\tau_h,q_h)), (\mathbf{v}_h,\eta_h)| + |C(\tau_h,q_h), (\mathbf{v}_h,\eta_h)| \right] \geq c_3, \quad (4.4.56)$$

where $\|(\mathbf{s}_h,\mathbf{v}_h,\eta_h)\|_{T \times U \times \mathbb{R}} = \|\mathbf{s}_h\|_T + \|\mathbf{v}_h\|_U + \|\eta_h\|_{\mathbb{R}}$. 

112
**Proof:** If \((\tau_h, q_h) \in Z_{1h}\), then for all \((v_h, \eta_h) \in U_h \times \mathbb{R}\), we have \([C(\tau_h, q_h), (v_h, \eta_h)] = 0\), thus (4.4.56) follows immediately from Lemma 4.4.1.

For \(K \in T_h\), let \(\Pi_{0,K} : T'_\text{div,} h(K) \rightarrow T_h(K)\) be the \(P_0(K)\) interpolation operator defined by

\[
\int_K (\tau_h - \Pi_{0,K} \tau_h) \, dK = 0, \quad \forall \tau_h \in T'_\text{div,} h(K).
\]

For \(\tau_h \in T'_\text{div,} h(\Omega)\), let \(\hat{\tau} = \Pi_0 \tau_h = \bigcup_{K \in T_h} \Pi_{0,K} \tau_h\). From [12, 26], we have that there exists a constant \(\hat{C}\) such that

\[
\|\tau_h - \hat{\tau}\|_{0,r',\Omega} = \left( \sum_{K \in T_h} \|\tau_h - \hat{\tau}\|_{0,r',K}^{r'} \right)^{1/r'} \leq \hat{C} h \left( \sum_{K \in T_h} |\tau_h|_{1,r',K}^{r'} \right)^{1/r'}
\]

\[
= \hat{C} h \left( \sum_{K \in T_h} \|\nabla \tau_h\|_{0,r',K}^{r'} \right)^{1/r'}. \quad (4.4.57)
\]

Note that, since \(\tau_h|_K \in (RT_0(K))^2\) for all \(K \in T_h\), the partial derivatives of \(\nabla \tau_h\) that are not also present in \(\text{div} \ \tau_h\) are zero, and thus

\[
\left( \sum_{K \in T_h} \|\nabla \tau_h\|_{0,r',K}^{r'} \right)^{1/r'} = \left( \sum_{K \in T_h} (\sqrt{n} \|\text{div} \ \tau_h\|_{0,r',K})^{r'} \right)^{1/r'}
\]

\[
= \sqrt{n} \|\text{div} \ \tau_h\|_{0,r',\Omega}. \quad (4.4.58)
\]

Combining (4.4.57) and (4.4.58) we can bound the error in approximating \(\tau_h\) with \(\hat{\tau}\) by

\[
\|\tau_h - \hat{\tau}\|_{0,r',\Omega} \leq \sqrt{n} \hat{C} h \|\text{div} \ \tau_h\|_{0,r',\Omega}. \quad (4.4.59)
\]

We will assume that \(\text{div} \ \tau \neq 0\), for if \(\text{div} \ \tau = 0\) then \(\tau\) is piecewise constant and \(\hat{\tau} = \tau\).

**Case 1:** \(\|\tau_h\|_{T'_\text{div}} \leq \|q_h\|_P\)

Let

\[
\mathbf{s}_h = \frac{-1}{\|q_h\mathbf{I} + \hat{\tau}\|_{T'_\text{div}}} |q_h\mathbf{I} + \hat{\tau}|^{r'/r-1}(q_h\mathbf{I} + \hat{\tau}).
\]
Note that \(s_h \in T_h\) and \(\|s_h\|_T = 1\). Then we have

\[
[B(s_h), (\tau_h, q_h)] = \int_{\Omega} \frac{q_h \mathbf{I} + \hat{\tau}}{\|q_h \mathbf{I} + \hat{\tau}\|_{T'}}^{|r'|/r-1} (\tau_h : (q_h \mathbf{I} + \hat{\tau}) + q_h \text{tr}(q_h \mathbf{I} + \hat{\tau})) \, d\Omega
\]

\[
= \int_{\Omega} \frac{q_h \mathbf{I} + \hat{\tau}}{\|q_h \mathbf{I} + \hat{\tau}\|_{T'}}^{|r'|/r-1} ((q_h \mathbf{I} + \hat{\tau}) : (q_h \mathbf{I} + \hat{\tau}) + (q_h \mathbf{I} + \hat{\tau}) : (\tau_h - \hat{\tau})) \, d\Omega
\]

\[
= \|q_h \mathbf{I} + \hat{\tau}\|_{T'} - \int_{\Omega} \frac{q_h \mathbf{I} + \hat{\tau}}{\|q_h \mathbf{I} + \hat{\tau}\|_{T'}}^{|r'|/r-1} ((q_h \mathbf{I} + \hat{\tau}) : (\tau_h - \hat{\tau})) \, d\Omega
\]

\[
\geq \|q_h \mathbf{I} + \hat{\tau}\|_{T'} - \|s_h\|_T \|\tau_h - \hat{\tau}\|_{T'}
\]

\[
\geq n^{1/r'} \|q_h\|_P - \|\mathbf{I}\|_{T'} - 2\sqrt{n} \hat{C} \|\text{div} \tau_h\|_{0,r',\Omega}
\]

\[
\geq n^{1/r'} \|q_h\|_P - \|\tau_h\|_{T'} - 2\sqrt{n} \hat{C} \|\text{div} \tau_h\|_{0,r',\Omega}
\]

\[
\geq (n^{1/r'} - 1) \|q_h\|_P - 2\sqrt{n} \hat{C} \|\text{div} \tau_h\|_{0,r',\Omega}.
\]

(4.4.60)

Let

\[
v_h = \frac{-(n^{1/r'} - 1)|\text{div} \tau_h|_{T')/r-1} (\text{div} \tau_h),
\]

and note that \(v_h \in U_h\) and \(\|v_h\|_U = n^{1/r'} - 1\). Let \(\eta_h = 0\), then we have

\[
[C(\tau_h, q_h), (v_h, \eta_h)] = (n^{1/r'} - 1) \int_{\Omega} |\text{div} \tau_h|_{T')/r-1} (\text{div} \tau_h) \cdot (\text{div} \tau_h) \, d\Omega
\]

\[
= (n^{1/r'} - 1) \|\text{div} \tau_h\|_{0,r',\Omega}.
\]

(4.4.61)

Thus, from (4.4.60) and (4.4.61), we have

\[
\frac{[B(s_h), (\tau_h, q_h)] + [C(\tau_h, q_h), (v_h, \eta_h)]}{\|(s_h, v_h, \eta_h)\|_{T \times U \times R}} \geq \frac{1}{n^{1/r'}} \left( (n^{1/r'} - 1) \|q_h\|_P + (n^{1/r'} - 1 - 2\sqrt{n} \hat{C}) \|\text{div} \tau_h\|_{0,r',\Omega} \right),
\]

(4.4.62)

and, for \(h\) small enough to satisfy \(n^{1/r'} - 1 - 2\sqrt{n} \hat{C} > 0\), we have that

\[
\frac{[B(s_h), (\tau_h, q_h)] + [C(\tau_h, q_h), (v_h, \eta_h)]}{\|(s_h, v_h, \eta_h)\|_{T \times U \times R}} \geq C\|(\tau_h, q_h)\|_{T' \times 0},
\]

for some constant \(C > 0\).

Case 2: \((\|\tau_h\|_{T'} \geq \|q_h\|_P)\)

First note that for \(\tau \in T'_{\text{div}}\), we have by Hölder’s inequality,

\[
\int_{\Omega} \text{tr}(\tau) \, d\Omega \leq \sqrt{n} |\Omega|^{1/r} \|\tau\|_{T'},
\]

114
and thus
\[
\left| \int_{\Omega} tr( \hat{\tau} - \tau_h ) \, d\Omega \right| \leq \hat{C} n |\Omega|^{1/r} h \| \text{div } \tau_h \|_{0,r',\Omega} . \tag{4.4.63}
\]

Given the piecewise constant interpolant $\hat{\tau}$, let
\[
\hat{\tau} = \hat{\tau} - \frac{1}{n} \left( \int_{\Omega} tr( \hat{\tau} ) \, d\Omega \right) \mathbf{I},
\]
and note that $\text{div } \hat{\tau} = \text{div } \tau$. From (4.4.63) we have
\[
\| \hat{\tau} - \tau \|_{T'} = \left| \frac{1}{n} \left( \int_{\Omega} tr( \hat{\tau} ) \, d\Omega \right) \mathbf{I} \right|_{T'} \leq \sqrt{n} \hat{C} |\Omega| h \| \text{div } \tau_h \|_{0,r',\Omega} + n^{-1/2} |\Omega|^{1/r'} \left| \int_{\Omega} tr( \tau_h ) \, d\Omega \right| . \tag{4.4.64}
\]

We have, from (4.4.59) and (4.4.64),
\[
\| \hat{\tau} - \tau_h \|_{T'} = \| \hat{\tau} - \tau \|_{T'} + \| \hat{\tau} - \tau_h \|_{T'} \leq \sqrt{n} \hat{C} (1 + |\Omega|) h \| \text{div } \tau_h \|_{0,r',\Omega} + n^{-1/2} |\Omega|^{1/r'} \left| \int_{\Omega} tr( \tau_h ) \, d\Omega \right| . \tag{4.4.65}
\]

Lemma 4.3.2 applies to $\hat{\tau}$ and we have that, for
\[
\tau^0 = \hat{\tau} - \frac{1}{n} tr( \hat{\tau} ) \mathbf{I},
\]
there exists a constant $C_0$ such that
\[
\| \hat{\tau} \|_{0,r'} \leq C_0 ( \| \tau^0 \|_{0,r'} + \| \text{div } \hat{\tau} \|_{-1,r'} ) , \tag{4.4.66}
\]
and, since $\text{div } \hat{\tau} = \text{div } \tau$, we have that
\[
\| \text{div } \tau \|_{-1,r'} \leq \| \text{div } \tau_h \|_{-1,r'} + \| \text{div } ( \tau_h - \hat{\tau} ) \|_{-1,r'} \leq \| \text{div } \tau_h \|_{0,r'} + \| \text{div } ( \tau_h - \hat{\tau} ) \|_{-1,r'} . \tag{4.4.67}
\]
Observe that
\[
\| \text{div} (\tau_h - \hat{\tau}) \|_{-1,r'} = \sup_{g \in W_0^{1,r}(\Omega)} \frac{\langle \text{div} (\tau_h - \hat{\tau}), g \rangle}{\| g \|_{1,r}} = \sup_{g \in W_0^{1,r}(\Omega)} \frac{-\langle \tau_h - \hat{\tau}, \nabla g \rangle}{\| g \|_{1,r}} \leq \| \tau_h - \hat{\tau} \|_{T'} .
\]
(4.4.68)

Combining (4.4.66) - (4.4.68) we have
\[
\| \tilde{\tau} \|_{0,r'} \leq C_0 \left( \| \tau_0 \|_{0,r'} + \| \text{div} \tau_h \|_{0,r'} + \sqrt{n} \hat{C} h \| \text{div} \tau_h \|_{0,r'} \right)
\leq C_0 \left( \| \tau_0 \|_{0,r'} + (1 + \sqrt{n} \hat{C} h) \| \text{div} \tau_h \|_{0,r'} \right) .
\]
(4.4.69)

Rearranging and using (4.4.65),
\[
\| \tau_0 \|_{T'} \geq \frac{1}{C_0} \| \tilde{\tau} \|_{T'} - (1 + \sqrt{n} \hat{C} h) \| \text{div} \tau_h \|_{0,r'} \\
\geq \frac{1}{C_0} \left( \| \tau_h \|_{T'} - \| \tilde{\tau} - \tau_h \|_{T'} \right) - (1 + \sqrt{n} \hat{C} h) \| \text{div} \tau_h \|_{0,r'} \\
\geq \frac{1}{C_0} \| \tau_h \|_{T'} - \left( 1 + \frac{\sqrt{n} \hat{C} (1 + C_0 + |\Omega|) h}{C_0} \right) \| \text{div} \tau_h \|_{0,r'} \\
- \frac{n^{-1/2} |\Omega|^{1/r'}}{C_0} \left( \int_{\Omega} \text{tr}(\tau_h) d\Omega \right) .
\]
(4.4.70)

Let
\[
\varsigma_h = -\frac{|\tau_0|_{r'/r'-1}}{\| \tau_0 \|_{T'}^{r'/r'-1}} \tau_0 ,
\]
and note that \( \varsigma_h \in T_h, \text{tr}(\varsigma_h) = 0, \) and \( \| \varsigma_h \|_T = 1. \) Let
\[
\nu_h = -2 \frac{|\text{div} \tau_h|_{r'/r'-1}}{\| \text{div} \tau_h \|_{0,r',\Omega}^{r'-1}} (\text{div} \tau_h) ,
\]
and note that \( \nu_h \in U_h \) and \( \| \nu_h \|_U = 2. \) Let
\[
\eta_h = \text{sgn} \left( \int_{\Omega} \text{tr}(\tau_h) d\Omega \right) \left( 1 + \frac{1}{C_0} \right) n^{-1/2} |\Omega|^{1/r'} .
\]
Then we have, from (4.4.65) and (4.4.70),

\[
\mathbf{B}(s_h, (\tau_h, q_h)) = \int_\Omega \frac{|\tau_0|^{r'/r-1}}{||\tau_0||_T^{r'/r-1}} \tau_0 : \tau_h \, d\Omega
\]

\[
= \int_\Omega \frac{|\tau_0|^{r'/r-1}}{||\tau_0||_T^{r'/r-1}} \tau_0 : \tilde{\tau} \, d\Omega - \int_\Omega \frac{|\tau_0|^{r'/r-1}}{||\tau_0||_T^{r'/r-1}} \tau_0 : (\tilde{\tau} - \tau_h) \, d\Omega
\]

\[
\geq \int_\Omega \frac{|\tau_0|^{r'/r-1}}{||\tau_0||_T^{r'/r-1}} \tau_0 : \tau_0 \, d\Omega - ||s_h||_T ||\tilde{\tau} - \tau_h||_{T'}
\]

\[
\geq \frac{1}{C_0} ||\tau_h||_{T'} - \left(1 + \sqrt{n} \hat{C} \left(1 + C_0 + |\Omega| / C_0\right) \right) ||\text{div} \, \tau_h||_{0,r',\Omega}
\]

\[
- \frac{n^{-1/2}|\Omega|^{1/r'}}{C_0} \left|\int_\Omega \text{tr}(\tau_h) \, d\Omega \right|
\]

\[
- \sqrt{n} \hat{C} \left(1 + |\Omega| / C_0\right) h ||\text{div} \, \tau_h||_{0,r',\Omega} - n^{-1/2}|\Omega|^{1/r'} \left|\int_\Omega \text{tr}(\tau_h) \, d\Omega \right|
\]

\[
= \frac{1}{C_0} ||\tau_h||_{T'} - \left(1 + \sqrt{n} \hat{C} h \left(1 + 2C_0 + (1 + C_0)|\Omega| / C_0\right) \right) ||\text{div} \, \tau_h||_{0,r',\Omega}
\]

\[
- \left(1 + \frac{1}{C_0}\right) n^{-1/2}|\Omega|^{1/r'} \left|\int_\Omega \text{tr}(\tau_h) \, d\Omega \right|
\]

and

\[
\mathbf{C}(\tau_h, q_h, (v_h, \eta_h)) = \int_\Omega \frac{2||\text{div} \, \tau_h||^{r'/r-1}}{||\text{div} \, \tau_h||_{0,r',\Omega}^{r'/r-1}} \left(\text{div} \, \tau_h\right) \cdot (\text{div} \, \tau_h) \, d\Omega
\]

\[
+ \left(1 + \frac{1}{C_0}\right) n^{-1/2}|\Omega|^{1/r'} \left|\int_\Omega \text{tr}(\tau_h) \, d\Omega \right|
\]

\[
= 2||\text{div} \, \tau_h||_{0,r',\Omega} + \left(1 + \frac{1}{C_0}\right) n^{-1/2}|\Omega|^{1/r'} \left|\int_\Omega \text{tr}(\tau_h) \, d\Omega \right|
\]

and

\[
||s_h||_T + ||v_h||_U + |\eta_h| = 3 + \left(1 + \frac{1}{C_0}\right) n^{-1/2}|\Omega|^{1/r'} = \hat{C}.
\]

Thus, (4.4.71)-(4.4.73) and \(h\) small enough to guarantee that

\[
C_0 > \sqrt{n} \hat{C} h (1 + 2C_0 + (1 + C_0)|\Omega|) \]

117
Let solve (4.4.5)-(4.4.7). Assume Theorem 4.4.3, Lemmas 4.4.1, 4.4.2, and the properties (4.4.75)-(4.4.79). Thus we have the following a priori error estimate, the proof of which follows directly from (4.4.56) is shown.

Standard approximation properties for the discrete spaces are shown in [26, 73]:

For all \((\varsigma, \tau, q, v) \in (W^{1,r}(\Omega))^{2 \times 2} \times \left(\left(W^{1,r'}(\Omega)\right)^{2 \times 2} \times W^{1,r'}(\Omega) \times (W^{1,r}(\Omega))^2\right)\), there exists \((\varsigma_h, \tau_h, q_h, v_h) \in T_h \times T'_{div,h} \times P_h \times U_h\) satisfying

\[
\|\varsigma - \varsigma_h\|_T \leq C h \|\varsigma\|_{(W^{1,r}(\Omega))^{2 \times 2}}, \quad \forall \varsigma \in (W^{1,r}(\Omega))^{2 \times 2},
\]
\[
\|\tau - \tau_h\|_{T'} \leq C h \|\tau\|_{\left(W^{1,r'}(\Omega)\right)^{2 \times 2}}, \quad \forall \tau \in \left(W^{1,r'}(\Omega)\right)^{2 \times 2},
\]
\[
\|\text{div } (\tau - \tau_h)\|_{T'} \leq C h \|\text{div } \tau\|_{\left(W^{1,r'}(\Omega)\right)^2}, \quad \forall (\text{div } \tau) \in \left(W^{1,r'}(\Omega)\right)^2,
\]
\[
\|q - q_h\|_P \leq C h \|q\|_{L^r(\Omega)}, \quad \forall q \in W^{1,r}(\Omega),
\]
\[
\|v - v_h\|_U \leq C h \|v\|_{(W^{1,r}(\Omega))^2}, \quad \forall v \in (W^{1,r}(\Omega))^2.
\]

Thus we have the following a priori error estimate, the proof of which follows directly from Theorem 4.4.2, Lemmas 4.4.1, 4.4.2, and 4.4.3, and the properties (4.4.75)-(4.4.79).

**Theorem 4.4.3** Let \(f \in \left(L^r(\Omega)\right)^2\) and \(u_f \in (W^{1-1/r,r}(\Gamma))^2\). Let \((\phi, \psi, p, u, \lambda) \in T \times T'_{div} \times P \times U \times \mathbb{R}\) solve (4.2.16)-(4.2.18) and let \((\phi_h, \psi_h, p_h, u_h, \lambda_h) \in T_h \times T'_{div,h} \times P_h \times U_h \times \mathbb{R}\) solve (4.4.5)-(4.4.7). Assume \((\phi, \psi, p, u) \in (W^{1,r}(\Omega))^{2 \times 2} \times \left(\left(W^{1,r'}(\Omega)\right)^{2 \times 2} \times W^{1,r'}(\Omega) \times (W^{1,r}(\Omega))^2\right)\) with \(\text{div } \psi \in \left(W^{1,r'}(\Omega)\right)^2\). Then there exists a positive constant \(C\) such that

\[
\|\phi - \phi_h\|_T + \|u - u_h\|_U + |\lambda - \lambda_h| \leq C \left\{ h^{r/2} \left( \|\phi\|_{(W^{1,r}(\Omega))^{2 \times 2}} + \|u\|_{(W^{1,r}(\Omega))^2} \right) + h \left( \|\psi\|_{(W^{1,r'}(\Omega))^2} + \|p\|_{L^r(\Omega)} \right) \right\}, \quad (4.4.80)
\]
\[
\|\psi - \psi_h\|_{T_{\text{div}}'} + \|p - p_h\|_P \leq C \left\{ h^{r'/r} \left( \|\phi\|_{(W^{1,r}(\Omega))^2} + \|u\|_{(W^{1,r}(\Omega))^2} \right) + h^{2/r'} \left( \|\psi\|_{(W^{1,r}(\Omega))^2} + \|p\|_{L^{r'}(\Omega)} \right) \right\}. \tag{4.4.81}
\]

**Remark 4.4.1** For the case \( n = 3 \), the above choices of approximation spaces produce the same error estimate as stated in Theorem 4.4.3 with the restriction that \( 6/5 < r < 2 \). This is due to a restriction on the domain of the Raviart-Thomas interpolation operator (see [26]).

### 4.4.3 Higher Order Approximation

In this section, approximation spaces of higher order are considered. For \( k \geq 1 \), define the following discrete spaces:

\[
T_h := \{ \phi \in T : \phi|_K \in (P_k(K))^2, \quad \forall K \in T_h \},
\]

\[
T_{\text{div},h}' := \left\{ \psi \in T_{\text{div}}' : \psi = (\psi_1 \quad \psi_2)^T|_K \in (RT_k(K))^2, \right. \quad (\psi_1 \quad \psi_2)^T|_K \in RT_k(K), \quad \forall i \in \{1, 2\}, \quad \forall K \in T_h \},
\]

\[
P_h := \{ p \in P : p|_K \in P_k(K), \quad \forall K \in T_h \},
\]

\[
U_h := \{ u \in U : u|_K \in (P_k(K))^2, \quad \forall K \in T_h \}.
\]

**Remark 4.4.2** Note that there is no interelement continuity requirement on the spaces \( T_h, U_h \), and \( P_h \).

Let \( s > 1 \) and let \( T_{h,k}^k : (W^{1,s}(\Omega))^2 \longrightarrow T_{\text{div},h}' \) be the \( k \)-th order Raviart-Thomas interpolation operator [13], defined by, for row \( j = 1, 2 \) of \( \tau \in T_{\text{div}}' \),

\[
\int_{e_i} (\tau_j - T_{h,k}^k \tau_j) \cdot n_i v_k \ ds = 0, \quad \forall v_k \in P_k(K), \quad \forall e_i \in \partial K, \quad i = 1, 2, 3, \quad \forall K \in T_h,
\]

\[
\int_K (\tau_j - T_{h,k}^k \tau_j) \cdot v_{k-1} \ dK = 0, \quad \forall v_{k-1} \in (P_{k-1}(K))^2, \quad \forall K \in T_h.
\]

Then, for \( 0 \leq m \leq k + 1 \), we have

\[
\|\tau - T_{h,k}^k \tau\|_{0,r',\Omega} \leq C h^m |\tau|_{m,r',\Omega}, \tag{4.4.82}
\]

\[
\|\text{div} (\tau - T_{h,k}^k \tau)\|_{0,r',\Omega} \leq C h^m |\text{div} \tau|_{m,r',\Omega}. \tag{4.4.83}
\]
and, for $v \in U$,
\[
\int_{\Omega} v \cdot \text{div}(\tau - \tau_h^k) \, d\Omega = 0, \quad \forall \tau \in T'_{\text{div}}. \tag{4.4.84}
\]

In the lowest-order case, the special functions that were constructed to show the inf-sup conditions (4.4.47) and (4.4.56), for example
\[
\phi^* = -\frac{|q_h I + \tau_h|^{r'/r-1} (q_h I + \tau_h)}{\|q_h I + \tau_h\|^{r'-1}_{T'}},
\]
were readily available in the appropriate piecewise constant function spaces. However, for higher-order approximation, the analogous functions do not lie in polynomial spaces for $1 < r < 2$. Nevertheless, one can find functions in the appropriate polynomial spaces that share the same important features of these special functions, which are related to the norm and $L^2$ inner product.

From [1], if $f \in L^2(\Omega)$, then $f \in L^r(\Omega)$ for $1 < r < 2$, and we also have
\[
\|f\|_{0,r,\Omega} \leq \|f\|_{0,2,\Omega} \cdot \frac{\Omega^{\frac{2}{r}}}{\Omega^{\frac{2}{r'}}}.
\]

Another result regarding the relationship of norms in $L^2(\Omega)$ and $L^r(\Omega)$ is now presented.

**Lemma 4.4.4** Let $f, g \in L^2(\Omega)$ and $1 < r < 2$. If $\|f\|_{0,r} = \mu$ and $\|f\|_{0,2} = \|g\|_{0,2}$, then $\|g\|_{0,r} \leq \mu$.

**Proof:** Assume $\|f\|_{0,r} = \mu$ and $\|f\|_{0,2} = \|g\|_{0,2}$. Let $\alpha = r/2$. Then $1/\alpha = 2/r$ and $1/\alpha + 1/\beta = 1$ implies $\beta = r/(r - 2)$. Then by Hölder’s inequality we have
\[
\|f\|^2_{0,2} = \int_{\Omega} |f|^2 \, d\Omega \\
\leq \left( \int_{\Omega} (|f|^2)^{r/2} \, d\Omega \right)^{2/r} \left( \int_{\Omega} (1)^{r/(r-2)} \, d\Omega \right)^{(r-2)/r} \\
= \left( \int_{\Omega} |f|^r \, d\Omega \right)^{2/\alpha} \left( \int_{\Omega} 1 \, d\Omega \right)^{(r-2)/\alpha} \\
= (\|f\|_{0,r}^r)^{2/\alpha} |\Omega|^{(r-2)/\alpha} \\
= (\mu^r)^{2/\alpha} |\Omega|^{(r-2)/\alpha} \\
= \mu^2 |\Omega|^{(r-2)/r},
\]
thus
\[ \|f\|_{0,2} \leq \mu |\Omega|^{(r-2)/2r}. \] (4.4.85)

Now if \( \alpha = 2/r \), then \( 1/\alpha + 1/\beta = 1 \) implies \( \beta = 2/(2 - r) \). Therefore,

\[
\|g\|_{r,0} = \int_{\Omega} |g|^r \, d\Omega \\
\leq \left( \int_{\Omega} |g|^{r/2} \, d\Omega \right)^{r/2} \left( \int_{\Omega} (1)^{(2-r)/2} \, d\Omega \right)^{(2-r)/2} \\
= \left( \int_{\Omega} |g|^2 \, d\Omega \right)^{r/2} \left( \int_{\Omega} 1 \, d\Omega \right)^{(2-r)/2} \\
= \|g\|_{0,2} |\Omega|^{(2-r)/2} \\
= \|f\|_{0,2} |\Omega|^{(2-r)/2} \\
\leq \left( \mu |\Omega|^{(r-2)/2r} \right)^r |\Omega|^{(2-r)/2} \\
= \mu^r |\Omega|^{(r-2)/2} |\Omega|^{(2-r)/2} \\
= \mu^r |\Omega|^{(r-2+2-2)/r} = \mu^r.
\]

Hence \( \|g\|_{0,r} \leq \mu \).

The inf-sup conditions (4.4.47) and (4.4.48) are now shown to hold for \( k \geq 1 \).

**Lemma 4.4.5** For the choices of \( T_h, T_{\text{div},h}, P_h, \) and \( U_h \) above, there exists a positive constant \( c_1 \) such that

\[
\inf_{(\tau_h,q_h) \in Z_{1h}} \sup_{\phi_h \in T_h} \frac{[B(\phi_h),(\tau_h,q_h)]}{\|\phi_h\|_T \|\tau_h,q_h\|_{T_{\text{div}} \times P}} \geq c_1.
\]

**Proof:** Note that for \((\phi_h,q_h) \in Z_{1h}, \text{div} \tau_h = 0 \) implies \( \tau_h|_K \in (P_k(K))^{2\times2} \) for all \( K \in T_h \).

We also have that \((\tau_h + q_h \mathbf{I})|_K \in (P_k(K))^{2\times2} \) for all \( K \in T_h \). Thus \((\tau_h,q_h) \in Z_{1h}\) implies \( \tau_h \in T_h \) and \((\tau_h + q_h \mathbf{I}) \in T_h \).

Assume that \( \|q_h\|_P \leq \|\tau_h\|_{T_{\text{div}}} \). Let \( \tau_h^0 = \tau_h - \frac{1}{n} \text{tr}(\tau_h) \mathbf{I} \), and

\[
\phi^* = -|\tau_h^0|^{r'/r-1} \tau_h^0/\|\tau_h^0\|_{T_{\text{div}}}^{r'/r-1}.
\]

Then \( \|\phi^*\|_T = 1 \) and

\[
(\phi^*,\phi_h) = \int_{\Omega} \phi^* : \phi_h \, d\Omega, \quad \forall \phi_h \in T_h
\]
defines a continuous linear functional on $T_h$. Note that $T_h$ equipped with the $L^2$ inner product is a Hilbert space. Then by the Riesz Representation Theorem there exists a $\varsigma_h \in T_h$ such that

$$(\varsigma_h, \phi_h) = \int_{\Omega} \varsigma_h : \phi_h \, d\Omega = \int_{\Omega} \phi^* : \phi_h \, d\Omega = (\phi^*, \phi_h), \quad \forall \phi_h \in T_h,$$

with

$$\|\varsigma_h\|_{0,2} = \|\phi^*\|_{0,2}.$$  

From Lemma 4.4.4, we have that $\|\varsigma_h\|_{T} \leq 1$. Then $[B(\varsigma_h), (\tau_h, p_h)] = [B(\phi^*), (\tau_h, p_h)]$ for all $(\tau_h, q_h) \in Z_{1h}$. Continuing as in (4.3.22), the result is shown as in Case 1 of Lemma 4.3.3.

Now assume $\|q_h\|_{P} \geq \|\tau_h\|_{T_{div}}$. Let

$$\phi^* = \frac{-|q_hI + \tau_h|^{r'/r-1} (q_hI + \tau_h)}{\|q_hI + \tau_h\|_{T_{div}}^{r'-1}}.$$  

Again let $\varsigma_h \in T_h$ satisfy $(\varsigma_h, \phi_h) = (\phi^*, \phi_h)$ for all $\phi_h \in T_h$ and $\|\phi^*\|_{0,2} = \|\varsigma_h\|_{0,2} \leq \|\phi^*\|_{T} = 1$. Continuing as in the proof of Case 2 of Lemma 4.3.3, the result is shown.  

**Lemma 4.4.6** For the choices of $T_h$, $T_{div,h}$, $P_h$, and $U_h$ above, there exists a positive constant $c_2$ such that

$$\inf_{(u_h, \lambda_h) \in U_h \times R} \sup_{(\tau_h, q_h) \in T_{div,h} \times P_h} \frac{[C(\tau_h, q_h), (u_h, \lambda_h)]}{\|\tau_h, q_h\|_{T_{div}} \times P_h \| (u_h, \lambda_h) \|_{U \times R}} \geq c_2.$$  

**Proof:** The result is shown in a manner identical to the proof of Lemma 4.4.2, with the $k$-th order interpolation operator $I_h^k$.  

Before showing the inf-sup condition (4.4.11) holds for the chosen approximation spaces, we first discuss some properties of the Raviart-Thomas elements of order $k \geq 1$. Let $K \in T_h$ and let $r \in \mathcal{RT}_k(K)$. Then $r$ can be written as $r = r^k + r^*$, where $r^k \in (\mathcal{P}_k(K))^2$ and the components of $r^*$ consist of polynomial terms of degree $k + 1$ only. In fact, $r^*$ can...
be written as
\[
\mathbf{r}^* = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} k \sum_{j=0}^{k} \gamma_j x_1^{k-j} x_2^j = \left[ \sum_{j=0}^{k} \gamma_j x_1^{k-j+1} x_2^j \right],
\]
for some constants \( \gamma_j, j = 0, \ldots, k \).

We can also write \( \text{div } \mathbf{r} = \text{div } \mathbf{r}^k + \text{div } \mathbf{r}^* \), where \( \text{div } \mathbf{r}^k \) is a polynomial of degree at most \( k - 1 \) and \( \text{div } \mathbf{r}^* \) is a polynomial with terms of degree \( k \) only. It is important to note that if \( \text{div } \mathbf{r} = 0 \), then \( \text{div } \mathbf{r}^* = 0 \) (as the polynomials in \( \mathbf{r}^* \) are linearly independent of the polynomials in \( \mathbf{r}^k \)), and thus
\[
0 = \text{div } \mathbf{r}^* = \frac{\partial r^*_1}{\partial x_1} + \frac{\partial r^*_2}{\partial x_2}
= \sum_{j=0}^{k} (k-j+1) \gamma_j x_1^{k-j} x_2^j + \sum_{j=0}^{k} (j+1) \gamma_j x_1^{k-j} x_2^j = (k+2) \sum_{j=0}^{k} \gamma_j x_1^{k-j} x_2^j,
\]
which implies that \( \gamma_j = 0 \) for all \( 0 \leq j \leq k \). Hence \( \text{div } \mathbf{r} = 0 \Rightarrow \mathbf{r} \in (P_k(K))^2 \) and thus \( \mathbf{r}^* = 0 \).

The following lemma is a result from the general theory of finite-dimensional normed spaces (see [52]).

**Lemma 4.4.7** Let \( \{\mathbf{v}_0, \ldots, \mathbf{v}_n\} \) be a linearly independent set of vectors in a normed space \( X \) of dimension at least \( n + 1 \). Then, there is a constant \( C_* > 0 \) such that for every choice of scalars \( \gamma_0, \ldots, \gamma_n \), we have
\[
\|\gamma_0 \mathbf{v}_0 + \cdots + \gamma_n \mathbf{v}_n\| \geq C_*(|\gamma_0| + \cdots + |\gamma_n|).
\]

The preceeding lemma is used to show that on each triangle, the norm of the gradient of the highest-degree terms of a Raviart-Thomas element can be bounded by the norm of the divergence, which will be used in establishing an approximation property.

**Lemma 4.4.8** Let \( K \in T_h \) and let \( \mathbf{r} = \mathbf{r}^k + \mathbf{r}^* \in RT_k(K) \) where the components of \( \mathbf{r}^* \) consist of polynomial terms of degree \( k + 1 \) only. Then there exists a constant \( \tilde{C} > 0 \),
independent of $K$, such that
\[
\|\nabla r^*\|_{0,r',K} \leq \tilde{C} \| \text{div} r \|_{0,r',K} .
\] (4.4.87)

**Proof:** Let the finite-dimensional vector space $X$ be defined by
\[
X = \text{span} \left\{ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} : x_1^{k-j} x_2^j, \ j = 0, \ldots, k \right\} = \text{span} \{ v_j, \ j = 0, \ldots, k \}.
\]
Define the functional $\| \cdot \|_{\text{grad}} : X \to \mathbb{R}$ by,
\[
\| v \|_{\text{grad}} = \int_K \left| \frac{\partial v_1}{\partial x_1} \right| + \left| \frac{\partial v_2}{\partial x_2} \right| + \left| \frac{\partial v_1}{\partial x_2} \right| + \left| \frac{\partial v_2}{\partial x_1} \right| \ dK .
\] (4.4.88)

We now show that $\| \cdot \|_{\text{grad}}$ defines a norm on the space $X$. Of course, $\| v \|_{\text{grad}} \geq 0$ for all $v \in X$. To show $\| v \|_{\text{grad}} = 0$ if and only if $v = 0$, suppose that there exists a nonzero $v \in X$ with $\| v \|_{\text{grad}} = 0$. Then (4.4.88) requires all of the partial derivatives of the components of $v$ to be zero. Note that since $k \geq 1$, the partial derivatives of the components of $v$ are functions with polynomial terms of degree $k$ only. But since $v = \gamma_0 v_0 + \cdots + \gamma_k v_k \neq 0$, there is at least one nonzero $\gamma_j, 0 \leq j \leq k$, and thus the partial derivative $\partial v_1 / \partial x_1$ contains a term of the form $(k-j+1)\gamma_j x_1^{k-j} x_2^j$, a contradiction. It is easy to see that $\| \alpha v \|_{\text{grad}} = |\alpha| \| v \|_{\text{grad}}$ and that the triangle inequality holds for $\| \cdot \|_{\text{grad}}$. Thus $\| \cdot \|_{\text{grad}}$ defines a norm on $X$. In fact, we have that $\| v \|_{\text{grad}} = \| \nabla v \|_{0,1,K}$.

Now define the functional $\| \cdot \|_{\text{div}} : X \to \mathbb{R}$ by,
\[
\| v \|_{\text{div}} = \int_K \left| \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} \right| \ dK .
\]
This functional also defines a norm, in particular because $\text{div} v = 0$ implies that $v = 0$. We also have that $\| v \|_{\text{div}} = \| \text{div} v \|_{0,1,K}$. By the equivalence of norms on a finite-dimensional vector space, we have that, there exist constants $C_1, C_2$, and $C_3$, depending only on the dimension of the space (in this case $k + 1$), such that
\[
\| \nabla v \|_{0,r',K} \leq C_1 \| \nabla v \|_{0,1,K} \leq C_2 \| \text{div} v \|_{0,1,K} \leq C_3 \| \text{div} v \|_{0,r',K}.
\]
Thus there is a $C_K > 0$ such that

$$\|\nabla r^*\|_{0,r',K} \leq C_K \|\text{div} r^*\|_{0,r',K} \quad (4.4.89)$$

for all $K \in T_h$. The dependence of $C_K$ on $K \in T_h$ is due to the integral over $K$. The condition (4.4.1) guarantees that $T_h$ is a quasi-uniform triangulation of $\Omega$, thus we can find a global constant $C$, independent of $K$, such that

$$\|\nabla r^*\|_{0,r',K} \leq C \|\text{div} r^*\|_{0,r',K} \quad (4.4.90)$$

for all $K \in T_h$.

Now, let $X_k$ be the finite dimensional vector space spanned by the polynomials of degree $k$ only, and let $X = P_k(K)$. Note that $X = P_{k-1}(K) \oplus X^k$, and that $\text{div} r \in X$, $\text{div} r^k \in P_{k-1}(K)$, and $\text{div} r^* \in X^k$. Let $\{v_0, \ldots, v_k, \ldots, v_n\}$ be a basis for $X$ where $\{v_0, \ldots, v_k\}$ is also a basis for $X_k$. From Lemma 4.4.7, there is a constant $C_* > 0$ such that, for all $v = \gamma_0v_0 + \cdots + \gamma_nv_n \in X$,

$$\|v\|_{0,r',K} \geq C_* (|\gamma_0| + \cdots + |\gamma_n|).$$

Define the functional $\| \cdot \|_* : X \to \mathbb{R}$ for $v = \gamma_0v_0 + \cdots + \gamma_nv_n$ by

$$\|v\|_* = C_* (|\gamma_0| + \cdots + |\gamma_n|).$$

It is straightforward to show that $\| \cdot \|_*$ defines a norm on $X$. Then, by the definition of $r, r^*$, the equivalence of norms on a finite-dimensional space, and the quasi-uniform triangulation $T_h$, we have that there is a constant $C_4$, dependent only upon $n$ (which itself is dependent only upon $k$) such that

$$\|\text{div} r^*\|_{0,r',K} \leq C_4\|\text{div} r^*\|_* = C_4C_* (|\gamma_0| + \cdots + |\gamma_k|)$$

$$\leq C_4C_* (|\gamma_0| + \cdots + |\gamma_k| + \cdots + |\gamma_n|) = C_4\|\text{div} r\|_* \leq C_4\|\text{div} r\|_{0,r',K}. \quad (4.4.91)$$

Combining (4.4.90) and (4.4.91) the result is shown.

The above results are easily extended to the tensor space $T_{\text{div},h}^\prime$ to obtain, for $\tau_h = \tau^k + \tau^*$
where the components of $\tau^*$ consist of polynomial terms of degree $k + 1$ only,

$$
\| \nabla \tau^* \|_{0,r,K} \leq \tilde{C} \| \text{div} \tau_h \|_{0,r',K}, \quad \forall K \in T_h. \tag{4.4.92}
$$

Let $\Pi_k : T^\prime_{\text{div},h} \rightarrow T_h$ be the classical Lagrangian $P_k$ interpolation operator ([26]) and define

$$
\hat{\tau} = \tau^k + \Pi_k \tau^*. \tag{4.4.93}
$$

Note that $\hat{\tau}|_K \in (P_k(K))^{2 \times 2}$ for all $K \in T_h$, and $\text{div} \tau_h = 0$ implies $\tau^* = 0$ and $\hat{\tau} = \tau_h$.

Then, using (4.4.92) and standard polynomial approximation properties [12, 26], the error associated in the approximation of $\tau_h$ by $\hat{\tau}$ is given by

$$
\| \tau_h - \hat{\tau} \|_{0,r',\Omega} = \| \tau^* - \Pi_k \tau^* \|_{0,r',\Omega}
\leq C \ h \left( \sum_{K \in T_h} \| \nabla \tau^* \|_{0,r',K} \right)^{1/r'}
\leq C \ h \left( \sum_{K \in T_h} \tilde{C} \| \text{div} \tau_h \|_{0,r',K} \right)^{1/r'}
\leq C \tilde{C} \ h \| \text{div} \tau_h \|_{0,r',\Omega} = \tilde{C} \ h \| \text{div} \tau_h \|_{0,r',\Omega}. \tag{4.4.94}
$$

**Lemma 4.4.9** For $h$ sufficiently small, there is a constant $c_3 > 0$ such that

$$
\inf_{(\tau_h,q_h) \in T^\prime_{\text{div},h} \times P_h} \sup_{(\mathbf{s}_h,\mathbf{v}_h,\eta_h) \in T_h \times \mathbf{U}_h \times \mathbb{R}} \left[ \mathbf{B}(\mathbf{s}_h, (\tau_h,q_h)) + [\mathbf{C}(\tau_h,q_h), (\mathbf{v}_h,\eta_h)] \right] \geq c_3. \tag{4.4.95}
$$

where $\| (\mathbf{s}_h,\mathbf{v}_h,\eta_h) \|_{T \times U \times \mathbb{R}} = \| \mathbf{s}_h \|_T + \| \mathbf{v}_h \|_U + \| \lambda_h \|_{\mathbb{R}}$.

**Proof:** The approach is similar to the proof of Lemma 4.4.3.

Case 1: ($\| \tau_h \|_{T^\prime_{\text{div}}} \leq \| q_h \|_P$)

Let $\check{\tau}$ be as defined in (4.4.93). Let

$$
\phi^* = \frac{-1}{\| q_h I + \check{\tau} \|_{T^\prime}^{r'-1}} q_h I + \check{\tau} |^{r'/r'-1} (q_h I + \check{\tau}).
$$
Note that \(\|\phi^*\|_T = 1\). Now, there exists a \(s_h \in T_h\) such that
\[
\int_{\Omega} s_h : \phi_h \, d\Omega = \int_{\Omega} \phi^* : \phi_h \, d\Omega, \quad \forall \phi_h \in T_h,
\]
with \(\|s_h\|_T \leq \|\phi^*\|_T = 1\). Then, since \(q_h I + \hat{\tau} \in T_h\), we have that
\[
\int_{\Omega} s_h : (q_h I + \hat{\tau}) \, d\Omega = \int_{\Omega} \phi^* : (q_h I + \hat{\tau}) \, d\Omega.
\]
Then we have
\[
\frac{[B(s_h), (\tau_h, q_h)]}{\|s_h\|_T} = -\int_{\Omega} s_h : (q_h I + \tau_h) \, d\Omega
\]
\[
= -\int_{\Omega} s_h : (q_h I + \hat{\tau}) \, d\Omega - \int_{\Omega} s_h : (\tau_h - \hat{\tau}) \, d\Omega
\]
\[
= \int_{\Omega} \|q_h I + \hat{\tau}\|_{r'/r-1} (q_h I + \hat{\tau}) : (q_h I + \hat{\tau}) \, d\Omega
\]
\[
- \int_{\Omega} s_h : (\tau_h - \hat{\tau}) \, d\Omega
\]
\[
\geq \|q_h I + \hat{\tau}\|_{T'} - \|s_h\|_T \|\tau_h - \hat{\tau}\|_{T'}
\]
\[
\geq (n^{1/r'} - 1) \|q_h\|_P - 2\hat{C}h \|\text{div} \, \tau_h\|_{0,r',\Omega}. \tag{4.4.96}
\]

Let
\[
u_* = \frac{-(n^{1/r'} - 1)\|\text{div} \, \tau_h\|_{r'/r-1}}{\|\text{div} \, \tau_h\|_{0,r',\Omega}} (\text{div} \, \tau_h),
\]
and note that \(\|\nu^*\|_U = n^{1/r'} - 1\). Recall that \(\text{div} \, \psi_h \in U_h\) for all \(\psi_h \in T'_{\text{div}, h}\). Thus, there is a \(v_h \in U_h\) such that
\[
\int_{\Omega} v_h : w_h \, d\Omega = \int_{\Omega} \nu^* : w_h \, d\Omega, \quad \forall w_h \in U_h,
\]
with \(v_h \in U_h\) and \(\|v_h\|_U \leq \|\nu^*\|_U = n^{1/r'} - 1\). Let \(\eta_h = 0\), then we have
\[
[C(\tau_h, q_h), (v_h, \eta_h)] = (n^{1/r'} - 1) \int_{\Omega} \frac{\|\text{div} \, \tau_h\|_{r'/r-1}}{\|\text{div} \, \tau_h\|_{0,r',\Omega}} (\text{div} \, \tau_h) : (\text{div} \, \tau_h) \, d\Omega
\]
\[
= (n^{1/r'} - 1) \|\text{div} \, \tau_h\|_{0,r',\Omega}. \tag{4.4.97}
\]
Thus, from (4.4.96) and (4.4.97), we have

\[
\left[ B(\varsigma_h), (\tau_h, q_h) \right] + \left[ C(\tau_h, q_h), (v_h, \eta_h) \right] \geq \frac{1}{n^{1/r'}} \left( (n^{1/r'} - 1)\|q_h\|_P + (n^{1/r'} - 1 - 2\hat{c}h)\|\text{div}\,\tau_h\|_0, r', \Omega \right),
\]

(4.4.98)

and, for \( h \) small enough to satisfy \( n^{1/r'} - 1 - 2\hat{c}h > 0 \), we have that

\[
\left[ B(\varsigma_h), (\tau_h, q_h) \right] + \left[ C(\tau_h, q_h), (v_h, \eta_h) \right] \geq C\| (\tau_h, q_h) \|_{T' \times P},
\]

for some constant \( C > 0 \).

**Case 2:** \( (\|\tau_h\|_{T'_d} \geq \|q_h\|_P) \)

Let

\[
\tilde{\tau} = \tilde{\tau} - \frac{1}{n} \left( \int_{\Omega} \text{tr}(\tilde{\tau}) \, d\Omega \right) \mathbf{I},
\]

and

\[
\tau^0 = \tilde{\tau} - \frac{1}{n} \text{tr}(\tilde{\tau}) \mathbf{I}.
\]

Let

\[
\phi^* = \frac{-|\tau^0|^{r'/r - 1}}{\|\tau^0\|^{r'/r - 1}} \tau^0,
\]

and note that there is an \( \varsigma_h \in T_h \) satisfying

\[
\int_{\Omega} \varsigma_h : \phi_h \, d\Omega = \int_{\Omega} \phi^* : \phi_h \, d\Omega, \quad \forall \phi_h \in T_h,
\]

with \( \|\varsigma_h\|_T \leq \|\phi^*\|_T = 1 \). Note that \( \tilde{\tau}, \tau^0 \in T_h \). Let

\[
u^* = \frac{-2|\text{div}\,\tau_h|^{r'/r - 1}}{\|\text{div}\,\tau_h\|^{r'/r - 1}} (\text{div}\,\tau_h),
\]

and note that there is a \( v_h \in U_h \) such that

\[
\int_{\Omega} v_h \cdot w_h \, d\Omega = \int_{\Omega} u^* \cdot w_h \, d\Omega, \quad \forall w_h \in U_h,
\]

with \( \|v_h\|_U \leq \|u^*\|_U = 2 \). Let

\[
\eta_h = \text{sgn} \left( \int_{\Omega} \text{tr}(\tau_h) \, d\Omega \right) \left( 1 + \frac{1}{C_0} \right) n^{-1/2} |\Omega|^{1/r'}.
\]
Continuing as in (4.4.71)-(4.4.74), replacing $\sqrt{n} \hat{C}$ with $\hat{C}$, if $h$ is small enough to guarantee
\[ C_0 > \hat{C} h (1 + 2 C_0 + (1 + C_0) |\Omega|), \]
then the result is shown.

From the standard approximation properties [26, 13], the following error estimate is derived.

**Theorem 4.4.4** Let $f \in \left( L^r'(\Omega) \right)^2$ and $u_\Gamma \in \left( W_{1-1/r}^1(\Omega) \right)^2$. Let $(\phi, \psi, p, u, \lambda) \in T \times T'_{\text{div}} \times P \times U \times \mathbb{R}$ solve (4.2.16)-(4.2.18) and let $(\phi_h, \psi_h, p_h, u_h, \lambda_h) \in \mathcal{T}_h \times \mathcal{T}'_{\text{div}, h} \times P_h \times U_h \times \mathbb{R}$ solve (4.4.5)-(4.4.7). Assume $1 \leq m \leq k + 2$ and $(\phi, \psi, p, u) \in (W_{m,r}^m(\Omega))^{2 \times 2} \times (W_{m,r'}^m(\Omega))^2 \times (W_{m,r'}^m(\Omega))^2$ with $\text{div} \psi \in (W_{m,r'}^m(\Omega))^2$. Then there exists a positive constant $C$ such that
\[ \| \phi - \phi_h \|_T + \| u - u_h \|_U + |\lambda - \lambda_h| \leq C \left\{ h^{m/r/2} \left( \| \phi \|_{m,r,\Omega} + \| u \|_{m,r,\Omega} \right) + h^m \left( \| \psi \|_{m,r',\Omega} + \| p \|_{m,r',\Omega} \right) \right\}, \quad (4.4.99) \]
\[ \| \psi - \psi_h \|_{T'_{\text{div}}} + \| p - p_h \|_P \leq C \left\{ h^{m/r'} \left( \| \phi \|_{m,r,\Omega} + \| u \|_{m,r,\Omega} \right) + h^{2m/r'} \left( \| \psi \|_{m,r',\Omega} + \| p \|_{m,r',\Omega} \right) \right\}. \quad (4.4.100) \]

**Remark 4.4.3** The above estimate is pessimistic if $|\phi|$ and $|\phi_h|$ can be bounded away from zero. Then, if $\phi \in (W_{m,r}^m(\Omega))^{2 \times 2}$ for $m \geq 2$, the quantity $\| \phi - \phi_h \|_\infty$ can be bounded in (4.4.18) and (4.4.37) to improve (4.4.99) and (4.4.100).

### 4.5 Numerical Experiments

In this section we describe numerical experiments that support the theoretical results outlined in Sections 4.3 and 4.4. The first example illustrates the theoretical rate of convergence of the solution method and the second example illustrates the computed approximation for a benchmark physical problem. Computations are performed using the FreeFEM++ finite element software package [45]. All computations below are performed in the lowest-order case ($k = 0$).
4.5.1 Example 1

For this example (taken from [37]) approximations are computed for a power law fluid with $\nu_0 = 1.0$. The computational domain is $\Omega = [0, 2] \times [0, 2]$, with $f$ and $u_\Gamma$ chosen so that the exact solution of (4.2.13)-(4.2.15) is given by

$$u = \begin{bmatrix} \frac{-1}{(4.1-x_1-x_2)^{1/3}} \\ \frac{1}{(4.1-x_1-x_2)^{1/3}} \end{bmatrix} \quad \text{and} \quad p = x_1 + x_2,$$

for all $(x_1, x_2) \in \Omega$. Computations are performed on uniform meshes of decreasing size $h$ and for selected values of $r$. For values of $1 < r < 2$, the resulting system of equations is nonlinear, and a fixed-point iteration is used to compute approximations. The fixed-point iteration is terminated when the pointwise maximum absolute difference in successive approximations falls below $10^{-8}$. The quantity $N$ represents the total number of degrees of freedom for the approximating system. The results are shown in Table 4.5.1. It is observed that, for all choices of $r$, the convergence rate approaches 1 for all solution components as the mesh is refined. Thus the observed rate of convergence is higher than the theoretical rate shown in (4.4.80)–(4.4.81) for $r < 2$.

4.5.2 Example 2

This example is the benchmark driven cavity problem. For $\Omega = [0, 1] \times [0, 1]$, we have that $f = 0$ and

$$u_\Gamma = 0 \quad \text{on} \quad \Gamma \setminus \Gamma_{\text{top}} \quad \text{and} \quad u_\Gamma = [1 \ 0]^T \quad \text{on} \quad \Gamma_{\text{top}},$$

where $\Gamma_{\text{top}}$ is the portion of the boundary satisfying $0 \leq x_1 \leq 1$ and $x_2 = 1$. Computations were performed for a power law fluid with $\nu_0 = 1.0$ and selected values of $r$. Figures 4.5.1, 4.5.2, and 4.5.3 show plots of the streamlines computed for $h = 1/32$ for $r = 2$, $r = 1.5$, and $r = 1.1$, respectively. As the power $r$ in the constitutive law is decreased, we see a movement of the central vortex toward the top of the cavity, corresponding to an increase in viscosity. Table 4.5.2 shows the norms of computed approximations for the solution components, along with the relative difference in the norms on successive meshes. For example, the relative
difference in solution norms for $\mathbf{u}_h$ on meshes $h$ and $h/2$ is computed as

$$relative\ difference = \frac{\|\mathbf{u}_h\|_U - \|\mathbf{u}_{h/2}\|_U}{\|\mathbf{u}_{h/2}\|_U}.$$ 

It is observed that, as the mesh is refined, the relative difference in the norms of the computed approximations decreases.

![Figure 4.5.1 Streamlines for $r = 2.0$, driven cavity](image)

Figure 4.5.1 Streamlines for $r = 2.0$, driven cavity
Figure 4.5.2  Streamlines for $r = 1.5$, driven cavity

Figure 4.5.3  Streamlines for $r = 1.1$, driven cavity
<table>
<thead>
<tr>
<th>$r$</th>
<th>$N$</th>
<th>1</th>
<th>1/2</th>
<th>1/4</th>
<th>1/8</th>
<th>1/16</th>
<th>1/32</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>89</td>
<td>0.232668</td>
<td>0.126054</td>
<td>0.0661146</td>
<td>0.0345017</td>
<td>0.0177864</td>
<td>0.00902893</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.243553</td>
<td>0.122596</td>
<td>0.0611162</td>
<td>0.0306009</td>
<td>0.0153377</td>
<td>0.0076742</td>
</tr>
<tr>
<td>3/2</td>
<td>337</td>
<td>0.245281</td>
<td>0.125628</td>
<td>0.0660665</td>
<td>0.0344038</td>
<td>0.0176294</td>
<td>0.0090314</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.257042</td>
<td>0.128567</td>
<td>0.069924</td>
<td>0.0373716</td>
<td>0.019509</td>
<td>0.009976</td>
</tr>
<tr>
<td>5/4</td>
<td>1313</td>
<td>0.258248</td>
<td>0.12961</td>
<td>0.0708665</td>
<td>0.0383333</td>
<td>0.020482</td>
<td>0.0100356</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.269402</td>
<td>0.132553</td>
<td>0.0748108</td>
<td>0.0413157</td>
<td>0.021462</td>
<td>0.0108936</td>
</tr>
<tr>
<td>11/10</td>
<td>5185</td>
<td>0.296871</td>
<td>0.137459</td>
<td>0.0655466</td>
<td>0.031775</td>
<td>0.0155934</td>
<td>0.00771504</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.308182</td>
<td>0.142385</td>
<td>0.070492</td>
<td>0.0364716</td>
<td>0.0182974</td>
<td>0.0097192</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.321332</td>
<td>0.147309</td>
<td>0.0754376</td>
<td>0.0411676</td>
<td>0.0209994</td>
<td>0.0107248</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.334879</td>
<td>0.152233</td>
<td>0.0803832</td>
<td>0.0458636</td>
<td>0.0236914</td>
<td>0.0117304</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.348857</td>
<td>0.157157</td>
<td>0.0853288</td>
<td>0.0505596</td>
<td>0.0263834</td>
<td>0.012736</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.362916</td>
<td>0.162081</td>
<td>0.0902744</td>
<td>0.0552556</td>
<td>0.0290754</td>
<td>0.0137416</td>
</tr>
</tbody>
</table>

Table 4.5.1 Approximation errors and rates of convergence for Example 1.
<table>
<thead>
<tr>
<th>$r$</th>
<th>$h$</th>
<th>$1/2$</th>
<th>$1/4$</th>
<th>$1/8$</th>
<th>$1/16$</th>
<th>$1/32$</th>
<th>$1/64$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td></td>
<td>89</td>
<td>337</td>
<td>1313</td>
<td>5185</td>
<td>20609</td>
<td>82177</td>
</tr>
<tr>
<td>$|u_h|_U$</td>
<td>$r.\ d.$</td>
<td>0.225263</td>
<td>0.219973</td>
<td>0.238059</td>
<td>0.251018</td>
<td>0.256354</td>
<td>0.258239</td>
</tr>
<tr>
<td>$|p_h|_P$</td>
<td>$r.\ d.$</td>
<td>1.10583</td>
<td>2.09444</td>
<td>3.30095</td>
<td>4.43678</td>
<td>5.44463</td>
<td>6.33794</td>
</tr>
<tr>
<td>$|\phi_h|_T$</td>
<td>$r.\ d.$</td>
<td>1.79782</td>
<td>2.34254</td>
<td>2.86713</td>
<td>3.33377</td>
<td>3.74988</td>
<td>4.12608</td>
</tr>
<tr>
<td>$|\psi_h|_{T'}$</td>
<td>$r.\ d.$</td>
<td>2.38283</td>
<td>3.77636</td>
<td>5.47841</td>
<td>7.10521</td>
<td>8.56444</td>
<td>9.86729</td>
</tr>
<tr>
<td>$|\text{div}\psi_h|_0,r'$</td>
<td>$r.\ d.$</td>
<td>4.46E-16</td>
<td>1.99E-15</td>
<td>6.65E-15</td>
<td>1.85E-14</td>
<td>4.99E-14</td>
<td>1.10E-13</td>
</tr>
<tr>
<td>$|u_h|_U$</td>
<td>$r.\ d.$</td>
<td>0.159846</td>
<td>0.129975</td>
<td>0.136218</td>
<td>0.148649</td>
<td>0.154646</td>
<td>0.156886</td>
</tr>
<tr>
<td>$|p_h|_P$</td>
<td>$r.\ d.$</td>
<td>0.762013</td>
<td>1.34566</td>
<td>2.02864</td>
<td>2.60278</td>
<td>3.03252</td>
<td>3.347</td>
</tr>
<tr>
<td>$|\phi_h|_T$</td>
<td>$r.\ d.$</td>
<td>1.52356</td>
<td>1.8643</td>
<td>2.18568</td>
<td>2.4316</td>
<td>2.60511</td>
<td>2.72616</td>
</tr>
<tr>
<td>$|\psi_h|_{T'}$</td>
<td>$r.\ d.$</td>
<td>1.58266</td>
<td>2.27248</td>
<td>3.15535</td>
<td>3.92185</td>
<td>4.50248</td>
<td>4.9306</td>
</tr>
<tr>
<td>$|\text{div}\psi_h|_0,r'$</td>
<td>$r.\ d.$</td>
<td>3.60E-16</td>
<td>2.49E-15</td>
<td>4.75E-15</td>
<td>1.21E-14</td>
<td>2.98E-14</td>
<td>6.27E-14</td>
</tr>
<tr>
<td>$|u_h|_U$</td>
<td>$r.\ d.$</td>
<td>0.0947944</td>
<td>0.0505626</td>
<td>0.0274854</td>
<td>0.0186618</td>
<td>0.0215235</td>
<td>0.0267499</td>
</tr>
<tr>
<td>$|p_h|_P$</td>
<td>$r.\ d.$</td>
<td>0.61963</td>
<td>1.26868</td>
<td>1.95128</td>
<td>2.52066</td>
<td>3.0358</td>
<td>3.46015</td>
</tr>
<tr>
<td>$|\phi_h|_T$</td>
<td>$r.\ d.$</td>
<td>1.13417</td>
<td>1.20787</td>
<td>1.28602</td>
<td>1.36521</td>
<td>1.43154</td>
<td>1.47279</td>
</tr>
<tr>
<td>$|\psi_h|_{T'}$</td>
<td>$r.\ d.$</td>
<td>1.20842</td>
<td>1.88946</td>
<td>2.8378</td>
<td>3.64846</td>
<td>4.36081</td>
<td>4.95145</td>
</tr>
<tr>
<td>$|\text{div}\psi_h|_0,r'$</td>
<td>$r.\ d.$</td>
<td>6.54E-16</td>
<td>1.94E-15</td>
<td>3.13E-15</td>
<td>8.86E-15</td>
<td>2.85E-14</td>
<td>7.82E-14</td>
</tr>
</tbody>
</table>

Table 4.5.2 Approximation norms and relative differences for Example 2.
5.1 Defect-Correction Methods for Johnson-Segalman Fluids

A defect-correction method for the solution of viscoelastic flows has been presented in Chapter 2. The method was shown to preserve the order of convergence of the standard nonlinear solution approach as shown in [8]. Numerical experiments for a problem with an analytic solution support the theory. For the benchmark 4:1 contraction flow problem, when good initial approximations are not available the defect-correction method can be used to extend the range of Weissenberg number for which solutions can be computed.

Further studies with the method should include a posteriori error estimates as well as algorithms for choosing the defect parameters. Alternate correction strategies should also be considered, in particular those that allow for approximations that maintain the optimal order of convergence.

5.2 Continuation Methods for Johnson-Segalman Fluids

The continuation methods studied in Chapter 3 provide some insight into the behavior of numerical approximations of steady Johnson-Segalman fluids at high Weissenberg number. For the choices of parameters studied here, a limiting Weissenberg number is found in each case, with or without the presence of large changes in the solutions for small changes in the Weissenberg number.

In particular, for the Oldroyd-B constitutive model, the solution manifold of the discretized system of equations is seen to exhibit large changes in solutions for small changes in Weissenberg number near the limiting value. Nonsmooth flow behavior and sharp stress gradients have been observed for increasing values of the Weissenberg number. A turning point was observed for the Oldroyd-B model on the coarsest mesh, however no turning point was observed otherwise.

It is also observed that the limiting Weissenberg values are similar across the range of methods examined. For pseudo-arclength continuation methods, a spherical constraint
was seen to be more efficient than an orthogonal constraint in a region of high curvature of the solution manifold, while both constraints performed similarly in regions of low or moderate curvature. However, care must be taken to ensure that the spherical constraint does not skip over solutions and merely recompute previous points on the curve.

Further studies related to continuation methods for viscoelastic flows may include the analysis and development of steplength selection algorithms, and the examination of further pseudo-arclength constraint variations. Applications of multiparameter continuation, in which the inflow velocity, $a$, and $\alpha$ are included, as well as studies involving time-dependent formulations of the modeling equations, may provide further understanding of the behavior of the discretized system of equations.

5.3 Dual-Mixed Approximation of Shear-Thinning Fluids

The dual-mixed approximation method studied in Chapter 4 extends the current body of research on such methods by analyzing the problem in the appropriate Sobolev spaces. The analysis produced an error estimate that is dependent on the nonlinearity of the constitutive model, and numerical observations support the theoretical results. The lowest-order finite element approximation produced a method that is well-suited for adaptive computation, and the method can be used with higher-order elements.

Future work related to the dual-mixed approximation method should include the development of a posteriori error estimators that can be useful in adaptive computation strategies. Numerical experiments with higher-order Raviart-Thomas elements should produce better approximations. The theory of the method supports the use of tensor finite elements that preserve the symmetry of the extra stress tensor, and thus the development and implementation of such elements from elasticity theory may yield a more efficient approximation method. In addition, the block structure of the approximating system may yield preconditioning strategies that increase the efficiency of iterative linear solvers used with the method.


